Statistical Inference Course Assignment

Tony Bredehoeft

September 21, 2014

In this assignment we were asked to focus on two things: simulation of the exponential distribution and evaluating the ToothGrowth data set provided in R. I'll start with the simulation of the exponential distribution and highlight some properties of the simulation and how they relate to their theoretical counterparts.

I'll start by running 1000 simulations of the exponential distribution, each with 40 values and a lambda of 0.2. The theoretical mean and standard deviation of the exponential distribution are both 1/lambda, meaning the theoretical mean and standard deviation are both 5 in our simulation.

```
set.seed(1)
exp.data <- data.frame()
i = 1
while(i <= 1000){
  exp.data <- rbind(exp.data,rexp(40,.2))
  i = i + 1
}
```

Now we have our 1000 simulations of the exponential distribution in a new data frame. Next we'll look at the sample means from each of our simulations, take the mean of these sample means and compare that value to the theoretical mean of 5.

```
mean(rowMeans(exp.data))
```

```
## [1] 4.99
```

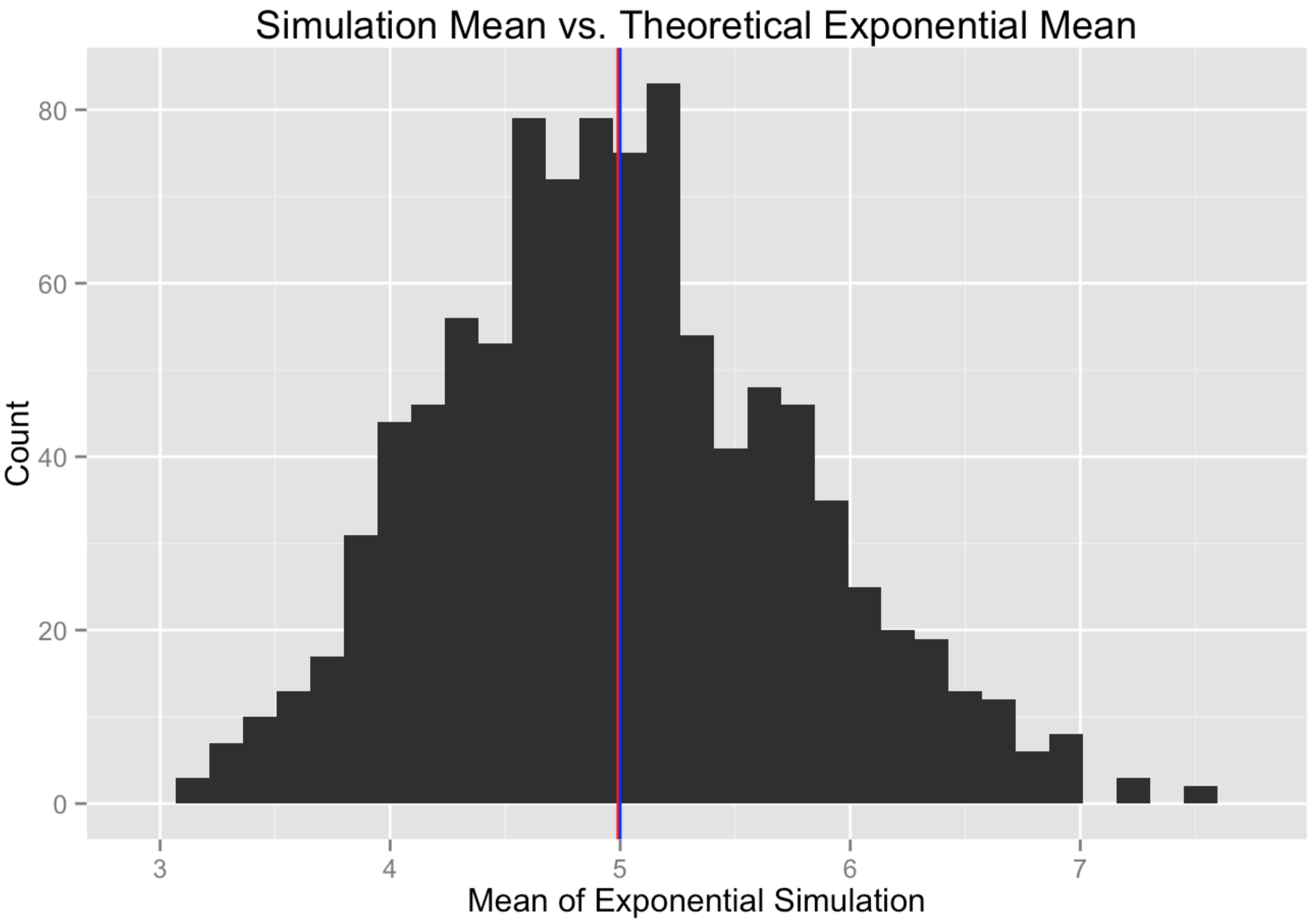
Here we can see the mean value provided by our simulations is 4.99, which is very close to the theoretical value.

Next I'll plot a histogram of the sample means provided by each simulation, and show how the simulation mean compares to the theoretical mean of 5.

Note that the theoretical mean is the blue line and the mean of the sample means is the red line below.

```
library(ggplot2)
ggplot(exp.data,aes(x=rowMeans(exp.data))) +
  geom_histogram() +
  geom_vline(xintercept = mean(rowMeans(exp.data)),color='red') +
  geom_vline(xintercept = 5,color='blue') +
  xlab('Mean of Exponential Simulation') +
  ylab('Count') +
  ggtitle('Simulation Mean vs. Theoretical Exponential Mean')
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



We can see that both the simulation and theoretical means are very close (in fact the lines overlap) and that the distribution of the sample means approximates a normal distribution.

Next let’s take a look at how the variance of the sample means compares to the theoretical variance.

```
var(rowMeans(exp.data))
```

```
## [1] 0.6111
```

```
(1/.2)^2/40
```

```
## [1] 0.625
```

Here we can see that the variance of the sample means of our simulation is slightly less than the theoretical variance provided by the formula (1/lambda)^2/n.

Now we want to evaluate the coverage provided by our simulation. As in the class lectures, for each simulation in our 1000 simulations I will calculate the confidence interval for the sample mean.

Next I will determine what percentage of these confidence intervals contain the theoretical mean of 5. Given we are looking at a 95% confidence interval for the sample means, the coverage value should be around .95.

```
coverage <- apply(exp.data,1,function(x) {
  ll <- mean(x) - 1.96*5/sqrt(40)
  ul <- mean(x) + 1.96*5/sqrt(40)
  mean(ll<5 & ul>5)
})
sum(coverage)/1000
```

```
## [1] 0.949
```

Here we can see that 0.949 of our sample confidence intervals (95%) contained the true population mean.

ToothGrowth Data Evaluation

Now we'll move into the second part of the assignment where I'll take a look at the ToothGrowth data set available in R. This data set tracks the tooth length of guinea pigs when given doses of Vitamin C in two delivery methods (orange juice or ascorbic acid).

First let's glance at the data set and the variables provided.

```
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 ...
```

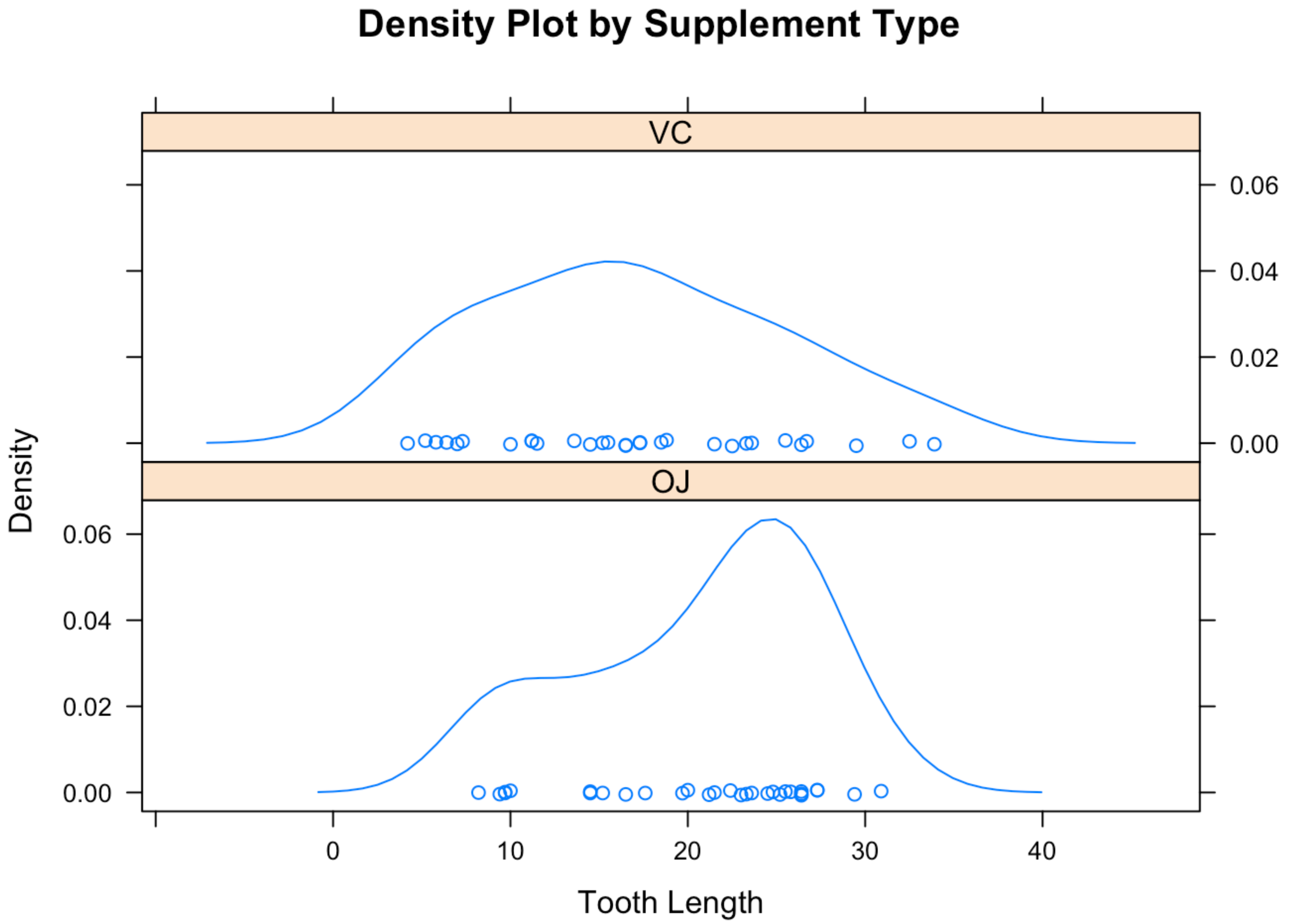
```
head(ToothGrowth)
```

```
##    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
```

Here, we can see that there are 3 variables in the data, where “len” is the the tooth length, “supp” is the delivery method and “dose” is the dose amount in milligrams.

Now I’m going to plot the density of tooth length by delivery method.

```
library(lattice)
with(ToothGrowth,densityplot(~ len|supp,
            main="Density Plot by Supplement Type",
            xlab="Tooth Length",
            layout=c(1,2)))
```

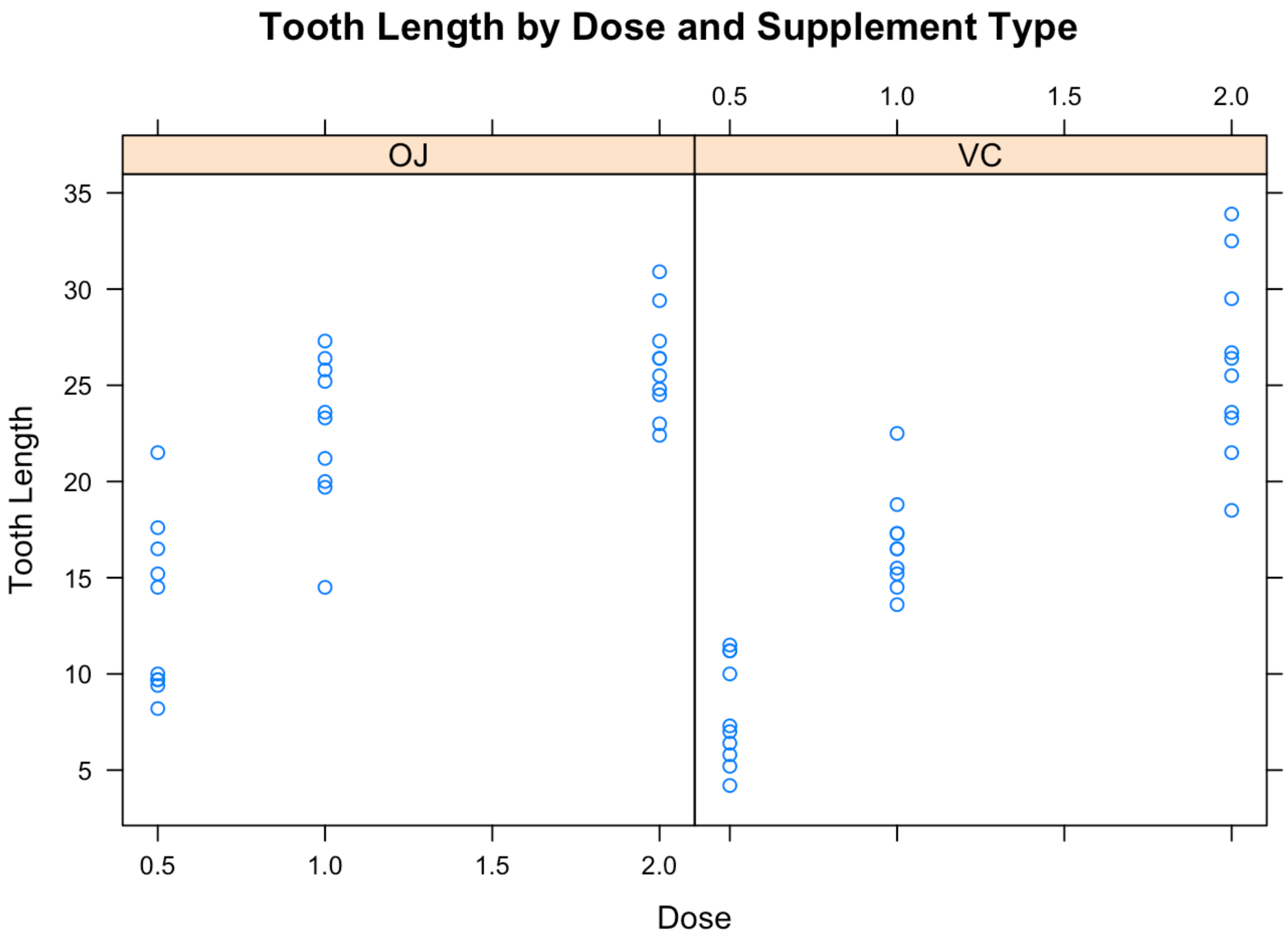


Here we can see that tooth length for guinea pigs delivered asorbic acid has a much wider distribution than those that were delivered orange juice. Also, the distribution for orange is almost bimodal, peaking at both 10 and 25.

Overall, it looks like the tooth length values are slightly higher for the orange juice distribution as well.

Now I'm going to add dose amount to the plot to see how this affects tooth length values.

```
with(ToothGrowth,xyplot(len ~ dose|supp,
            main="Tooth Length by Dose and Supplement Type",
            xlab="Dose",
            ylab="Tooth Length",
            layout=c(2,1)))
```



For both the asorbic acid and orange juice delivery methods, we can see that dose amount is positively correlated with tooth growth. Also, we can see that the asorbic acid dose of 2 milligrams looks to have a much higher variance in guinea pig tooth size.

Now let's look at a brief summary of the data.

```
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

```
describeBy(ToothGrowth,ToothGrowth$supp)
```

```
## group: OJ
##       vars  n  mean   sd median trimmed  mad min  max range  skew kurtosis
## len      1 30 20.66 6.61   22.7   21.04 5.49 8.2 30.9  22.7 -0.52    -1.03
## supp*    2 30  1.00 0.00    1.0    1.00 0.00 1.0  1.0   0.0   NaN      NaN
## dose     3 30  1.17 0.63    1.0    1.15 0.74 0.5  2.0   1.5  0.36    -1.60
##         se
## len   1.21
## supp* 0.00
## dose  0.12
## --------------------------------------------------------
## group: VC
##       vars  n  mean   sd median trimmed  mad min  max range skew kurtosis
## len      1 30 16.96 8.27   16.5   16.57 9.27 4.2 33.9  29.7 0.28    -0.93
## supp*    2 30  2.00 0.00    2.0    2.00 0.00 2.0  2.0   0.0  NaN      NaN
## dose     3 30  1.17 0.63    1.0    1.15 0.74 0.5  2.0   1.5 0.36    -1.60
##         se
## len   1.51
## supp* 0.00
## dose  0.12
```

Here we can see that the mean for the orange juice delivery method is slightly higher than that for the asorbic acid delivery method. Also, as we saw in our density plot above, the orange juice distribution has a noticeable left skew.

Now we'll begin to compare tooth length by dose and supplement using the moethods taught in class. In the description of the data, it suggests that 10 guinea pigs were used in the experiment, though it doesn't specify whether it is the same guinea pigs used across the different dose and supplement values. For this reason I'm using two sample t-tests and not paired sample t-tests.

I'll start by evaluating if there is a significant difference between values for the 2 delivery mechanisms (asorbic acid and orange juice) with a two sample t-test.

```
with(ToothGrowth, t.test(len ~ supp, paired = FALSE))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.915, df = 55.31, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.171  7.571
## sample estimates:
## mean in group OJ mean in group VC
##            20.66            16.96
```

We can see that the t statistic is 1.91, which is less than 1.96 which represents the right boundary of the 95% confidence interval, meaning the difference between tooth length values of the asorbic acid and orange juice groups is not significant at the 95% confidence level. This is also reflected in the p-value, which is greater than 0.05, and in the 95% confidence interval for the difference which contains 0.

Next I will compare the dose values (0.5,1 and 2 milligrams) with the same two sample t-tests. Note that we are focused on t-tests at the moment, so I can only compare 2 levels of values. This means I will have to create subsets of the data to compare 2 dose levels at a time.

```
TG.sub.1 <- subset(ToothGrowth, dose %in% c(0.5, 1))
TG.sub.2 <- subset(ToothGrowth, dose %in% c(0.5, 2))
TG.sub.3 <- subset(ToothGrowth, dose %in% c(1, 2))

with(TG.sub.1, t.test(len ~ dose, paired = FALSE))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.477, df = 37.99, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.984  -6.276
## sample estimates:
## mean in group 0.5   mean in group 1
##             10.61             19.73
```

```
with(TG.sub.2, t.test(len ~ dose, paired = FALSE))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.8, df = 36.88, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.16 -12.83
## sample estimates:
## mean in group 0.5   mean in group 2
##             10.61             26.10
```

```
with(TG.sub.3, t.test(len ~ dose, paired = FALSE))
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.901, df = 37.1, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996 -3.734
## sample estimates:
## mean in group 1 mean in group 2
##           19.73           26.10
```

Above we can see that each of the comparisons between dose levels were highly significant, with p-values below 0.05 and confidence intervals that do not contain 0. Overall, both dose level and delivery method were statistically significant in corresponding tooth length values for guinea pigs.

A next step would be to combine multiple factors in our comparison, as well as to include a baseline/placebo value for a group of guinea pigs. Thanks for reading, please feel free to comment.