A Statistical Look at R's ToothGrowth Data Set

Jen Becker

July 23, 2015

Exploratory Data Analysis

The ToothGrowth data set contains a 60x3 data frame containing the results of testing dose and delivery method of vitamin C on the tooth length of guinea pigs. The data frame contains columns for the length of the tooth, the supplement used to deliver the vitamin C (orange juice or ascorbic acid), and the dose of vitamin C (0.5, 1.0, or 2.0 milligrams).

There are 10 length measurements for each combination of dose and supplement.

Tooth length varies from 4.20 to 33.90, with a mean of 18. A summary for each type of supplement shows that, at first glance, tooth length seems to be greater with orange juice (22.7 vs 16.5). See Appendix for supporting R code.

A quick graph shows that tooth length increases as the vitamin C dose increases. We can also see that it appears that the orange juice supplement results in longer teeth for lower doses of vitamin C, while the effects seem more equal at the highest dose.

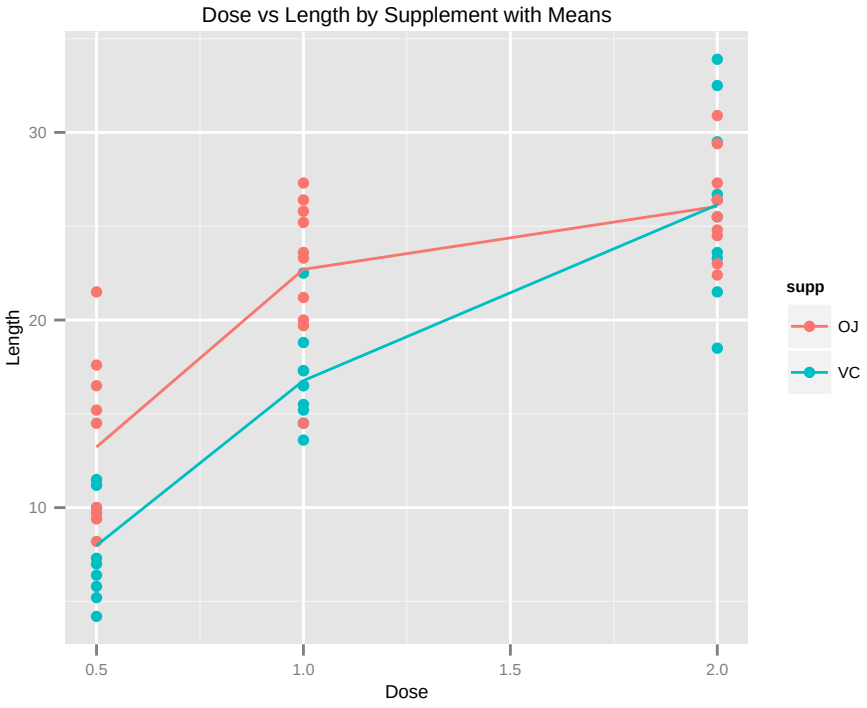

Figure 1: See Appendix for supporting R code

Comparisons between doses of vitamin C

We first analyze whether increasing doses of vitamin C are associated with longer teeth, regardless of the supplement used to deliver it. To do this, we assume that the samples are independent and identically

distributed.

The null hypothesis is that there is no difference in tooth growth as the dose of vitamin C is varied.

To discover this, we construct 95% confidence intervals using t-tests, comparing the differences in length for each dose (ignoring the supplement).

```
t.test(ToothGrowth[ToothGrowth$dose==0.5,"len"], ToothGrowth[ToothGrowth$dose==1.0, "len"])$conf.int
```

```
## [1] -11.983781  -6.276219
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(ToothGrowth[ToothGrowth$dose==1.0,"len"], ToothGrowth[ToothGrowth$dose==2.0, "len"])$conf.int
```

```
## [1] -8.996481 -3.733519
## attr(,"conf.level")
## [1] 0.95
```

In both cases, the confidence intervals for the difference in mean do not include 0, so we reject the null hypthesis that there is not a statistical difference in tooth growth associated with different doses of vitamin C.

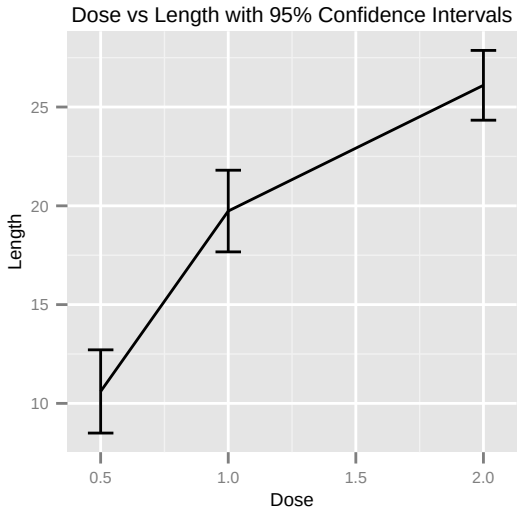

Figure 2: See Appendix for supporting R code

Comparisons between supplements within a dose

Another method of analyzing this data is to compare the differences in tooth length between the 2 supplements within a given dose. This will give us a clue as to whether different supplements may have an impact on the effectiveness of vitamin C on tooth growth. To do this, we assume that the samples are independent and identically distributed.

The null hypothesis is that there is no difference in tooth growth between supplements if the dose remains the same.

To discover this, we construct 95% confidence intervals using t-tests, comparing the differences in length for each dose between supplements.

Orange juice vs Ascorbic acid, 0.5 milligram dose:

```
## [1] -8.780943 -1.719057
## attr(,"conf.level")
## [1] 0.95
```

Orange juice vs Ascorbic acid, 1.0 milligram dose:

```
## [1] -9.057852 -2.802148
## attr(,"conf.level")
## [1] 0.95
```

Orange juice vs Ascorbic acid, 2.0 milligram dose:

```
## [1] -3.63807  3.79807
## attr(,"conf.level")
## [1] 0.95
```

By looking at these results, we can reject the null hypothesis that the supplement does not affect tooth growth for doses equal to 0.5 and 1.0 milligrams, but we cannot reject the null hypothesis for doses equal to 2.0 milligrams.

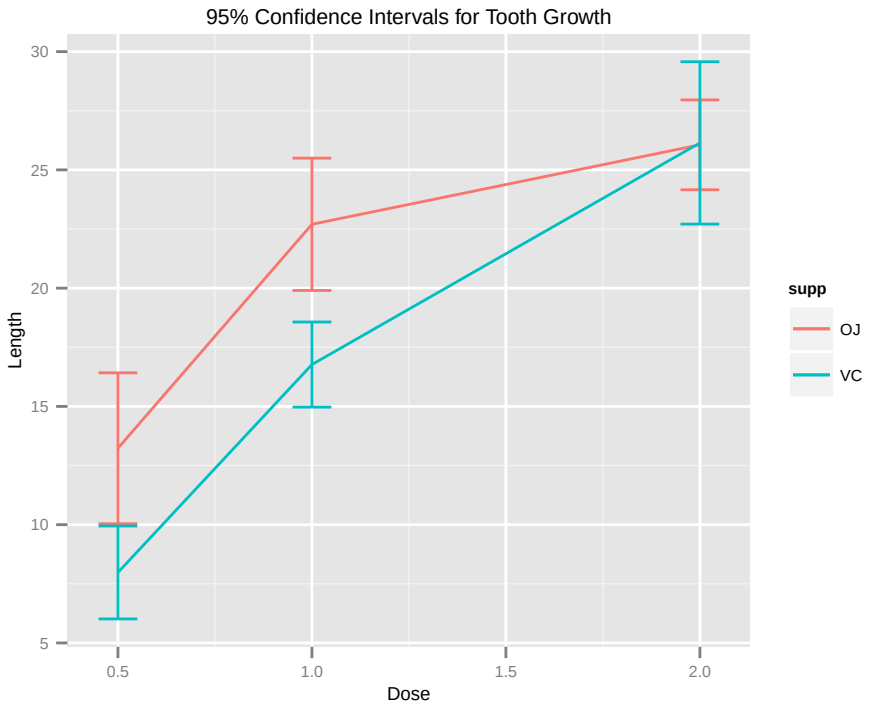

Figure 3: See Appendix for supporting R code

Appendix

Exploratory data anaylsis code

```
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 ...
```

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

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

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

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

```
summary(ToothGrowth[ToothGrowth$supp == "OJ",]$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20   15.52   22.70   20.66   25.72   30.90
```

```
summary(ToothGrowth[ToothGrowth$supp == "VC",]$len)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20   11.20   16.50   16.96   23.10   33.90
```

Figure 1

```
library(plyr); library(ggplot2)
## Create a data frame that also includes the mean and confidence intervals
## per dose/supplement
tgmean <- ddply(ToothGrowth, .(dose, supp), summarize,
                mean=mean(len),
                confmin=t.test(len)$conf.int[1],
                confmax=t.test(len)$conf.int[2])
## Plot the dose vs supplement with length points and lines connecting the means
g <- ggplot(ToothGrowth, aes(dose, len)) +
    geom_point(data=ToothGrowth, mapping=aes(x=dose, y=len, col=supp)) +
    geom_line(data=tgmean, mapping=aes(x=dose, y=mean, col=supp)) +
    theme(text = element_text(size = 7)) +
    xlab("Dose") + ylab("Length") + ggtitle("Dose vs Length by Supplement with Means")
g
```

Figure 2

```
library(plyr); library(ggplot2)
## Create a data frame with means and confidence intervals per dose only
doseonly <- ddply(ToothGrowth, .(dose), summarize,
                  mean=mean(len),
                  confmin=t.test(len)$conf.int[1],
                  confmax=t.test(len)$conf.int[2])
## Plot dose vs length means with confidence intervals
g1 <- ggplot() +
    geom_line(data=doseonly, mapping=aes(x=dose, y=mean)) +
    geom_errorbar(data=doseonly, width=0.1,
                  mapping=aes(x=dose, ymin=confmin, ymax=confmax)) +
    theme(text = element_text(size = 7)) +
    xlab("Dose") + ylab("Length") + ggtitle("Dose vs Length with 95% Confidence Intervals")
g1
```

Figure 3

```
library(ggplot2)
## Plot dose vs length with means and confidence intervals per supplement
g2 <- ggplot() +
    geom_line(data=tgmean, mapping=aes(x=dose, y=mean, col=supp))+
    geom_errorbar(data=tgmean, width=0.1,
                  mapping=aes(x=dose, ymin=confmin, ymax=confmax, col=supp)) +
    theme(text = element_text(size = 7)) +
    xlab("Dose") + ylab("Length") + ggtitle("95% Confidence Intervals for Tooth Growth")
g2
```