Inferential Statistics Course Project

Wally Thornton

August 22, 2015

This course project consists of two sections, one a simulation exercise and the second an inferential data analysis.

Section 1: A Simulation Exercise

Using the simulation of an exponential distribution, we will explore the relationships between the sample and the theoretical population (including mean and variance), and demonstrate that the distribution of sample means adheres to a Gaussian (normal) distribution, even though the original distribution is exponential.

Run the simulations

The first step is to set up the simulation environment and load packages that might be needed. (Code not shown. Throughout report only code directly generating results will be shown, to save space.)

We then create an exponential distribution simulation with a sample size of 40 (n=40) and a rate of 0.2 (λ=0.2). We'll use the `rexp()` function in R, which randomly pull values from an exponential distribution with a mean of $1/\lambda$, or 5, and repeat this 1,000 times (r=1000) to get a nice, big matrix.

```
set.seed(42)
n <- 40
lambda <- 0.2
r <- 1000
my_samples <- matrix(rexp(n*r, lambda), r)
```

Calculate and compare the sample mean to the theoretical mean

Now that we have our simulation results, we'll calculate the mean for each 40-sample run and capture all 1,000.

```
sample_means <- apply(my_samples, 1, mean)
```

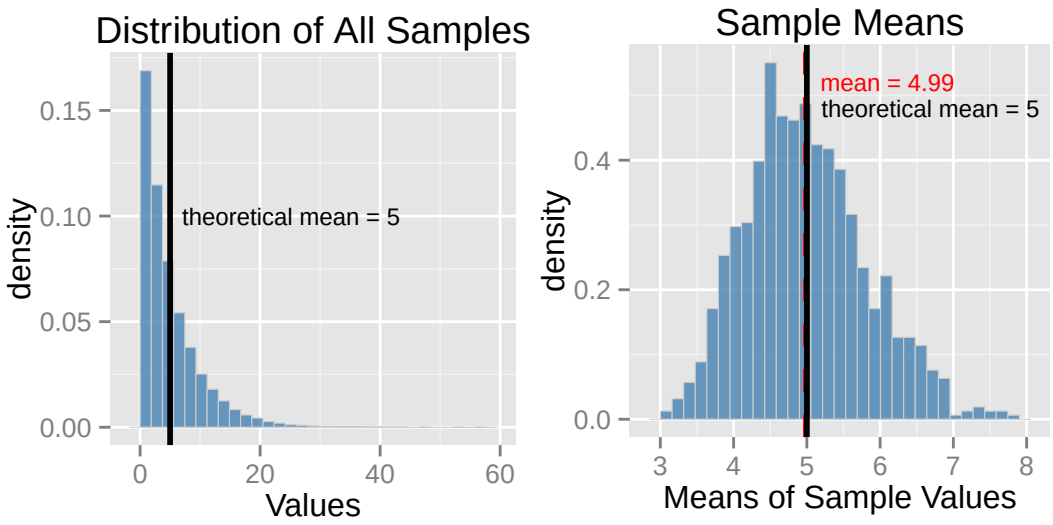
How do they compare to our theoretical mean of 5? While the mean of each sample ranges from 3.14 to 7.88, the mean of the sample means is 4.99, **very close to our theoretical mean of 5**.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.141   4.406   4.919   4.987   5.504   7.882
```

This is visually evident when we plot the distributions:

```
long_sample <- gather(data.frame(c(my_samples)), "sample", "x")
long_means <- gather(data.frame(sample_means), "sample", "x")
long_combined <- bind_rows(long_sample, long_means)
a <- ggplot(long_sample, aes(x=x, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
```

```
    geom_vline(xintercept=1/lambda, linetype="solid", color="black", size = 1) +
    annotate("text", size=3, hjust=0, x = 1/lambda+2, y=.1, color="black"
             , label=paste("theoretical mean =",1/lambda)) +
    xlab("Values") +
    ggtitle("Distribution of All Samples") +
    theme(plot.title=element_text(size=14))
b <- ggplot(long_means, aes(x=x, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
    geom_vline(xintercept=mean(sample_means), linetype="dashed", color="red", size=1) +
    geom_vline(xintercept=1/lambda, linetype="solid", color="black", size=1) +
    annotate("text", size=3, hjust=0, x = mean(sample_means)+.2, y=.52, color="red"
             , label=paste("mean =",round(mean(sample_means),2))) +
    annotate("text", size=3, hjust=0, x = 1/lambda+.2, y=.48, color="black"
             , label=paste("theoretical mean =",round(1/lambda,2))) +
    xlab("Means of Sample Values") +
    ggtitle("Sample Means") +
    theme(plot.title=element_text(size=14))
```



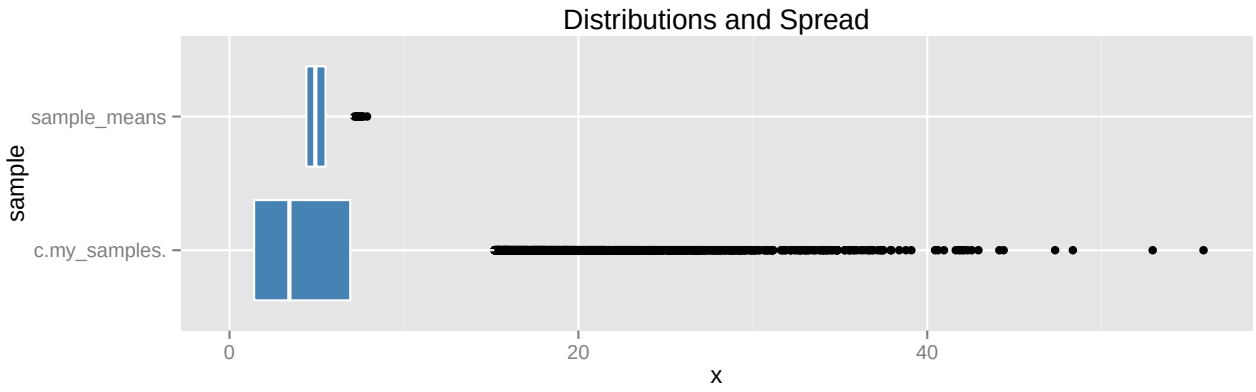
Calculate and compare the sample variance to the theoretical variance

The standard deviation of an exponential distribution is $1/\lambda$, so the variance is be $(1/\lambda)^2$, which results in the theoretical variance of the distribution of 25.

The variance of the sampling distribution of the mean (also known as the standard error of the mean) is defined as $\sigma_\mu^2 = \sigma^2/N$, that is, the population variance divided by the sample size. Plugging in our values results in a standard error of the mean of:

```
## [1] "Standard error of the mean:  0.625"
```

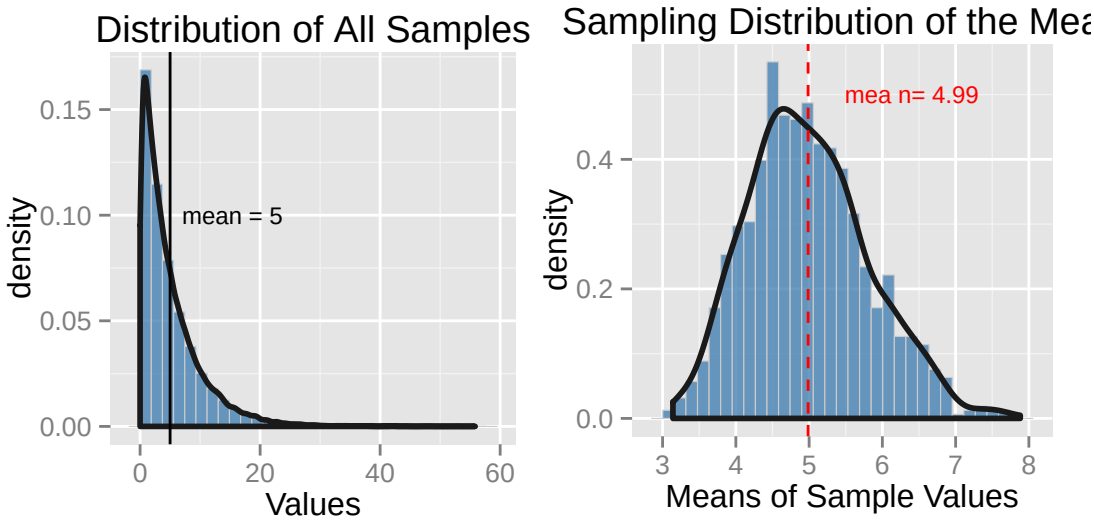
This is significantly less than the theoretical population variance. Why? This is because we are estimating how far each sample mean is likely to be from the population mean and as sample sizes get larger, the standard error will trend toward zero because the estimate improves. This difference in variance shows up in boxplots of the two distributions, with the sample means much more clustered while the population is much more spread out.



Analyze the distribution of the sample mean

Looking at the histogram of the 40,000 randomly generated exponents, it's obvious that the data are not normally distributed and actually follow an exponential distribution. If we plot the means of each 40-value observation, we find that they are not distributed exponentially, but rather follow a normal distribution. Here are the same plots as above, now with their respective density curves overlaid:

```
a <- ggplot(long_sample, aes(x=x, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
    geom_density(color="grey10", size=1) +
    geom_vline(xintercept=1/lambda, linetype="solid", color="black") +
    annotate("text", size=3, hjust=0, x = 1/lambda+2, y=.1, color="black"
             , label=paste("mean =",1/lambda)) +
    xlab("Values") +
    ggtitle("Distribution of All Samples") +
    theme(plot.title=element_text(size=14))
b <- ggplot(long_means, aes(x=x, y=..density..)) +
    geom_histogram(fill="steelblue", color="grey80", size=.2, alpha=.8) +
    geom_density(color="grey10", size=1) +
    geom_vline(xintercept=mean(sample_means), linetype="dashed", color="red") +
    annotate("text", size=3, hjust=0, x = mean(sample_means)+.5, y=.5, color="red"
             , label=paste("mea n=",round(mean(sample_means),2))) +
    xlab("Means of Sample Values") +
    ggtitle("Sampling Distribution of the Mean") +
    theme(plot.title=element_text(size=14))
```



Visually, the distribution of the sample means is much closer to normal than the distribution of the sample itself. If the distribution is normal, we expect the mean of the distribution to be equal to the median and about 95% of the results within 1.96 standard deviations, which we find to be approximately the case:

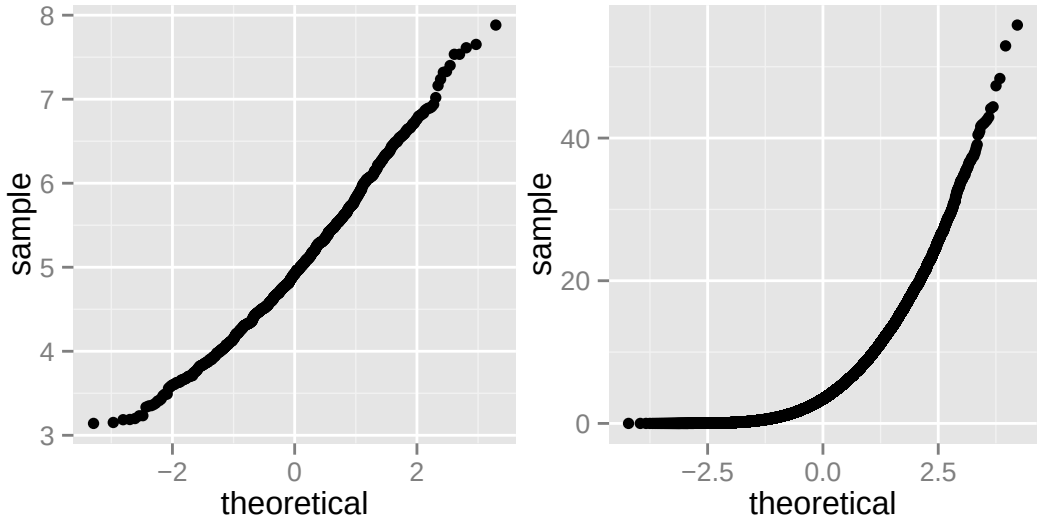
```
lower <- mean(sample_means) - 1.96*sd(sample_means)
upper <- mean(sample_means) + 1.96*sd(sample_means)
set_check <- sum(sample_means > lower & sample_means < upper)/length(sample_means)
```

```
## [1] "Mean:  4.99"
```

```
## [1] "Median:  4.92"
```

```
## [1] "Percentage of sampling means within 1.96 standard deviations:  95.4%"
```

To confirm these checks with something more concrete, we can run a quantile-quantile (Q-Q) plot. The Q-Q plot compares each point in the dataset to where they would be in a perfectly normal distribution with the same mean and standard deviation:



In a Q-Q plot, we look for general linearity as indicative of normality, which we roughly have with all four combinations of dosage and supplement. We could check specifically with a Shapiro-Wilk normality test, but

with large datasets (like our sample) this test can lead to concluding the data is not normal even though in reality it is quite normal.

Given the linearity of the Q-Q plot and what we see with the standard deviations, we can feel comfortable that the sampling distribution of the means is indeed normally distributed, consistent with the Central Limit Theorem.

Section 2: Basic Inferential Data Analysis

Using the ToothGrowth data from the R datasets package, we will first perform some exploratory data analysis to get a feel for the data set and then provide a basic summary. We'll then compare tooth growth by supp and dose, using confidence intervals and hypothesis testing. Based on this analysis, we'll show that Vitamin C results in increased tooth growth, regardless of whether it comes from orange juice or ascorbic acid, at least to 2.0mg dosages.

Exploratory Data Analysis

The first step is to load the packages we'll need, along with the dataset, and get a sense of the structure of ToothGrowth. R's documentation states that the ToothGrowth dataset is the effect of Vitamin C on tooth growth in guinea pigs. Ten guinea pigs were each given three dose levels of Vitamin C (0.5, 1 and 2 mg) with each of two delivery methods (orange juice and ascorbic acid), and their tooth growth measured after each test. Therefore, we'd expect to see 60 observations in the dataset.

```
ensurePkg("dplyr")
data("ToothGrowth")
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 ...
```

We see that there are 60 observations of three variables: * len, which is the length of observed tooth growth * supp, a variable with two values: "OJ" and "VC", which is the supplement type * dose, which is the dose in milligrams of orange juice or vitamin C given to each subject

To get a sense of the range of values for `len`, we run a quick summary of the variable:

```
summary(ToothGrowth$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   13.08   19.25   18.81   25.28   33.90
```

Taking a look at a few of the rows, it appears that there might be some correlation between `dose` and `len`:

```
head(ToothGrowth)
```

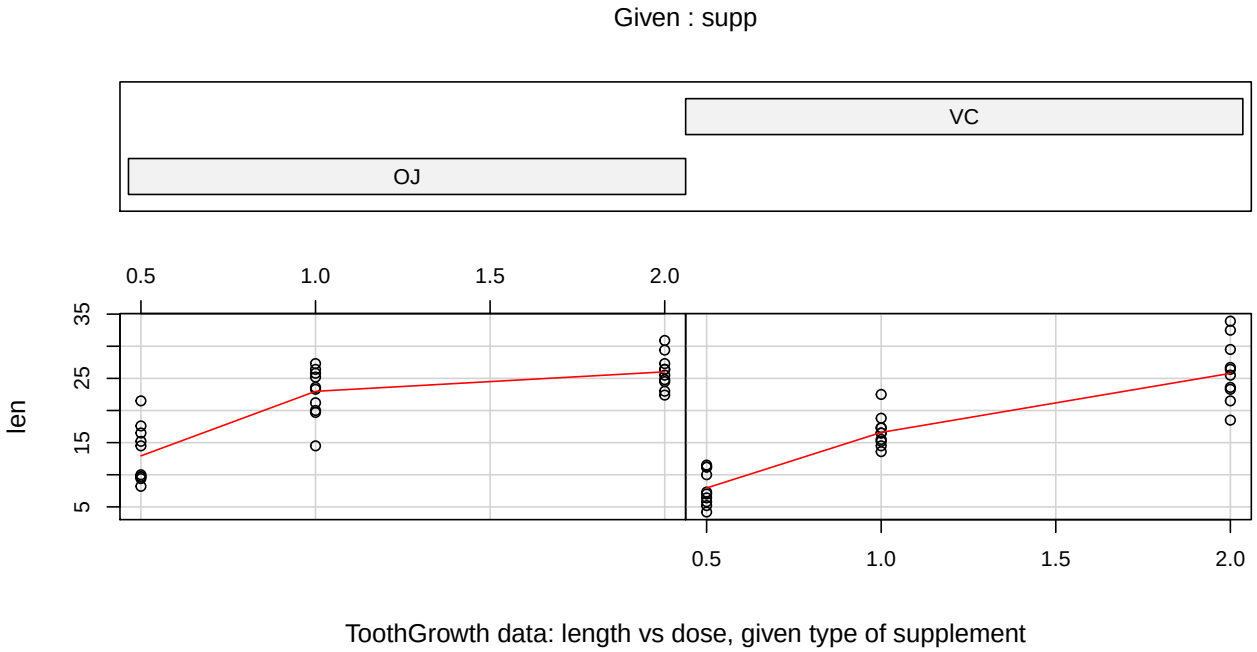
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
tail(ToothGrowth)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

Plotting the values for each combination of supplement and dose, it is at least visually suggestive that there is a correlation between greater doses of Vitamin C and tooth growth:

```
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
       xlab = "ToothGrowth data: length vs dose, given type of supplement")
```



The conditioning plot above shows each tooth growth length observation for each combination of dose, broken out by type of supplement. It illustrates the increase in growth length under higher Vitamin C doses for both supplements, with OJ appearing to do better at lower doses but VC closing the gap at 2.0mg with a similar mean (albeit greater variance).

To confirm the correlation we see in the conditioning plot, we'll calculate the correlation between dose and tooth growth. Since there is a difference in results between OJ and VC, we first group by `supp` and `dose` and calculate the means.

```
grouped_results <- ToothGrowth %>% group_by(supp, dose) %>% summarize(len.mean = mean(len))
print(grouped_results)
```

```
## Source: local data frame [6 x 3]
## Groups: supp
##
##   supp dose len.mean
## 1   OJ  0.5    13.23
## 2   OJ  1.0    22.70
## 3   OJ  2.0    26.06
## 4   VC  0.5     7.98
## 5   VC  1.0    16.77
## 6   VC  2.0    26.14
```

And then calculate the correlation between dose and mean length for each supplement:

```
grouped_cor_OJ <- cor(grouped_results[grouped_results$supp == "OJ", ]$dose, grouped_results[grouped_resu
grouped_cor_VC <- cor(grouped_results[grouped_results$supp == "VC", ]$dose, grouped_results[grouped_resu
```

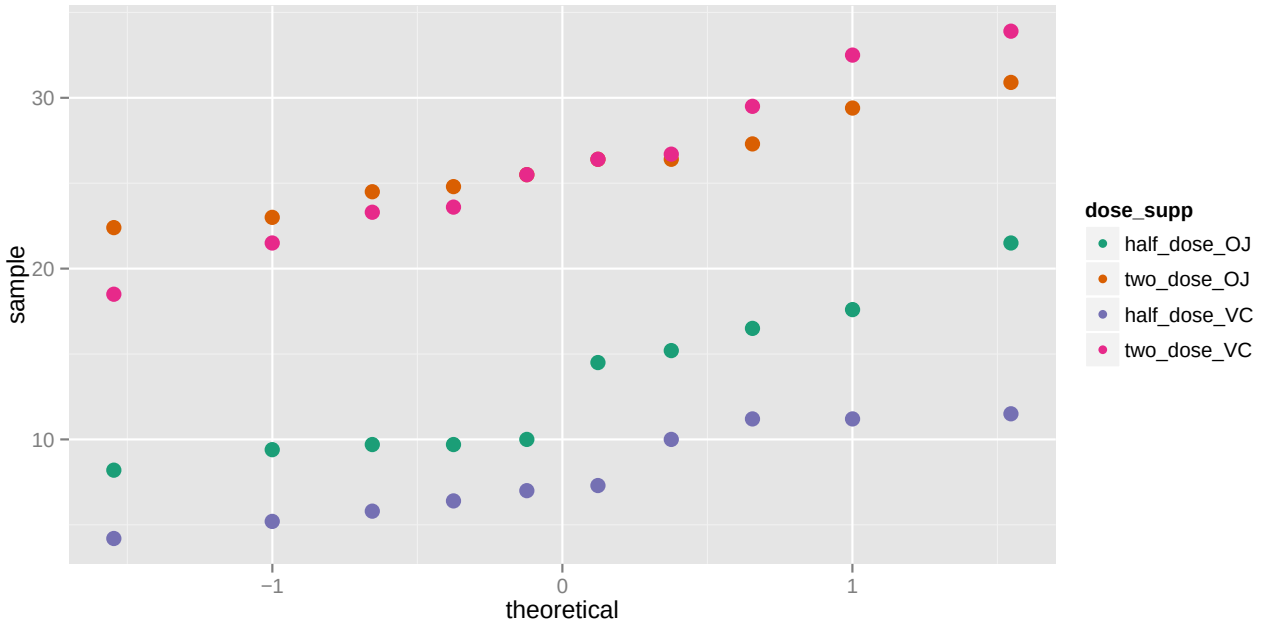
The means of tooth growth are very highly correlated with Vitamin C supplements, regardless of the source of the C (correlation for OJ: 0.8967416; correlation for VC: 0.9852978. But how can we be sure that this is statistically valid? In the next section, we will answer this question.

We also want to see whether or not the data are normally distributed, since this will affect which test statistic to use in our analysis. Given that each combination of `supp` and `dose` is unique, we cannot look at the entire `len` column so we'll break it down.

Let's subset by 0.5mg and 2.0mg dosages for each supplement and then create a quantile-quantile (Q-Q) plot to test for normality. The Q-Q plot compares each point in the dataset to where they would be in a perfectly normal distribution with the same mean and standard deviation. In this case, we'll plot the 2.0mg and 0.5mg doses for each supplement.

```
half_dose_OJ <- ToothGrowth[ToothGrowth$dose == 0.5 & ToothGrowth$supp == "OJ", ]$len
two_dose_OJ <- ToothGrowth[ToothGrowth$dose == 2 & ToothGrowth$supp == "OJ", ]$len
half_dose_VC <- ToothGrowth[ToothGrowth$dose == 0.5 & ToothGrowth$supp == "VC", ]$len
two_dose_VC <- ToothGrowth[ToothGrowth$dose == 2 & ToothGrowth$supp == "VC", ]$len
qq_df <- data.frame(cbind(half_dose_OJ, two_dose_OJ, half_dose_VC, two_dose_VC))
qq_df <- gather(qq_df, "dose_supp", "len")

p <- qplot(sample=len, data=qq_df, color=dose_supp, size=4)
p + scale_color_brewer(palette="Dark2") + guides(size=FALSE)
```



In a Q-Q plot, we look for general linearity as indicative of normality, which we roughly have with all four combinations of dosage and supplement. We can check specifically with a Shapiro-Wilk normality test, looking for a P-value greater than our chosen α of 0.5, which would lead us to accept the null hypothesis that the data are distributed normally.

```
shapiro.test(half_dose_OJ)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  half_dose_OJ
## W = 0.8927, p-value = 0.182
```

```
shapiro.test(two_dose_OJ)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  two_dose_OJ
## W = 0.9626, p-value = 0.8148
```

```
shapiro.test(half_dose_VC)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  half_dose_VC
## W = 0.89, p-value = 0.1696
```

```
shapiro.test(two_dose_VC)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  two_dose_VC
## W = 0.9733, p-value = 0.9194
```

Given approximate linearity of the Q-Q plots and the P-values > $α$, we are comfortable stating that the data are approximately normally distributed.

We next look at the doses that were administered and confirm that they were given equally in the form of both orange juice and ascorbic acid. This gives us confidence that we do not have any missing observations.

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

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

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

The final critical thing to notice about the dataset is that there is **no unique identifier for each guinea pig**. This means that although the same animals were used for each dose and supplement, we have no way to match them up and therefore no way to identify which results are paired.

Summary of the Data

Before moving on to comparing tooth growth results by dose and source of ascorbic acid, let's summarize what we know:

- There are 60 observations in total, with six per guinea pig
- Each guinea pig was tested with three doses each of orange juice and ascorbic acid
- There is a wide range of tooth growth, so presumably some effect is taking place
- There appears to be a correlation between the length of tooth growth and the dose
- The measurements of tooth growth do not appear to adhere strictly to any particular distribution, although the sample size is not large
- The samples are paired, with equal numbers and types of treatments given to each subject
- However, there is no unique identifier for each subject so we cannot pair the results; nothing would indicate, for example, that we can assume the guinea pig with the smallest length of growth under 0.5mg is the same guinea pig that showed the smallest growth when given 2.0mg.

Comparison of Tooth Growth Results by 'supp' and 'dose'

As discussed in the previous section, the means of tooth growth do increase in response to greater dosages of Vitamin C, regardless of the source of the C, but we want to confirm this statistically with a reasonable level of confidence.

Also discussed in the previous section, while the samples are paired, **we cannot identify which guinea pig is which so we cannot accurately pair them**. We therefore will not be able to use the paired t-test. If we had identifiers for which `len` result belonged to which guinea pig, we could have.

We could group both OJ and VC and compare tooth growth for the guinea pigs, but the means are quite different between the two supplements.

```
print(grouped_results)
```

```
## Source: local data frame [6 x 3]
## Groups: supp
##
##   supp dose len.mean
## 1   OJ  0.5    13.23
## 2   OJ  1.0    22.70
## 3   OJ  2.0    26.06
## 4   VC  0.5     7.98
## 5   VC  1.0    16.77
## 6   VC  2.0    26.14
```

To remove this issue, we will treat the OJ and VC treatments as separate experiments on the same subjects and analyze them separately. We will obtain the t statistic, confidence intervals and confidence intervals first formulaically for the OJ supplement and then directly using the R function `t.test()` for the VC.

If Vitamin C does increase tooth growth, we would expect greater dosages to yield greater growth (presumably up to some limit, but testing that limit is beyond the data provided). Therefore, for one or both supplements, mean tooth growth from dosages of 2.0mg should be greater than mean tooth growth from 0.5mg dosages. But if Vitamin C supplements do not result in increased tooth growth, the means should not differ significantly.

Our null hypothesis is therefore:

$$H_0 : \mu_2 = \mu_{0.5} \quad \text{or, equivalently,} \quad H_0 : \mu_2 - \mu_{0.5} = 0$$

If Vitamin C does enhance tooth growth, we would expect to see increased growth under the 2.0mg dose, so our alternative hypothesis is:

$$H_a : \mu_2 > \mu_{0.5} \quad \text{or, equivalently,} \quad H_a : \mu_2 - \mu_{0.5} > 0$$

To determine whether or not to reject the null hypothesis above, we'll set the level of significance (α) to 0.05.

With a sample size of only 10, we will use a two-sample t-test and as noted before, given the data, **we will not use a paired t-test**. Our first question is, are the variances equal or not? We can run a quick F test for equality of variances and look at the P-Value. The null hypothesis of the F test is that the variances are equal, so if the resulting P-Value is greater than our α of 0.05, we fail to reject this null hypothesis.

```
f_test_p <-var.test(half_dose_OJ, two_dose_OJ)
```

The resulting P-Value of 0.14 means that we can move forward with the two-sample t-test with equal variances.

The formulas we're using to calculate the test statistic are:

$$t = \frac{\bar{X}_2 - \bar{X}_{0.5}}{s_p\sqrt{\frac{1}{n_2} + \frac{1}{n_{0.5}}}}$$

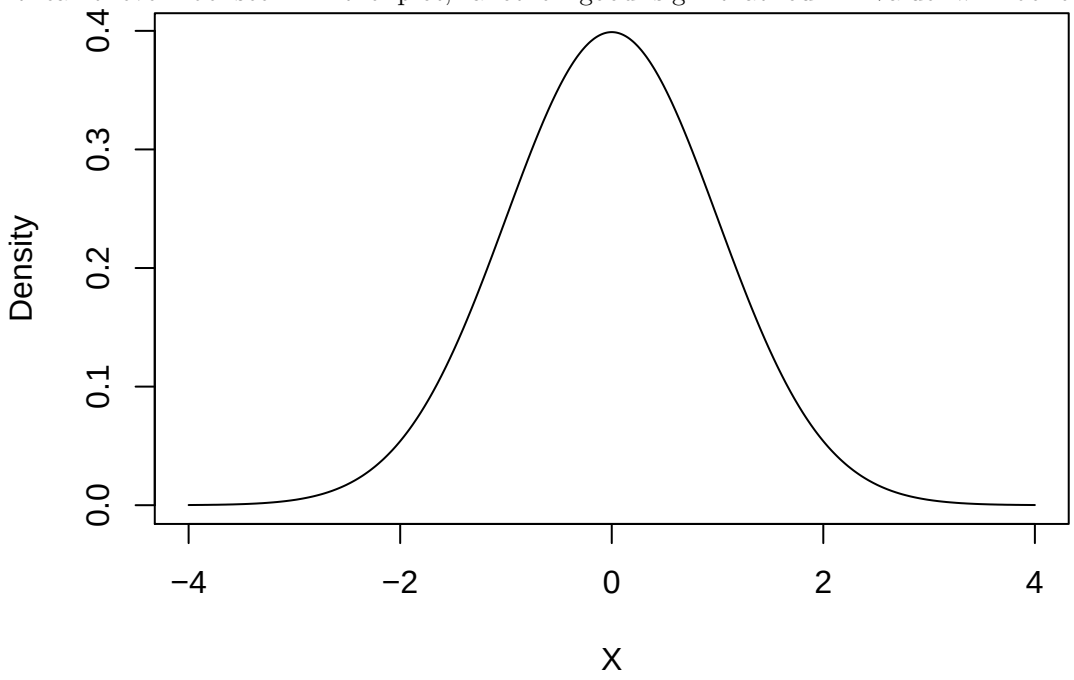
$$s_p = \sqrt{\frac{(n_2 - 1)s_2^2 + (n_{0.5} - 1)s_{0.5}^2}{n_2 + n_{0.5} - 2}}$$

Since we are not using the paired test, the sample size is 20. Plugging in the needed values into the above formula yields:

```r
n <- 20
x_two <- mean(two_dose_OJ)
x_half <- mean(half_dose_OJ)
s_two <- sd(two_dose_OJ)
s_half <- sd(half_dose_OJ)

sp <- sqrt( ((n-1)*s_two^2 + (n-1)*s_half^2)/(2*n-2) )
t <- (x_two-x_half)/(sp*sqrt(1/n + 1/n))
```

This results in a t statistic of 11.05, which is quite large and signals that we will have a very small P-Value. Since t distributions have the shape of a normal distribution, we can shade the portion of a normal curve that corresponds to our computed test statistic, but since it is so large, it can't even be seen in the plot, another good sign that our P-Value will be extremely small.



Even so, we calculate our P-Value as:

```r
t_test_OJ <- t.test(two_dose_OJ, half_dose_OJ, alternative = "greater", var.equal = TRUE)
t_test_OJ
```

```
##
##  Two Sample t-test
##
## data:  two_dose_OJ and half_dose_OJ
## t = 7.817, df = 18, p-value = 0.0000001701
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  9.983898      Inf
## sample estimates:
## mean of x mean of y
##     26.06     13.23
```

Our P-Value of $0.0000002 < \alpha$, so we reject the null hypothesis, in favor of the alternative hypothesis: there is evidence to suggest that Vitamin C when given by orange juice results in increased tooth growth.

To double-check our work, we look at the confidence intervals, using the formula:

$$\bar{X}_2 - \bar{X}_{0.5} \pm t_{df}\, s_p \sqrt{\frac{1}{n_2} + \frac{1}{n_{0.5}}}$$

But we can pull the 95% confidence interval from the `t.test()` results:

```
t_test_OJ_CI <- t.test(two_dose_OJ, half_dose_OJ, var.equal = TRUE)
```

Which results in a lower bound of 9.38 and an upper bound 16.28, which comfortably include the difference between our two sample means of 12.83, so the difference in our sample means is solidly within the range of our 95% confidence interval. Further, the entire interval is well above zero, which, under H_a, means that the difference between our two sample means is positive and therefore the means are not equivalent.

Now we will run the same tests for the ascorbic acid supplement, but using R functions directly.

```
t_test_VC <- t.test(two_dose_VC, half_dose_VC, alternative = "greater", var.equal = TRUE)
t_test_VC
```

```
##
##  Two Sample t-test
##
## data:  two_dose_VC and half_dose_VC
## t = 10.3878, df = 18, p-value = 0.000000002479
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  15.1285     Inf
## sample estimates:
## mean of x mean of y
##     26.14      7.98
```

Similar to OJ, the t statistic for VC is quite high (10.39), which would lead to a very low P-Value, which we see in the results: 0. This is well below our α of 0.5 so we reject the null hypothesis and conclude that there is evidence that Vitamin C given by ascorbic acid results in increased tooth growth.

```
t_test_VC_CI <- t.test(two_dose_VC, half_dose_VC, var.equal = TRUE)
```

Calculating the 95% confidence intervals for the VC supp yields [14.49, 21.83], which again comfortably includes our mean and is well above zero.

SHORT ANALYSIS IF THERE IS 2X THE TOOTH GROWTH FROM 1MG TO 2MG AS THERE IS FROM .5MG TO 1MG

Conclusion

Given the limitations of the data, we had to make a number of assumptions:

- From the description of the dataset, the 10 guinea pigs comprised the entire study, so while this is not technically a random sample, it is obviously a valid representation of the population
- The observations are independent, particularly that the supplement and/or dosage used in previous trials didn't affect the tooth growth in subsequent trials
- The data are normally distributed

- Since we could not identify which `len` result belonged to which subject, we had to treat the data as not paired
- The variances are approximately equal, which we checked with an F test

Even with these assumptions, the results show with high levels of confidence that Vitamin C in the form of orange juice and ascorbic acid does promote tooth growth.