Part 2 Basic Inferential Data Analysis

John B Cheadle

August 10, 2017

Overview

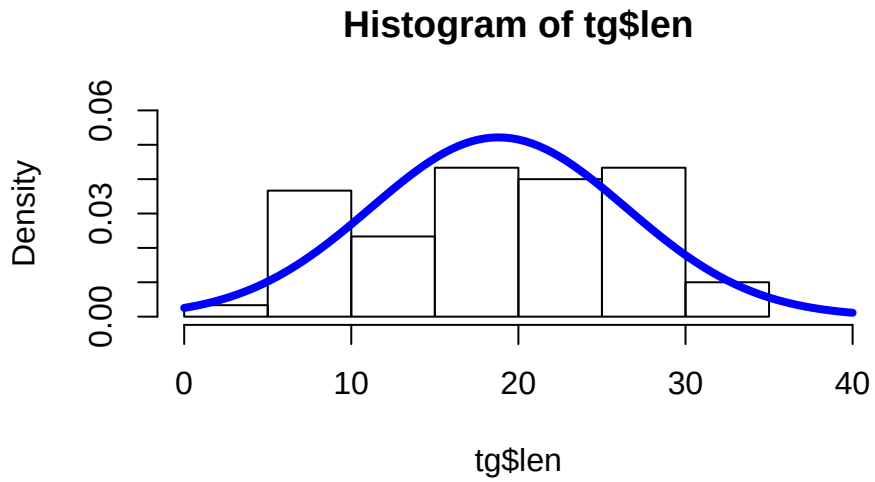
This report describes analysis of the ToothGrowth data in the R datasets package. Here we load and summarize the data using basic exploratory analysis, then answer the question of whether supp and/or dose affect tooth growth length using confidence intervals and hypothesis tests.

ToothGrowth Data

We load the ToothGrowth data frame from the R datasets package and explore some of its characteristics. See Appendix for more information.

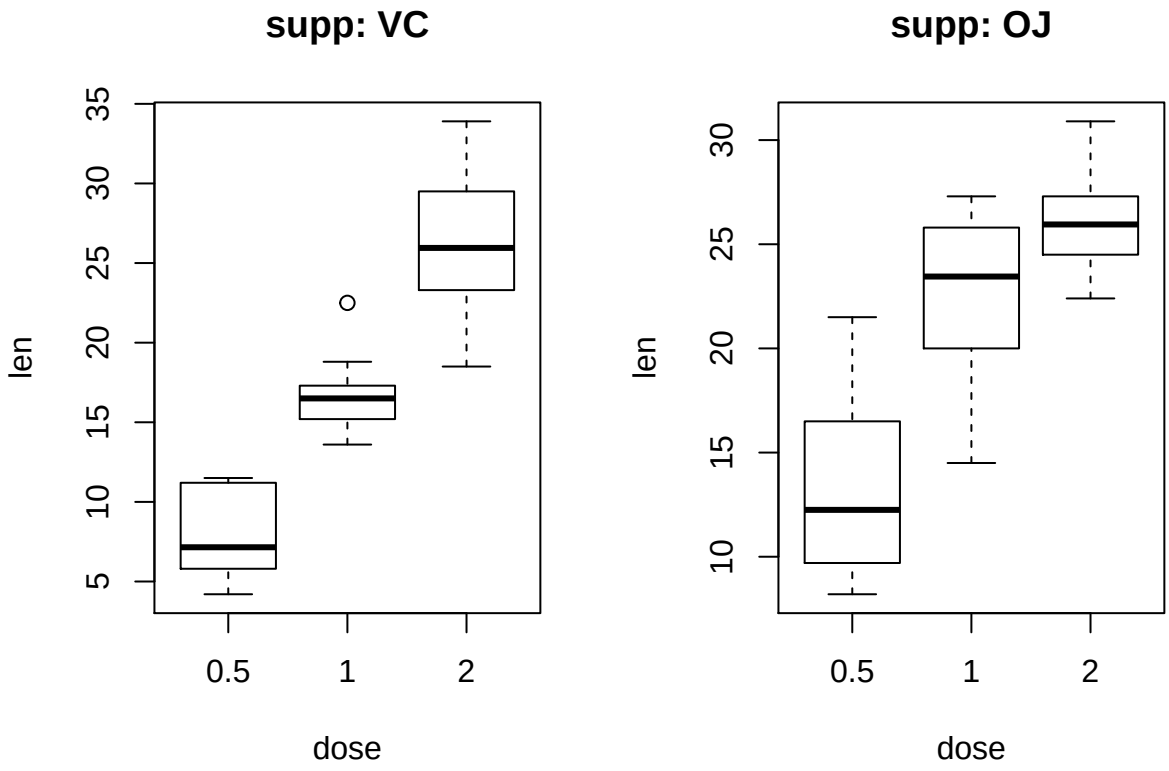
Briefly it appears that there are two supplements (supp), 'VC" and "OJ', with doses of 0.5, 1.0, and 2.0 for each. 'len' is the measurement of tooth growth length for each of these groups.

```
## Histogram of all data
hist(tg$len,breaks = 10, prob=TRUE, ylim=c(0, 0.06), xlim=c(0,40))
curve(dnorm(x, mean=mean(tg$len), sd=sd(tg$len)), add=TRUE, col="blue", lwd=4)
```




```
## Split data into supplements
VC_tg <- tg[tg$supp == "VC",]
OJ_tg <- tg[tg$supp == "OJ",]

## Plot boxplots of TG data for each supp
par(mfrow=c(1,2))
boxplot(formula=len ~ dose, data=VC_tg,names=unique(VC_tg$dose),
        main="supp: VC", xlab="dose", ylab="len")
boxplot(formula=len ~ dose, data=OJ_tg,names=unique(OJ_tg$dose),
        main="supp: OJ", xlab="dose", ylab="len")
```



Summary of ToothGrowth Data

We see that our distribution for all lengths very roughly follows a normal distribution, with the mean centered around **18.81**.

From our boxplots, we see a linear relationship between dose and len for both supplement groups (VC and OJ), though the difference between the supplement groups is not immediately apparent (see y-axes). We will probe this further in later sections.

Comparison of tooth growth by supp and dose

Using ?ToothGrowth in R, we see that the data set observes toothgrowth in 60 different Guinea pigs, which means **our observations are not paired.** For the analyses we set our alpha, the probability of Type I error, to 0.05.

Differences in supp

Our null hypothesis (Ho) is that there is no difference in mean len between VC and OJ. To address this, we perform an unpaired t test with a 95% confidence interval, using our VC_tg and OJ_tg data frames generated previously

```
test <- t.test(VC_tg$len, OJ_tg$len, paired=FALSE, var.equal=TRUE, conf.level=0.95)
test$conf.int
```

```
## [1] -7.5670064  0.1670064
## attr(,"conf.level")
## [1] 0.95
```

```
test$p.value
```

```
## [1] 0.06039337
```

Our confidence interval includes 0, so we can't rule out that our population difference is 0. Secondly, our p-value is greater than our designated alpha = 0.05.

Therefore, we fail to reject the null hypothesis that there is no difference in mean len between VC and OJ.

Differences in dose

We determined that there was no difference in mean len between the supp variables, VC and OJ. Next, we determine whether there is a difference between the dosages within these groups. Visually there appears to be (see boxplots above), however we again use an unpaired t-test using alpha = 0.05 to determine our differences.

Since there are three doses (0.5, 1.0, and 2.0) in each group, the appropriate test to use is *probably* an ANOVA. However we didn't go over this in the course, and therefore we will use multiple t tests to compare between doses of a group. We summarize this data in a table below. Data to generate the hypothesis tables are shown in the appendix.

```
kable(df)
```

supp	dose comparison	conf int	p value
VC	1.0 - 0.5	6.32 11.26	0.0000006
VC	2.0 - 1.0	5.77 12.97	0.0000340
OJ	1.0 - 0.5	5.53 13.41	0.0000836
OJ	2.0 - 1.0	0.22 6.5	0.0373628

Conclusions

In this report, we investigate the R dataset ToothGrowth answer the question of whether supp and/or dose affect tooth growth length. Several assumptions were made; first, we assumed the data roughly followed a normal distribution and showed this by overlaying a normal curve over the histogram. secondly, we assumed that the samples were indendent and therefore the t tests were unpaired. Finally, we set our alpha (type I error rate) to be 0.05, choosing a confidence interval of 95%.

The null hypothesis proposes there is no difference in the means of len between the two values of supp, VC and OJ, using a t test. Our confidence interval (-7.57, 0.17) contains 0, and our generated p-value, 0.0603934, is greater than our alpha value of 0.05. Therefore, we fail to reject the null hypothesis.

For dose difference, our null hypothesis states that there is no difference between the mean len of the three doses for each value of supp. By performing t tests, we find that for each dose compared, the confidence intervals do not contain zero, and every p-value is lower than our alpha of 0.05. Therefore, we reject the null hypothesis in favor of the alternative hypothesis that len is different for different doses in each supp.

Appendix

```
## ToothGrowth Data
tg <- ToothGrowth
dim(tg)
```

```
## [1] 60  3
```

```
head(tg)
```

```
##    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
```

```
## Table Generation

# VC Dose Tests
VC_0.5 <- VC_tg[VC_tg$dose == 0.5,]$len
VC_1.0 <- VC_tg[VC_tg$dose == 1.0,]$len
VC_2.0 <- VC_tg[VC_tg$dose == 2.0,]$len

testVC1 <- t.test(VC_1.0, VC_0.5, paired=FALSE, var.equal=TRUE, conf.level=0.95)
testVC1$conf.int
```

```
## [1]  6.315654 11.264346
## attr(,"conf.level")
## [1] 0.95
```

```
testVC1$p.value
```

```
## [1] 6.492265e-07
```

```
testVC2 <- t.test(VC_2.0, VC_1.0, paired=FALSE, var.equal=TRUE, conf.level=0.95)
testVC2$conf.int
```

```
## [1]  5.77104 12.96896
## attr(,"conf.level")
## [1] 0.95
```

```
testVC2$p.value
```

```
## [1] 3.397578e-05
```

```
# OJ Dose Tests
OJ_0.5 <- OJ_tg[OJ_tg$dose == 0.5,]$len
OJ_1.0 <- OJ_tg[OJ_tg$dose == 1.0,]$len
OJ_2.0 <- OJ_tg[OJ_tg$dose == 2.0,]$len

testOJ1 <- t.test(OJ_1.0, OJ_0.5, paired=FALSE, var.equal=TRUE, conf.level=0.95)
testOJ1$conf.int
```

```
## [1]  5.529186 13.410814
## attr(,"conf.level")
## [1] 0.95
```

```
testOJ1$p.value
```

```
## [1] 8.357559e-05
```

```
testOJ2 <- t.test(OJ_2.0, OJ_1.0, paired=FALSE, var.equal=TRUE, conf.level=0.95)
testOJ2$conf.int
```

```
## [1] 0.2194983 6.5005017
## attr(,"conf.level")
## [1] 0.95
```

```
testOJ2$p.value
```

```
## [1] 0.0373628
```

```
# Creating table
df <- data.frame()
df <- data.frame(c(rep("VC", 2), rep("OJ",2)))
df <- cbind(df, c(rep(c("1.0 - 0.5", "2.0 - 1.0"), 2)))
df <- cbind(df, c(paste(round(testVC1$conf.int[1:2],2),collapse="   "),
                  paste(round(testVC2$conf.int[1:2],2),collapse="   "),
                  paste(round(testOJ1$conf.int[1:2],2),collapse="   "),
                  paste(round(testOJ2$conf.int[1:2],2),collapse="   ")))
df <- cbind(df, c(testVC1$p.value, testVC2$p.value, testOJ1$p.value, testOJ2$p.value))
names(df) <- c("supp", "dose comparison", "conf int", "p value")
```