Statistical Inference- Course Project

Tooth Growth data

(Note: we need the `ggplot2` package for graphing and the `reshape2` package for the `dcast` function.)

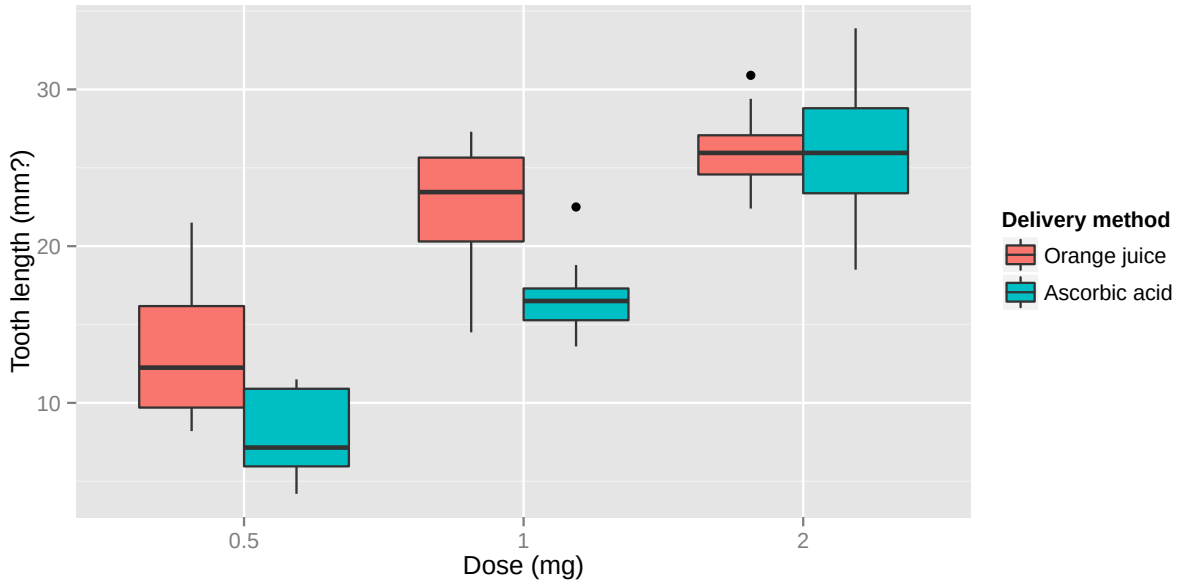
1. Load the ToothGrowth data and perform some basic exploratory data analyses.

The ToothGrowth data is located in the `datasets` package. In two pages, there is no room to show all the EDA. (I looked at a bunch of histograms and boxplots of the variables.) The most helpful summary of the data is given below.

2. Provide a basic summary of the data.

A good way summarize the data is a clustered boxplot. (The unit of measurement for tooth length is not specified—maybe millimeters?)

```
ggplot(ToothGrowth, aes(x = factor(dose), y = len, fill = supp)) +
    xlab("Dose (mg)") +
    ylab("Tooth length (mm?)") +
    scale_fill_discrete(name="Delivery method",
                        breaks=c("OJ", "VC"),
                        labels=c("Orange juice", "Ascorbic acid")) +
    geom_boxplot()
```



It appears that at each dose except the highest, delivering the supplemental Vitamin C by orange juice corresponds to longer teeth. There is a clear increase in tooth length as the dose increases.

3. Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose. (Use the techniques from class even if there's other approaches worth considering).

With two factors, a good way to analyze this data would be two-factor ANOVA; however, since we are limited to using tools we saw in the lecture, I will conduct two paired t-tests.

First we will look at a possible difference in tooth length based on the delivery method. For the supplement type, orange juice is coded as `OJ` and ascorbic acid is coded `VC` within the variable `supp`. If μ_d is the mean difference in tooth length (OJ minus VC), the null and alternate hypotheses are

$H_0 : \mu_d = 0$

$H_A : \mu_d \neq 0$

We reshape the data so that each row represents a single guinea pig. Tooth length with OJ and tooth length with VC will be separate variables (paired for each guinea pig), with each measurement averaged over the three dosing levels. The `dcast` function needs an ID variable, so we must add it first (assuming that the observations on the guinea pigs are in the same order for each combination of factors).

```
ID <- rep(1:10, 6)
ToothGrowth2 <- cbind(ID, ToothGrowth)
ToothGrowth_by_supp <- dcast(ToothGrowth2, ID ~ supp,
                             fun.aggregate = mean,
                             value.var = "len")
```

Now we run a t-test and get a t-interval for this paired data.

```
t.test(ToothGrowth_by_supp$OJ, ToothGrowth_by_supp$VC, paired = TRUE)
```

Next we will look at a possible difference in tooth length based on the dosage. A paired t-test can only compare two levels, so we will compare the lowest dose (0.5 mg) directly to the highest dose (2 mg). The hypotheses are as before and the reshaping process is likewise similar.

```
ToothGrowth_by_dose <- dcast(ToothGrowth2, ID ~ dose,
                             fun.aggregate = mean,
                             value.var = "len")
```

```
t.test(ToothGrowth_by_dose$`2`, ToothGrowth_by_dose$`0.5`, paired = TRUE)
```

(To get the report to fit in two pages, I did not have space to echo the results of the `t.test` functions.)

4. State your conclusions and the assumptions needed for your conclusions.

For the first test, there is evidence that that the teeth are longer when the dose is given through orange juice versus ascorbic acid. ($P < 0.001$) We are 95% confident that the true population difference is captured in the interval (2.3, 5.1).

For the second test, there is evidence that teeth are longer when the dose is 2 mg versus 0.5 mg. ($P < 0.001$) We are 95% confident that the true population difference is captured in the interval (12.6, 18.3).

For these conclusions to be valid, we must be assured that these 10 guinea pigs are representative of all guinea pigs so that the theoretical assumption of independent, identically distributed draws is satisfied. Since 10 is a small sample size, we also need to assume that the population of tooth length differences (whether such differences are due to delivery method or dose) in guinea pigs is normally distributed. This plausibility of the latter condition is nearly impossible to verify (graphically or otherwise) given only 10 data points. (Histograms and Q-Q plots of the differences show some evidence of skewness and potentially problematic outliers.)