Basic Inferential Data Analysis on ToothGrowth Dataset

Anthony Liu

Overview

In this report we perform an analysis on the ToothGrowth data set originally from an experiment in 1947. We perform multiple hypothesis tests using R 3.6.1 to determine whether the independent variables of the experiment have an effect on the measured response.

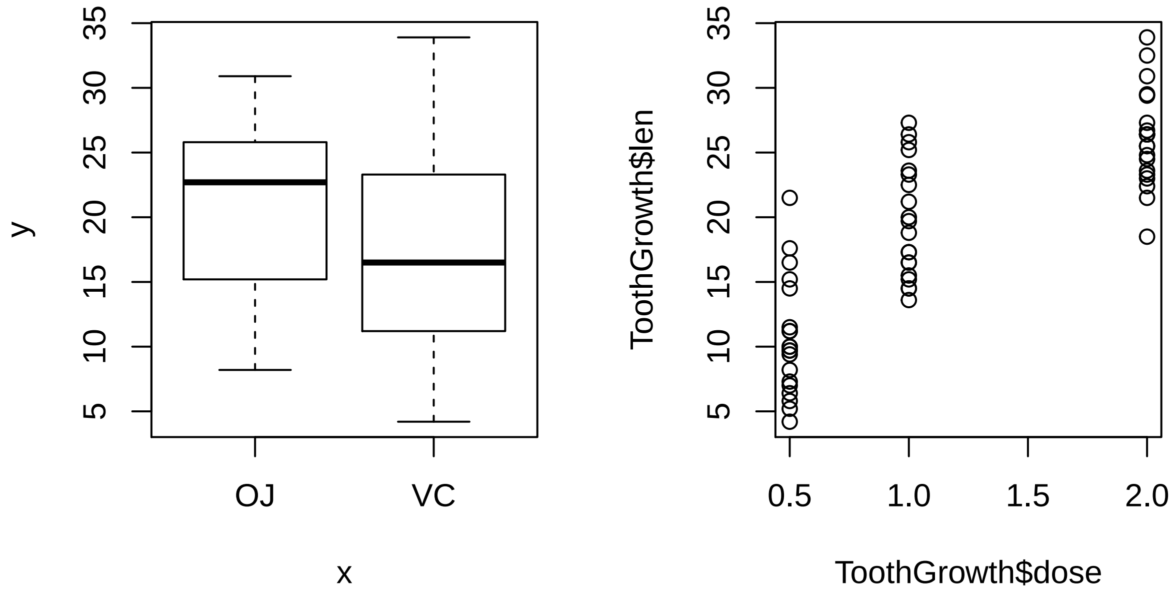
Loading ToothGrowth Data and Exploratory Data Analysis

Here we load the ToothGrowth dataset, create a table comparing the number of trials for each supplement type (supp) and dose level (dose), and make two exploratory plots comparing supplement type and dose level with tooth length.

```
data(ToothGrowth)
table(ToothGrowth$supp, ToothGrowth$dose)
```

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

```
par(mfrow = c(1, 2))
plot(ToothGrowth$supp, ToothGrowth$len)
plot(ToothGrowth$dose, ToothGrowth$len)
```



Basic Summary of Data

According to the documentation of the ToothGrowth dataset, the data here is on "The Effect of Vitamin C on Tooth Growth in Guinea Pigs" where "The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC)."

The data has three variables: tooth length (len, numeric variable), supplement type (supp, factor variable) and dose level (dose, numeric variable but only takes the values 0.5, 1.0 and 2.0). The exploratory table in the previous section shows that the 60 guinea pigs were divided into 6 groups of 10, each with a unique combination of supplement type and dose level. Looking at the exploratory plots there seems to be a positive correlation between dose level and tooth length, but the relationship between supplement type and tooth length is less clear.

Tooth growth by supplement type and dose level

Supplement type

To determine whether supplement type has a significant effect on tooth growth, we will perform a hypothesis test with null hypothesis $H_0 : \mu_{OJ} = \mu_{VC}$ and alternative hypothesis $H_a : \mu_{OJ} > \mu_{VC}$. Before this, we look at the standard deviation of tooth length for each supplement type to get an indication of whether we should use equal or unequal variances for our hypothesis test (see Appendix A.1).

We observe that the standard deviations (and hence variances) are reasonably different so we make the **assumption** of *unequal* variances for our hypothesis test. We also assume that the tooth lengths within each group are independent and identically distributed so that the Central Limit Theorem is applicable. Note that looking at the sample standard deviations of our data is only an *indicator* of whether the true population variances are equal, hence although reasonably motivated unequal variances is still an assumption. We now perform the hypothesis test using the one sided t-test for independent samples with 95% confidence (see Appendix A.2).

As zero lies outside of this confidence interval, we reject the null hypothesis and **conclude** that use of orange juice over ascorbic acid is correlated with greater tooth growth. (As an interesting note, a *two* sided t-test would have yielded a confidence interval where zero was just contained and we would have failed to reject the null hypothesis.)

Dose level

To determine whether dose level has a significant effect on tooth growth, we will perform two hypothesis tests. The first is with null hypothesis $H_0 : \mu_{2.0} = \mu_{1.0}$ and alternative hypothesis $H_a : \mu_{2.0} > \mu_{1.0}$. The second is with null hypothesis $H_0 : \mu_{1.0} = \mu_{0.5}$ and alternative hypothesis $H_a : \mu_{1.0} > \mu_{0.5}$. In this situation there are better methods such as ANOVA and use of the F statistic but we will restrict ourselves here to simple pairwise hypothesis tests. Before this, we look at the standard deviation of tooth length for each dose level to get an indication of whether we should use equal or unequal variances for our hypothesis test (see Appendix A.3).

We observe that the standard deviations (and hence variances) are reasonably similar so we make the **assumption** of *equal* variances for our hypothesis test. We also assume that the tooth lengths within each group are independent and identically distributed so that the Central Limit Theorem is applicable. Note that looking at the sample standard deviations of our data is only an *indicator* of whether the true population variances are equal, hence although reasonably motivated equal variances is still an assumption. We now perform the hypothesis tests using the one sided t-test for independent samples with 95% confidence (see Appendix A.4).

As neither confidence interval contains zero, we reject both null hypotheses and **conclude** that larger dosages of supplements are correlated with greater tooth growth.

Conclusions and assumptions

Supplement type

Assumptions: unequal variances between OJ and VC supplement type groups, tooth lengths are independent and identically distributed so that the Central Limit Theorem is applicable and the means of the groups approximate a normal distribution (more accurately a t-distribution)

Conclusion: use of orange juice over ascorbic acid is correlated with greater tooth growth

Dose level

Assumptions: equal variances between 0.5, 1.0 and 2.0 dose level groups, tooth lengths are independent and identically distributed so that the Central Limit Theorem is applicable and the means of the groups approximate a normal distribution (more accurately a t-distribution)

Conclusion: larger dosages of supplements are correlated with greater tooth growth

Appendix A

A.1 - Standard deviation across groups of supplement type

```r
suppOJ <- subset(ToothGrowth, supp == "OJ")$len
suppVC <- subset(ToothGrowth, supp == "VC")$len
sd(suppOJ)
```

```
## [1] 6.605561
```

```r
sd(suppVC)
```

```
## [1] 8.266029
```

A.2 - Hypothesis test for supplement type

```r
t.test(suppOJ, suppVC, alternative = "greater", paired = FALSE,
    var.equal = FALSE)$conf
```

```
## [1] 0.4682687       Inf
## attr(,"conf.level")
## [1] 0.95
```

A.3 - Standard deviation across groups of dose level

```r
supp0.5 <- subset(ToothGrowth, dose == 0.5)$len
supp1.0 <- subset(ToothGrowth, dose == 1)$len
supp2.0 <- subset(ToothGrowth, dose == 2)$len
sd(supp0.5)
```

```
## [1] 4.499763
```

```r
sd(supp1.0)
```

```
## [1] 4.415436
```

```r
sd(supp2.0)
```

```
## [1] 3.77415
```

A.4 - Hypothesis tests for dose level

```r
t.test(supp2.0, supp1.0, alternative = "greater", paired = FALSE,
    var.equal = FALSE)$conf
```

```
## [1] 4.17387     Inf
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test(supp1.0, supp0.5, alternative = "greater", paired = FALSE,
    var.equal = FALSE)$conf
```

```
## [1] 6.753323      Inf
## attr(,"conf.level")
## [1] 0.95
```