Statistical Inference Course Project Part 2

Evan Oman

02/22/2015

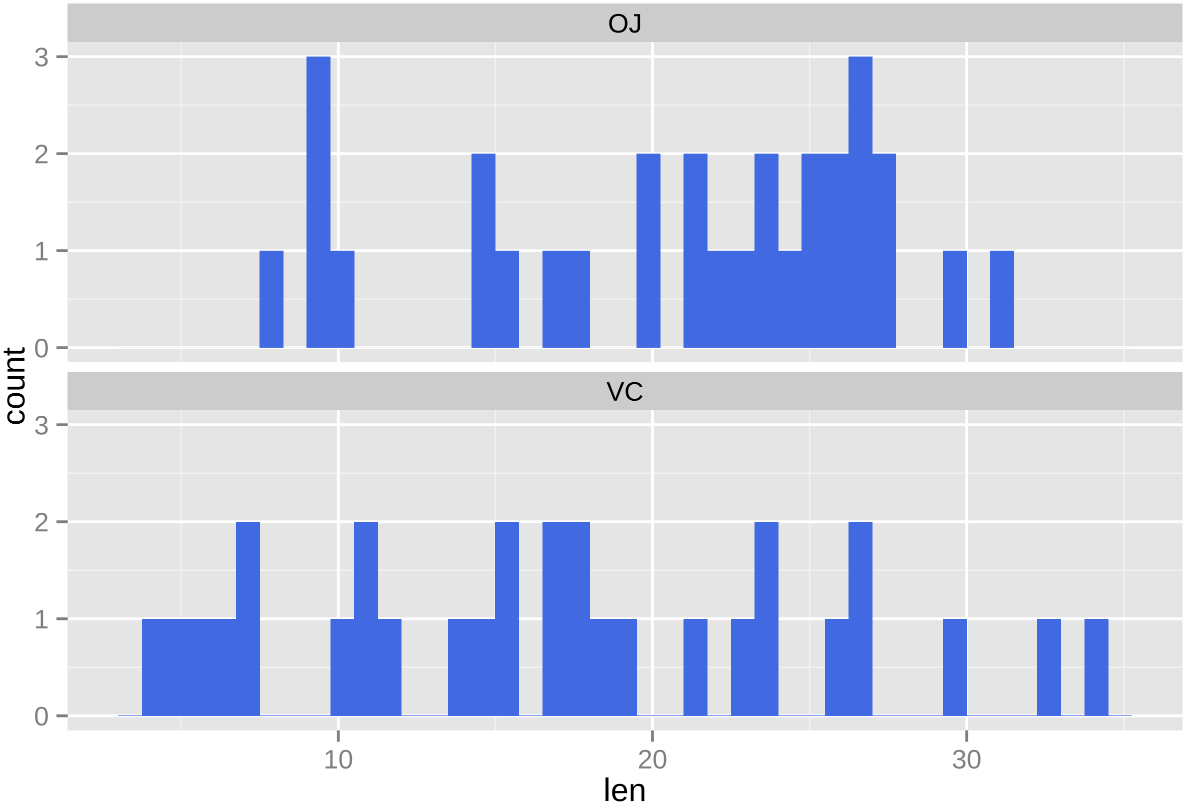
Introduction

The ToothGrowth dataset provided by *R* is a dataset collected in order to measure the effect on Guinea Pig tooth growth based on two different delivery supplements at three different dosages. The dataset consists of 60 observations with inputs: delivery methods(orange juice or ascorbic acid) and dosage: .5, 1, or 2 mg; and output variable: tooth length(more specifically the length of certain cells within the tooth). We are asked to determine whether or not tooth growth is affected by delivery method(supp) or by dosage(dose).

Exploratory Analysis

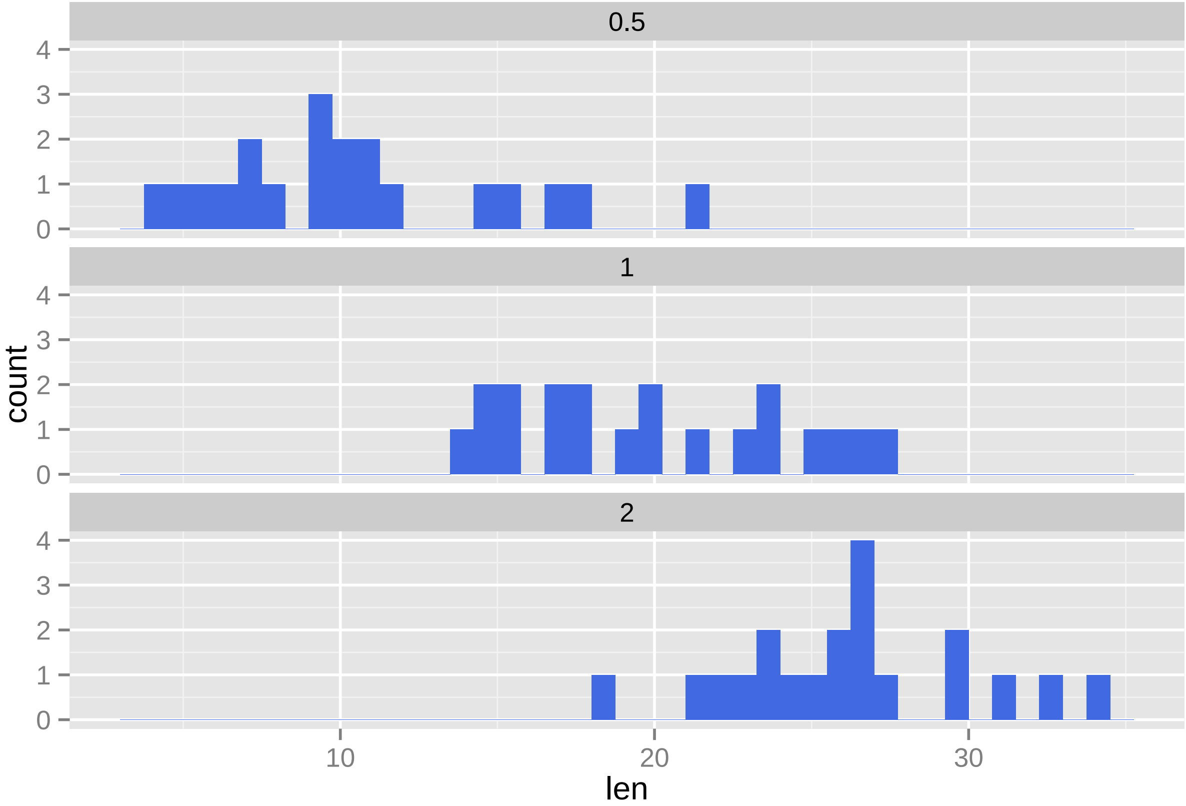
We begin by trying to get a rough idea of what the data looks like. First we plot the tooth length by supp:

```
library(ggplot2)
data(ToothGrowth)
ToothGrowth$supp <- as.factor(ToothGrowth$supp)
ggplot(ToothGrowth, aes(x=len)) +
    geom_histogram(fill = "royalblue", binwidth = 0.75) +
    facet_wrap( ~ supp, ncol=1)
```



Based on this plot I would say that there is definitely a difference in the variance of the tooth length between VC or OJ, but I am not exactly sure if there is a statistically significant difference between the means. Now we plot tooth length by dosage:

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x=len)) +
    geom_histogram(fill = "royalblue", binwidth = 0.75) +
    facet_wrap( ~ dose, ncol=1)
```



Here it is pretty clear that there is a noticable difference between the means of each dosage.

Hypothesis Testing

We will now perform hypothesis testing to answer concretely the questions posed above. For these tests we will use a T distribution(since we have a very small sample size) and an α level of .05.

First we will look at the supp variable. We define the hypotheses to be:

- H_0 : The means μ_{VC} and μ_{OJ} are the same, such that $\mu_{VC} - \mu_{OJ} = 0$
- H_a : The means μ_{VC} and μ_{OJ} are not the same, such that $\mu_{VC} - \mu_{OJ} \neq 0$

Here we can simply use a two-tailed T test with R:

```
t.test(len ~ supp, data = ToothGrowth)$p.value
```

```
## [1] 0.06063451
```

Thus we can see that our p value is .0606. Then since $p > \alpha$, we fail to reject the null hypothesis that the means are the same(though we do so just barely). Thus we can conclude that, in our sample, the supp variable does not affect the length of Guinea Pig teeth in a statistically significant manner.

Next we will consider the dosages given by the dose variable. Since this measurement is split into 3 levels, we will need to look at the difference in means for .05 mg vs 1 mg, 1 mg vs 2 mg, and .5 mg vs 2 mg(note that if we were not limited to the methods introduced in this course, this problem would be a prime candidate for ANOVA). The following code makes three data sets based on the comparisons mentioned above:

```
TG12 <- ToothGrowth[ToothGrowth$dose == .5 | ToothGrowth$dose == 1,]
TG13 <- ToothGrowth[ToothGrowth$dose == 1 | ToothGrowth$dose == 2,]
TG23 <- ToothGrowth[ToothGrowth$dose == .5 | ToothGrowth$dose == 2,]
```

Now that we have our data we define our hypotheses as:

- $H_0^{i,j}$: The means μ_i and μ_j are the same, such that $\mu_i - \mu_j = 0$
- $H_a^{i,j}$: The means μ_i and μ_i are not the same, such that $\mu_i - \mu_j \neq 0$

Where $i, j \in \{1, 2, 3\},\ i < j$. We can then perform T tests as we did before on each of these 3 data sets:

```
t.test(len ~ dose, data = TG12)$p.value
```

```
## [1] 1.268301e-07
```

```
t.test(len ~ dose, data = TG13)$p.value
```

```
## [1] 1.90643e-05
```

```
t.test(len ~ dose, data = TG23)$p.value
```

```
## [1] 4.397525e-14
```

Checking each of the p values, we see that each is well below our α level of .05 so we reject all three of our null hypotheses and conclude that, in our sample, the dosage level correlates to a statistically significant change in the length of Ginea Pig teeth.