Coursera Statistical Inference Project Part 2

Nicholas Pairolero

Overview

This project explores the dataset stock dataset toothgrowth. T tests are used to check whether two supplements provide the same mean growth at different dosages.

Exploratory Analysis

First load the ToothGrowth data set and check out the variable names

```
library(datasets)
data(ToothGrowth)
names(ToothGrowth)
```

```
## [1] "len"  "supp" "dose"
```

Next, I like to relabel the dataset because the title is too long to type.

```
data <- ToothGrowth
```

The data set gives the tooth length of 10 guinea pigs over two supplement types and three dosage levels. The supplement types are VC and OJ, at the .5, 1 and 2 mg dosage levels. Lets sum across the different supplment types and dosage levels to try and get a feel for which supplement gives the most growth controlling for dosage. First, check to make sure the supp and dose variables are factors.

```
class(data$supp)
```

```
## [1] "factor"
```

```
class(data$dose)
```

```
## [1] "numeric"
```

It turns out dose is not yet a factor so convert it.

```
data$dose <- factor(data$dose)
```

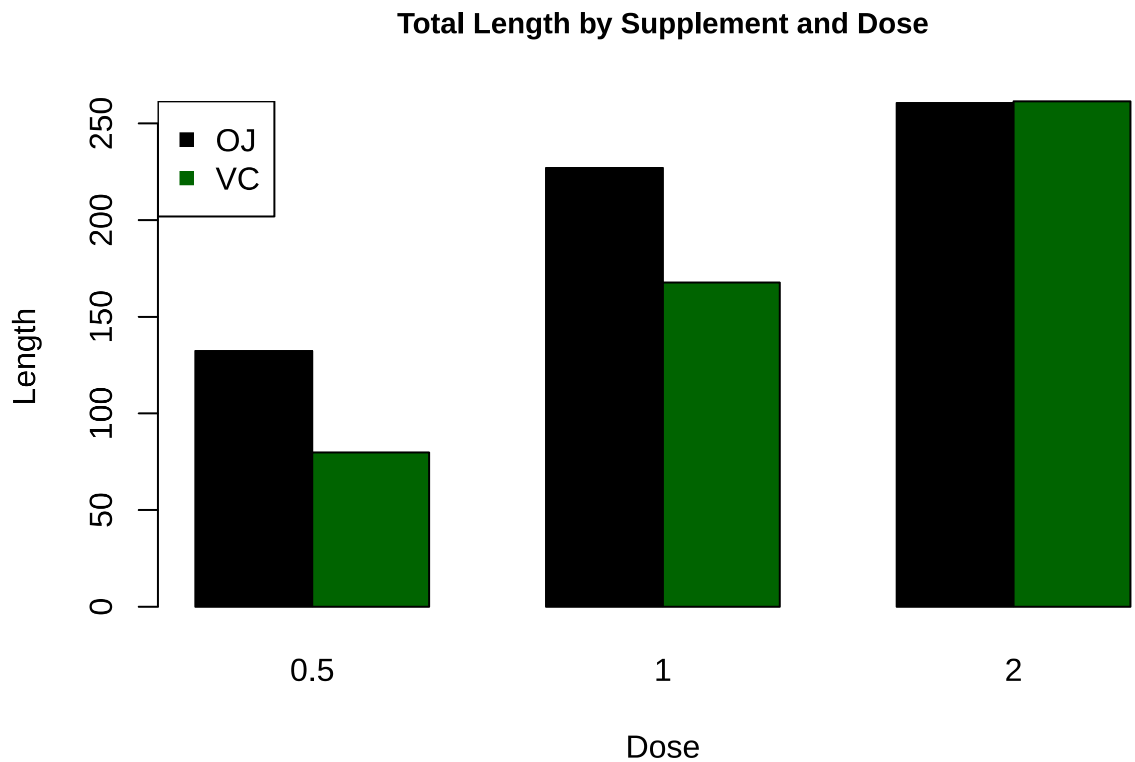
Next, use tapply to sum tooth growth over the factors dose and supp.

```
sums <- tapply(data$len, list(data$supp,data$dose), sum)
sums
```

```
##      0.5     1     2
## OJ 132.3 227.0 260.6
## VC  79.8 167.7 261.4
```

Construct a grouped bar plot to visually compare and contrast the two supplements across dosages.

```r
barplot(sums, main="Total Length by Supplement and Dose", xlab = "Dose", ylab = "Length",
        col = c("black", "darkgreen"), cex.main=.9, beside=TRUE)
legend("topleft", col = c("black", "darkgreen"), pch=15, legend = c("OJ", "VC"))
```



From this plot its relatively clear that OJ correlates with much larger tooth growth than VC at dosages less than 2. Tooth length at a dosage of 2 is relatively even on supplement type.

Hypothesis Testing

Next lets test three hypotheses. For each dosage level, test whether the means vary across supplement type. Run t tests on the means, assuming unequal variances. Its unclear whether the data is paired, so to be safe assume that it is not. The t test is used because the theoretical variance is unknown. First, test across supplements at dosage 0.5

```r
t.test(subset(data, supp == "OJ" & dose == 0.5)$len, subset(data, supp == "VC" & dose == 0.5)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(data, supp == "OJ" & dose == 0.5)$len and subset(data, supp == "VC" & dose == 0.5)$len
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
```

```
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

The p value of .006 is small enough to reject the null hypothesis of equal variances at the 1 percent level. Next, test for equal means across supplements at dosage of 1.

```
t.test(subset(data, supp == "OJ" & dose == 1)$len, subset(data, supp == "VC" & dose == 1)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(data, supp == "OJ" & dose == 1)$len and subset(data, supp == "VC" & dose == 1)$len
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

Again, here the null hypothesis of equal means can be rejected at the 1 percent level. Next, test for dosage of 2

```
t.test(subset(data, supp == "OJ" & dose == 2)$len, subset(data, supp == "VC" & dose == 2)$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(data, supp == "OJ" & dose == 2)$len and subset(data, supp == "VC" & dose == 2)$len
## t = -0.0461, 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
```

This test is very different than the last two. The null hypothesis cannot be rejected. This supports the intuitive idea from the grouped bar plot that OJ is much more favorable for growth at the .5 and 1 dosage levels, but growth is about the same at a dosage of 2.

Results

It can be rejected at the 1 percent level that length means are the same across supplements for dosages of .5 and 1. It cannot be rejected that means are the same across supplements at a dosage of 2 for even the 10 percent level. The p value of this test is around .96. The **assumptions** for the validity of the tests run is that the underlying data is normal or the sample of 10 is large enough for the central limit theorem to hold.