Inferential Statistics Course Project: Inferential Analysis

Wally Thornton

August 23, 2015

Project 2: An 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 supplement type 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 up 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 (code loading packages not shown, to save space).

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, 1.0 and 2.0 mg) with each of two delivery methods (orange juice and ascorbic acid), and their tooth growth measured after each test. Therefore, we expect to see 60 observations in the dataset.

```r
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 are the two supplement types
- dose, which is the dose in milligrams of orange juice or ascorbic acid given to each subject

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

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

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

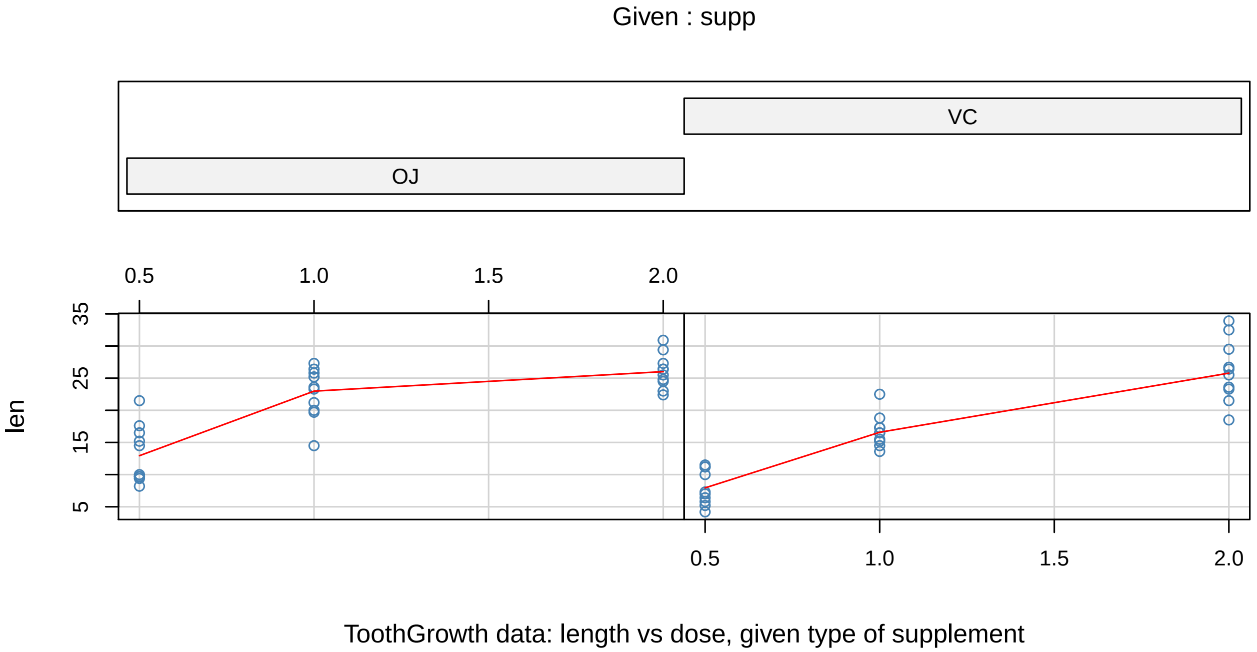
Taking a look at a few of the rows, it appears that there might be some correlation between `dose` and `len`:

```
ToothGrowth[c(1:3,58:60), ]
```

```
##     len supp dose
## 1   4.2   VC  0.5
## 2  11.5   VC  0.5
## 3   7.3   VC  0.5
## 58 27.3   OJ  2.0
## 59 29.4   OJ  2.0
## 60 23.0   OJ  2.0
```

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, col="steelblue",
       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 each supplement, both of which show a correlation between increased Vitamin C doses and growth length.

We'll next calculate this 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))
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:

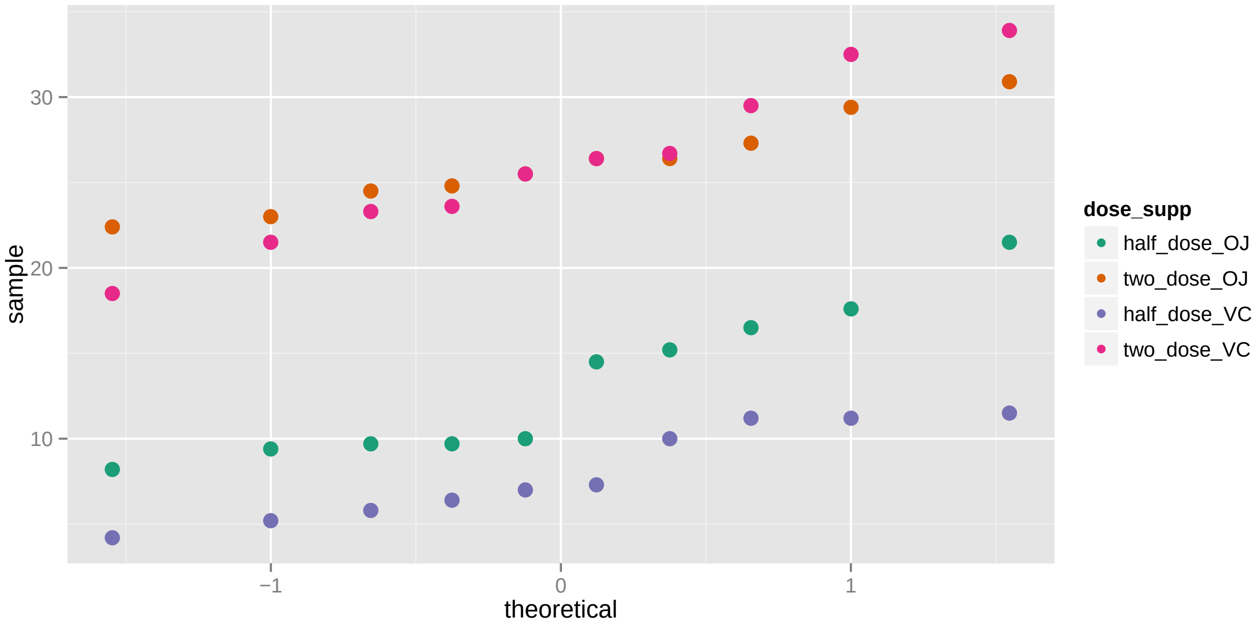
```
grouped_cor_OJ <- cor(grouped_results[grouped_results$supp == "OJ", ]$dose,
                      grouped_results[grouped_results$supp == "OJ", ]$len.mean)
grouped_cor_VC <- cor(grouped_results[grouped_results$supp == "VC", ]$dose,
                      grouped_results[grouped_results$supp == "VC", ]$len.mean)
```

The means of tooth growth are very highly correlated with Vitamin C supplements, regardless of the source of the C (correlation for OJ: 0.897; correlation for VC: 0.985). We will examine the statistical validity of these correlations in the analysis section.

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 subset the data.

We 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 0.05, which would lead us to accept the null hypothesis that the data are distributed normally. (Showing just VC tests to conserve space.)

```
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 the approximate linearity of the Q-Q plots and that all P-values > α, we are comfortable stating that the data are approximately normally distributed. This is an important assumption, as we will see.

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.

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

```
##     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 Exploratory Data Analysis

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
- Each combination of `supp` and `dose` is distributed approximately normally
- 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'

We saw that 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.

As already mentioned, 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.

The means for each supplement are quite different so we will treat OJ and VC treatments as separate experiments and analyze them separately. 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, there should be little to no difference between the means.

Our null hypothesis is therefore:

$$H_0 : \mu_{2.0} = \mu_{0.5} \quad \text{or, equivalently,} \quad H_0 : \mu_{2.0} - \mu_{0.5} = 0$$

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

$$H_a : \mu_{2.0} > \mu_{0.5} \quad \text{or, equivalently,} \quad H_a : \mu_{2.0} - \mu_{0.5} > 0$$

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 cannot use a paired t-test**. To test for equality of variances, we run an F test 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 and treat the variances as equal.

```
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:

$$t = \frac{\bar{X}_{2.0} - \bar{X}_{0.5}}{s_p \sqrt{\frac{1}{n_{2.0}} + \frac{1}{n_{0.5}}}}$$

$$s_p = \sqrt{\frac{(n_{2.0} - 1)s_{2.0}^2 + (n_{0.5} - 1)s_{0.5}^2}{n_{2.0} + n_{0.5} - 2}}$$

Plugging in the needed values into the above formula yields:

```
n <- 10 # both 2.0 and 0.5 sample sizes are the same
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 7.82, which is quite large and signals that we will have a very small P-Value. To confirm, we calculate our P-Value as:

```
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:

$$\bar{X}_{2.0} - \bar{X}_{0.5} \pm t_{df} s_p \sqrt{\frac{1}{n_{2.0}} + \frac{1}{n_{0.5}}}$$

And 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 above. This is well below our α of 0.05 so we reject the null hypothesis and conclude that there is evidence that Vitamin C given by ascorbic acid results in increased tooth growth.

```r
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 the difference between the means and is well above zero.

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 this particular population
- The observations are independent, in general
- In particular, 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

Given these assumptions, the results show with high levels of confidence that Vitamin C in the form of both orange juice and ascorbic acid did promote tooth growth in these guinea pigs.