Statistical Inference Course Project Part 2

Now in the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.

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

```
library('datasets')
```

2. Provide a basic summary of the data.

```
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

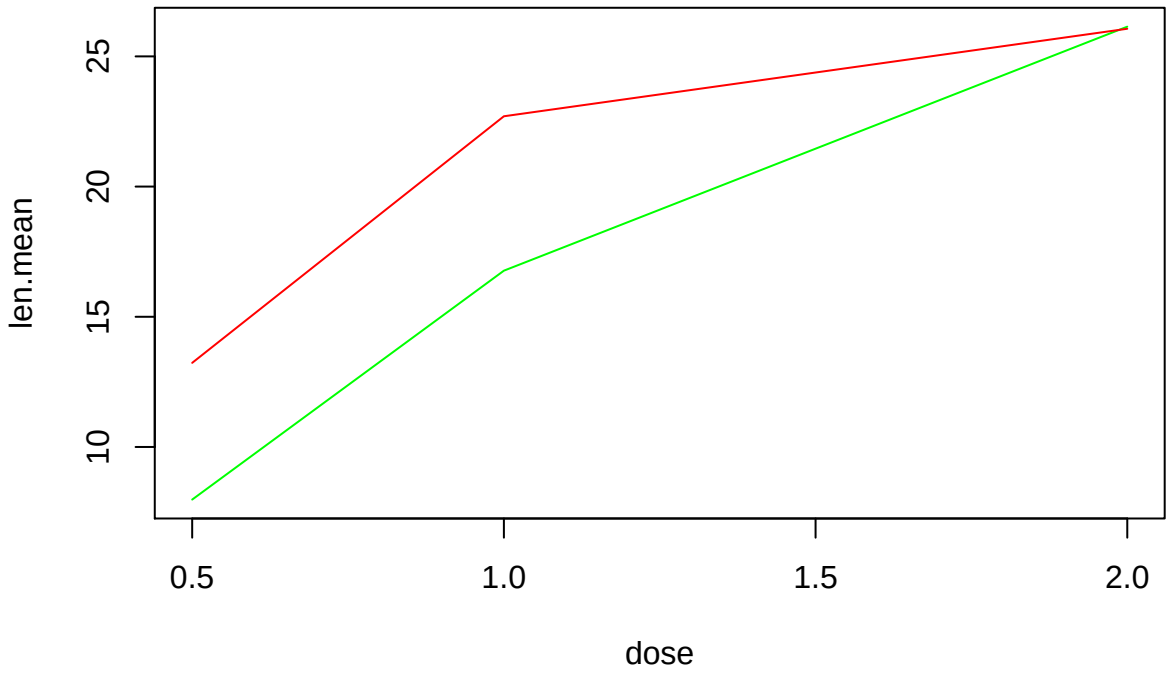
The data shows the tooth length of guinea pigs who had been given different amounts of Vitamin C (0.5, 1, and 2mg) with different delivery methods (orange juice or ascorbic acid). There are 6 possible combinations of dose and delivery method, and each combination has 10 samples. Here is the mean and standard deviation of each group:

```
df <- data.frame(aggregate(len ~ dose + supp, data=ToothGrowth, function(x){c(mean=mean(x),sd=sd(x))}))
df <- do.call(data.frame, df)
df
```

```
##   dose supp len.mean   len.sd
## 1  0.5   OJ    13.23 4.459709
## 2  1.0   OJ    22.70 3.910953
## 3  2.0   OJ    26.06 2.655058
## 4  0.5   VC     7.98 2.746634
## 5  1.0   VC    16.77 2.515309
## 6  2.0   VC    26.14 4.797731
```

Here it is plotted with Vitamin C in green and Ascorbic Acid in red.

```
plot(len.mean~dose, data=df[df$supp=="VC",], type='l', col='green')
lines(len.mean~dose, data=df[df$supp=="OJ",], col='red')
```



Based on the means, it would appear that increasing the dose leads to longer teeth, and delivery via Orange Juice is a bit more effective than via Ascorbic Acid, particularly at low doses.

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

Let us test the hypotheses (H_{a1}) that the group with a dose of 2mg has longer teeth than the group with dose 0.5mg, versus the null hypothesis (H_{01}) that those two groups have the same mean tooth length. The 95% confidence interval that we get is

```
ci <- function(xs, ys) {
    mx <- mean(xs)
    sx <- sd(xs)
    nx <- length(xs)
    my <- mean(ys)
    sy <- sd(ys)
    ny <- length(ys)
    # pooled variance
    sp <- sqrt(((nx-1)*sx^2 + (ny-1)*sy^2)/(nx+ny-2))
    my - mx + c(-1, 1) * qt(0.975, nx+ny-2) * sp * sqrt(1/nx+1/ny)
  }
xs <- ToothGrowth[ToothGrowth$dose==0.5, "len"]
ys <- ToothGrowth[ToothGrowth$dose==2, "len"]
ci(xs, ys)
```

```
## [1] 12.83648 18.15352
```

The confidence interval does not include 0, so we can reject the null hypothesis.

Now let us test the hypotheses (H_{a2}) that delivery by Ascorbic Acid leads to longer teeth than delivery by Orange Juice, versus the null hypothesis (H_{02}) that those two groups have the same mean tooth length. In this case, 95% confidence interval that we get is

```r
xs <- ToothGrowth[ToothGrowth$supp=="VC", "len"]
ys <- ToothGrowth[ToothGrowth$supp=="OJ", "len"]
ci(xs, ys)
```

```
## [1] -0.1670064  7.5670064
```

Since the confidence interval does include 0, we cannot reject the null hypothesis.

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

From this data, we have reason to believe that a 2mg dose of Vitamin C will increase guinea pig tooth length more than a 0.5mg dose will. The effect of delivery method is inconclusive when we aggregate across all the given dosages, though further studies with this data may find effects for particular dosages.

In drawing these conclusions, I am making a number of assumptions. First, I am assuming that the tooth length vs dose or delivery method is roughly symmetric and bell shaped, so that we can approximate it with a T distribution. Second, I am assuming that the data is not paired. It's possible that the data was drawn from the same 10 guinea pigs, but since that is not explicitly stated, I am assuming that it was drawn from 60 separate subjects. I am also assuming that the variances for all the different groups are the same.