ToothGrowth Data Analysis

Sean Dunagan

September 27, 2015

Background of ToothGrowth Dataset

The ToothGrowth data set contains data regarding “The Effect of Vitamin C on Tooth Growth in Guinea Pigs”. The two delivery methods (referred to as suppliers in my codeset and this report) were Orange Juice (hereby referred to as OJ) and Absorbic Acid (Vitamin C) and hereby referred to as VC. Three levels of dosage we used in the experiment: 0.5, 1, and 2 mg. The experiment measured the length of odontoblasts (teeth) in each of 10 guinea pigs at each dosage level and supplier method. The above information retrieved from the inside-R ToothGrowth (http://www.inside-r.org/r-doc/datasets/ToothGrowth) page

Basic Exploratory Data Analyses

My R code below displays output given by the functions defined in the toothgrowth_analysis.R file I wrote to help with the project. File toothgrowth_analysis.R assumes that the user has loaded the datasets R library

A brief inspection of the ToothGrowth dataset reveals the following regarding the data:

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

```
## [1] VC OJ
## Levels: OJ VC
```

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

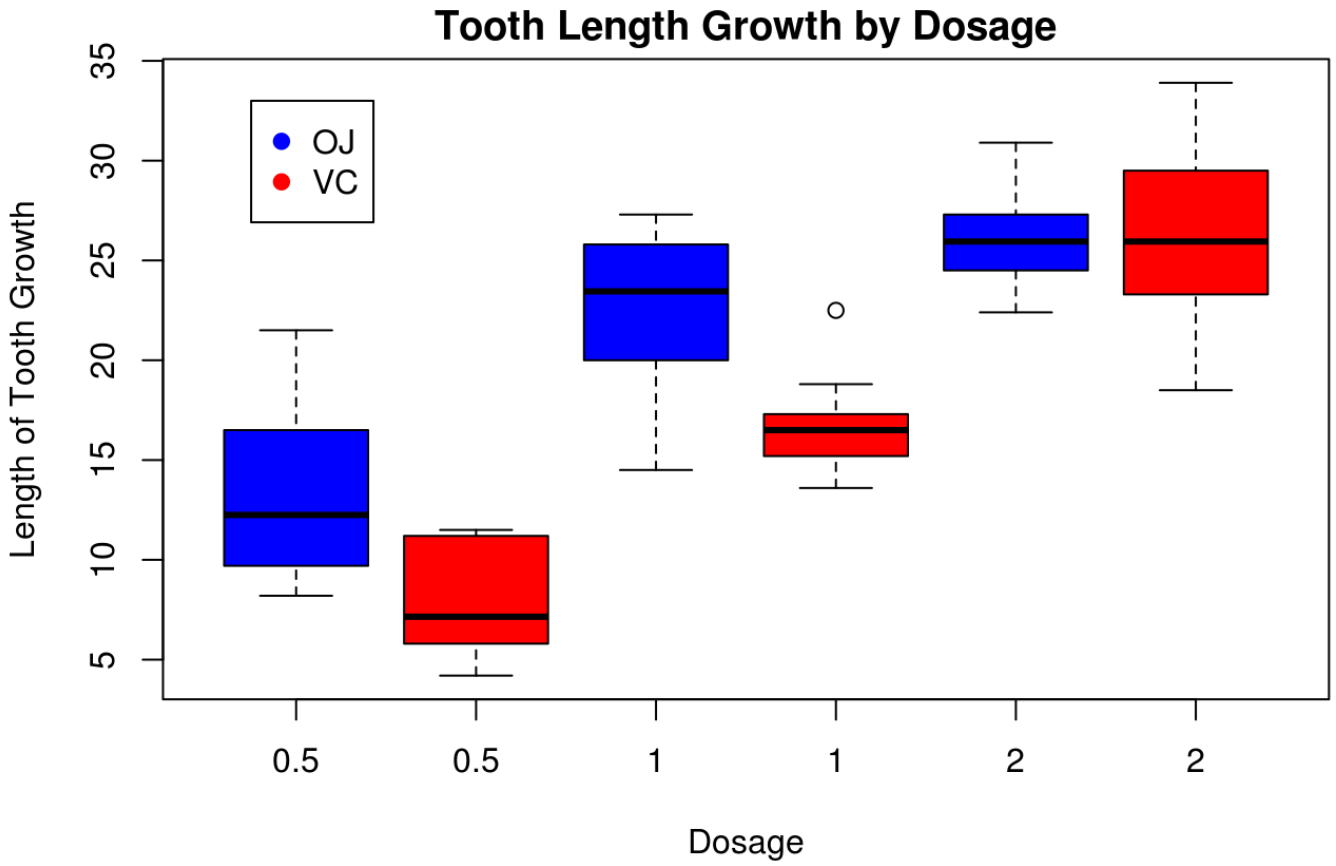
We naturally see 6 distinct subsets in this dataset, split between the suppliers and the dosages. The means and standard deviations of the 6 subsets are given below:

```
##    0.5   1     2
## OJ 13.23 22.7  26.06
## VC 7.98  16.77 26.14
```

```
##    0.5      1        2
## OJ 4.459709 3.910953 2.655058
## VC 2.746634 2.515309 4.797731
```

We can construct a boxplot of the data across these 6 subsets to get a quick visual representation of the distribution of the odontoblasts lengths

Figure 1.c



We can clearly see that higher dosage levels result in greater tooth length based on the result of getToothGrowthSubsetMeansMatrix(). However the box plot provides the best summary of the data in the ToothGrowth dataset. It displays the mean values of each of the 6 subsets and provides a visual representation of their distribution and variance. We can easily see that at dosages of 0.5 and 1 mg, OJ resulted in significantly higher tooth lengths. At 2 mgs however, the means of the lengths are very close, with supplier VC having a significantly larger variance. Naturally, our next step with this dataset would be to compare the difference between supplier tooth lengths at each dosage level.

Means and Confidence Intervals of difference between suppliers by dosage

As this dataset has a relatively small number of samples (10 in each subset) it makes sense to use the T distribution to calculate our confidence intervals regarding the difference in tooth length between supplier methods. We can easily see from the standard deviation matrix above that the variances among the delivery methods are clearly unequal at every dosage level; as such our model should not treat the groups as having a constant variance. Since different guinea pigs were treated with each supplier (as opposed to treating 10 pigs with one supplier, waiting a “clean-out” period of time and then treating the same 10 pigs with the other supplier) these groups are not paired.

We can use the R t.test() function to compute these confidence intervals. We will use the default confidence level of 95%. In all of the intervals, we measured the difference by calculating the VC values minus the OJ values (VC - OJ).

```
##      Mean Interval Low Interval High Attained Significance Level
## 0.5 -5.25    -9.374912     -1.125088                   0.9968207
## 1   -5.93    -9.581900     -2.278100                   0.9994808
## 2    0.08    -4.270131      4.430131                   0.4819258
```

As we see in the above, supplier OJ results in longer tooth lengths than supplier VC by a mean length of 5.25 and 5.93 at dosage levels 0.5mg and 1mg respectively. However, once we get to a dosage of 2mg, supplier VC actually has a mean tooth length slightly greater than value OJ (.08mg).

Conclusions based on the T-confidence intervals

We used a 95% confidence interval in our calculations. For dosages 0.5mg and 1mg, the entire 95% interval is negative. As such, we can say with 95% confidence that supplier OJ results in longer tooth lengths than supplier VC. Additionally, we can say with 95% confidence (based on this data) that supplier OJ will result in longer tooth lengths than supplier VC by at least 1.125 and 2.278 at dosage levels 0.5mg and 1mg respectively. (I could not find any units of measurement for the length online). Additionally, using the test statistic of (VC_mean - OJ_mean) / standard_error_esimate, we have calculated the attained level of significance for these intervals being 99.68% and 99.95% respectively. This means that we can say with 99.68% and 99.95% certinainty (given the data in the set) that the mean tooth length with OJ is longer than the mean tooth length of VC

However once we get to the dosage level of 2mg, the two supplier methods are very similar. Supplier VC actually has a larger mean tooth length by .08. The confidence interval at this dosage is -4.27,4.43. In English, this means that we can say with 95% certainty (based on our data) that supplier VC will result in tooth lengths that are less than -4.27 shorter and less than 4.43 longer than supplier OJ. As a result, we can not say with 95% confidence which supplier results in longer tooth lengths at the 2mg dosage level. In fact, the attained level of significance of (VC_mean - OJ_mean) at 2mg dosage is 48%, meaning we are less certain of (VC_mean < OJ_mean) than we are of (OJ_mean < VC_mean). This means that if our null hypothesis was OJ mean greater than VC mean, we would reject the null hypothesis in this scenario at any confidence level.

Assumptions required for these Conclusions

The assumption used with confidence intervals (and the Central Limit Theorem in general) is that the data you are working with consists of iid (independent and identically distributed) variables. In this experiment, we have 60 guinea pigs. Each of these guinea pigs can be considered an independent variable since they each are their own distinct test subject. In order to be considered indentically distributed, each guinea pig should have the same probability distribution as every other guinea pig in the test set. This means that we assume that all of the guinea pigs were of the same breed of guinea pig in order to remove breed as a confounding factor. In order to remove gender as a confouding factor, we would either select all 60 guinea pigs as the same gender, or we could assign each of the 6 subsets to have 5 males and 5 females each. We could also seek to remove size as a confounding factor by choosing guinea pigs all of about the same

length and weight; an alternative to this would be to normalize the length data by measuring the percent change in tooth growth (I was unable to find any units for the length measurement, so theoretically these lengths could be measured in perecent growth of the teeth).

In regards to working with the T distribution, one is assuming that the data is also Gaussian with the result that ((Estimate - Mean_of_Estimate) / (Standard_Error_of_esimate)) follows a t distrbiution with n-1 degrees of freedom.

R Code used for the assignment

```
splitToothGrowthIntoSubsets <- function()
{
    subsets <- split(ToothGrowth, list(ToothGrowth$supp, ToothGrowth$dose));
    subsets;
}

splitToothGrowthBySupplier <- function()
{
    subsets <- split(ToothGrowth, list(ToothGrowth$supp));
    subsets;
}

getToothGrowthSubsetSds <- function()
{
    subsets <- splitToothGrowthIntoSubsets();
    sds_of_subsets <- lapply(subsets, function(subset)
    {
        subset_tooth_lengths <- subset[['len']];
        subset_tooth_lengths_sd <- sd(subset_tooth_lengths);
        subset_tooth_lengths_sd
    });
}

getToothGrowthSubsetMeans <- function()
{
    subsets <- splitToothGrowthIntoSubsets();
    means_of_subsets <- lapply(subsets, function(subset)
    {
        subset_tooth_lengths <- subset[['len']];
        subset_tooth_lengths_mean <- mean(subset_tooth_lengths);
        subset_tooth_lengths_mean
    });
}

getToothGrowthSubsetSdsMatrix <- function()
{
    row_names <- c('OJ', 'VC');
    col_names <- c(0.5, 1, 2);
    dim_names <- list(row_names, col_names);
    subset_sds <- getToothGrowthSubsetSds();
    subset_sds_matrix <- matrix(subset_sds, nrow = 2, ncol = 3, dimnames = dim_nam
es);
    subset_sds_matrix;
}

getToothGrowthSubsetMeansMatrix <- function()
{
    row_names <- c('OJ', 'VC');
    col_names <- c(0.5, 1, 2);
    dim_names <- list(row_names, col_names);
```

```
    subset_means <- getToothGrowthSubsetMeans();
    subset_means_matrix <- matrix(subset_means, nrow = 2, ncol = 3, dimnames = di
m_names);
    subset_means_matrix;
}

getComparisonConfidenceInterval <- function(subtracted_from_lengths, subtracting_l
engths)
{
    confidence_interval <- t.test(x = subtracted_from_lengths, y = subtracting_len
gths,
                                  paired=FALSE, var.equal=FALSE, conf.level=.97
5);
    confidence_interval;
}

getSubsetIndexBySupplierAndDosage <- function(supplier, dosage)
{
    paste(supplier, dosage, sep='.');
}

getSupplierDifferenceTTestByDosage <- function(dosage)
{
    subsets <- splitToothGrowthIntoSubsets();
    vc_index <- getSubsetIndexBySupplierAndDosage('VC', dosage);
    vc_data <- subsets[[vc_index]];

    oj_index <- getSubsetIndexBySupplierAndDosage('OJ', dosage);
    oj_data <- subsets[[oj_index]];

    tTestResult <- getComparisonConfidenceInterval(vc_data$len, oj_data$len);
    tTestResult;
}

getSupplierDifferenceOfMeansByDosage <- function(dosage)
{
    subset_means <- getToothGrowthSubsetMeans();
    vc_index <- getSubsetIndexBySupplierAndDosage('VC', dosage);
    vc_mean <- subset_means[[vc_index]];

    oj_index <- getSubsetIndexBySupplierAndDosage('OJ', dosage);
    oj_mean <- subset_means[[oj_index]];

    difference_in_means <- vc_mean - oj_mean;
    difference_in_means
}

getSupplerDifferenceMeanAndConfidenceIntervalForDosage <- function(dosage)
{
    mean <- getSupplierDifferenceOfMeansByDosage(dosage);
```

```
    tTestResult <- getSupplierDifferenceTTestByDosage(dosage);
    interval_low_end <- tTestResult$conf.int[1];
    interval_high_end <- tTestResult$conf.int[2];

    degrees_of_freedom <- tTestResult$parameter;

    subset_means <- getToothGrowthSubsetMeans();
    vc_index <- getSubsetIndexBySupplierAndDosage('VC', dosage);
    vc_mean <- subset_means[[vc_index]];

    oj_index <- getSubsetIndexBySupplierAndDosage('OJ', dosage);
    oj_mean <- subset_means[[oj_index]];

    subset_sds <- getToothGrowthSubsetSds();
    vc_sd <- subset_sds[[vc_index]];
    oj_sd <- subset_sds[[oj_index]];

    attained_significance_level <- getPercentCertaintyOfHypothesis(degrees_of_free
dom, oj_mean, vc_mean, oj_sd, vc_sd, 10,10);

    c(mean, interval_low_end, interval_high_end, attained_significance_level);
}

getSupplierDifferencesMatrix <- function()
{
    col_names <- c('Mean', 'Interval Low', 'Interval High', 'Attained Significance
Level');
    dosage <- c(0.5, 1, 2);
    dim_names <- list(dosage, col_names);
    supplier_comparison <- lapply(dosage, getSupplerDifferenceMeanAndConfidenceInt
ervalForDosage);
    supplier_comparison_vector <- c(supplier_comparison[[1]],supplier_compariso
n[[2]],supplier_comparison[[3]]);
    comparisonMatrix <- matrix(supplier_comparison_vector, nrow=3, ncol=4, byrow=T
RUE, dimnames = dim_names);
    comparisonMatrix
}

getVariationOfMean <- function(S, n)
{
    variaton_of_mean <- (S^2) / n;
    invisible(variaton_of_mean);
}
# 95% confidence interval
getPercentCertaintyOfHypothesis <- function(df, OJ_mean, VC_mean, OJ_sd, VC_sd, O
J_n, VC_n)
{
    standard_error_estimate <- sqrt(getVariationOfMean(OJ_sd, OJ_n) + getVariation
OfMean(VC_sd, VC_n));
    difference_in_means <- VC_mean - OJ_mean;
```

```
    test_statistic <- difference_in_means / standard_error_estimate;
    attained_significance_level <- (1 - pt(test_statistic, df));
    attained_significance_level
}

createSupplierComparisonScatterPlot <- function()
{
    tg_by_supplier <- splitToothGrowthBySupplier();
    x_label <- 'Dosage';
    y_label <- 'Length of Tooth Growth';
    main_label <- 'Tooth Length Growth by Dosage';
    plot(tg_by_supplier$OJ$dose, tg_by_supplier$OJ$len, col="blue", xlab = x_labe
l, ylab = y_label, main=main_label, pch=18);
    points(tg_by_supplier$VC$dose, tg_by_supplier$VC$len, col="red", pch=16);

    legend(x = 1.25, y=25, legend=c('OJ', 'VC'), pch=c(18,16), col=c("blue", "re
d"));
}

createSupplierComparisonBoxPlot <- function()
{
    tg_by_supplier <- splitToothGrowthBySupplier();
    x_label <- 'Dosage';
    y_label <- 'Length of Tooth Growth';
    main_label <- "Figure 1.c\n\nTooth Length Growth by Dosage";
    box_names <- c('0.5', '0.5', '1', '1', '2', '2');
    boxplot(len~supp*dose, data=ToothGrowth, notch=FALSE, names=box_names,
            col=(c("blue","red")), main=main_label, xlab=x_label, ylab=y_label);

    legend(x = 0.75, y=33, legend=c('OJ', 'VC'), pch=19, col=c("blue", "red"));
}
```