Statistical Inference: Tooth Growth Analysis

Henrique Souza (github.com/htssouza)

Goals

1-) Load the ToothGrowth data and perform some basic exploratory data analyses 2-) Provide a basic summary of the data 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) 4-) State your conclusions and the assumptions needed for your conclusions

1-) Load the ToothGrowth data and perform some basic exploratory data analyses

```
# Load data
data(ToothGrowth)

# Data summary
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 ...
```

```
# A small sample of the data
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
```

2-) Provide a basic summary of the data.

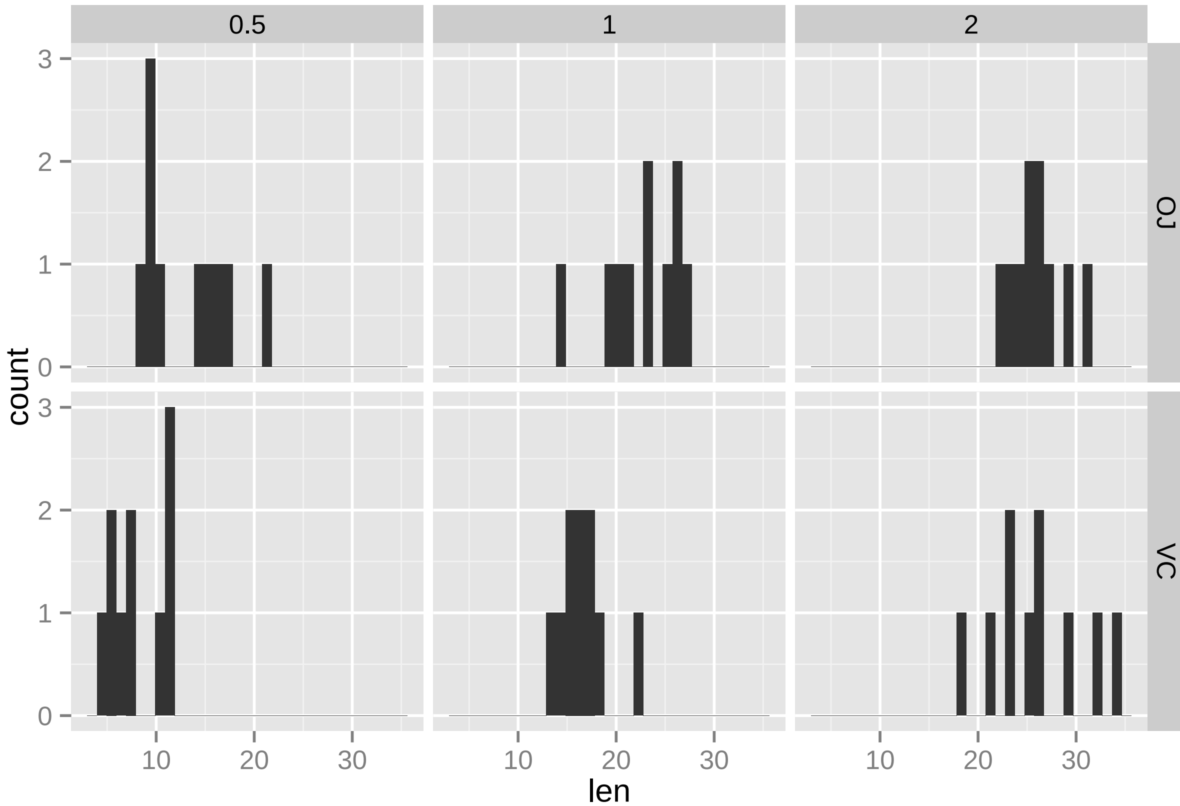
Dose is a factor (few different valyes) so make the preprocessing conversion:

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

Plotting some histograms:

```r
library(plyr)
library(ggplot2)
qplot(
  len,
  data = ToothGrowth,
  facets = .~supp~dose,
  geom="histogram")
```

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



3-) Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

Summarizing data to compute confidence intervals (using Student T interval):

```r
summaries <- ddply(ToothGrowth, .(supp, dose), summarize, mean = mean(len), sd = sd(len), n = length(le
summaries
```

```
##   supp dose  mean       sd  n
```

```
## 1   OJ  0.5 13.23 4.459709 10
## 2   OJ    1 22.70 3.910953 10
## 3   OJ    2 26.06 2.655058 10
## 4   VC  0.5  7.98 2.746634 10
## 5   VC    1 16.77 2.515309 10
## 6   VC    2 26.14 4.797731 10
```

Computing confidence intervals (95% of confidence):

```
summaries$min <- summaries$mean - (qt(.975, (summaries$n-1) * summaries$sd / sqrt(summaries$n)))
summaries$max <- summaries$mean + (qt(.975, (summaries$n-1) * summaries$sd / sqrt(summaries$n)))
summaries
```

```
##   supp dose  mean       sd  n       min      max
## 1   OJ  0.5 13.23 4.459709 10 11.064300 15.39570
## 2   OJ    1 22.70 3.910953 10 20.502166 24.89783
## 3   OJ    2 26.06 2.655058 10 23.730220 28.38978
## 4   VC  0.5  7.98 2.746634 10  5.664568 10.29543
## 5   VC    1 16.77 2.515309 10 14.415961 19.12404
## 6   VC    2 26.14 4.797731 10 23.990112 28.28989
```

4-) Conclusions

Based on the summary above we can say (with 95% of sure that):

- lens when using OJ (.5 dose) > VC (.5 dose)
- lens when using OJ (1.0 dose) > VC (1.0 dose)
- lens when using 2.0 of both OJ and VC are similar/equivalent

The assumptions are:

- current dataset (small, with only 60 rows)
- 95% of confidence