Coursera Statistical Inference Project - Part 2

Father Abraham

13/09/2014

Part 2

In this part of we're going to analyze the ToothGrowth data in the R datasets package. Load the datasets package into the variable "tg" and do some exploratory data analysis:

```
str(tg)
```

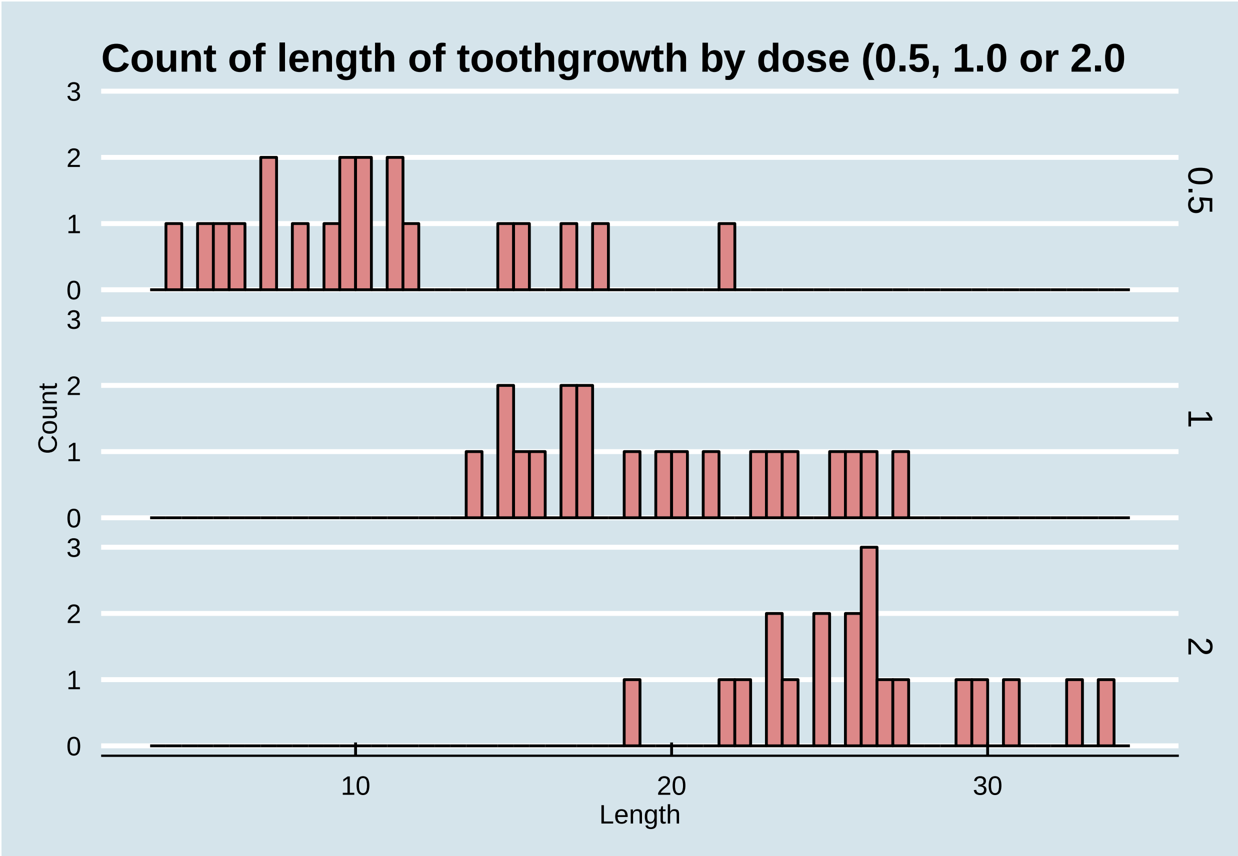
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
sqldf("select count(supp) from tg group by supp")
```

```
##   count(supp)
## 1          30
## 2          30
```

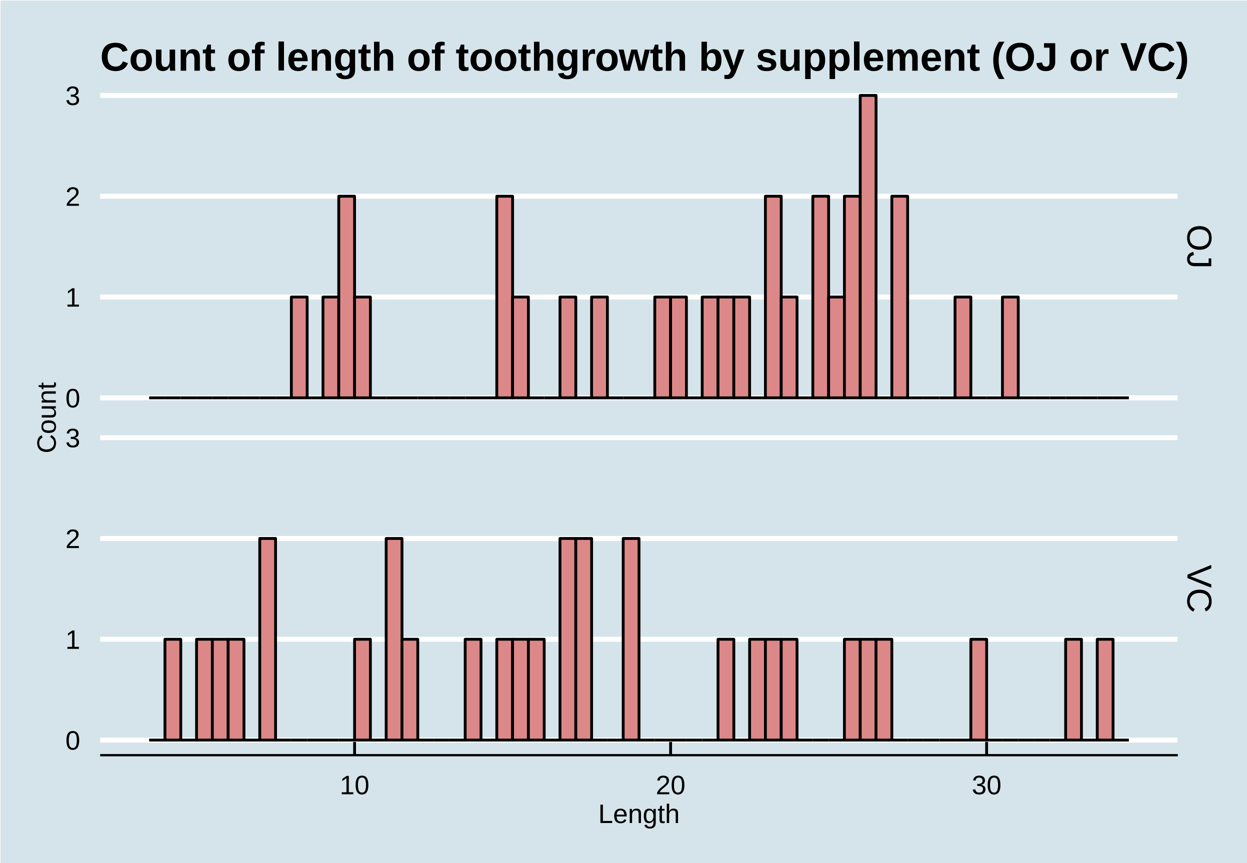
The last result is useful, as we now have the same sample size for each supplement, so we can use the t.test() function and replicate Student's sleep study. Break down the distribution of length by dose:

```
g <- ggplot(data=tg, aes(x=len)) + geom_bar(colour="black", stat="bin", fill="#DD8888", binwidth=0.5)
g <- g + facet_grid(dose ~ .) + xlab("Length") + ylab("Count")
g <- g + ggtitle("Count of length of toothgrowth by dose (0.5, 1.0 or 2.0")
g <- g + theme_economist()
g
```



Break down the distribution of length by supplement:

```
g <- ggplot(data=tg, aes(x=len)) + geom_bar(colour="black", stat="bin", fill="#DD8888", binwidth=0.5) +
g <- g + ggtitle("Count of length of toothgrowth by supplement (OJ or VC)")
g <- g + theme_economist()
g
```



Provide a basic summary of the data:

```
summary(tg)
```

```
##       len        supp         dose      
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00  
##  Mean   :18.8           Mean   :1.17  
##  3rd Qu.:25.3           3rd Qu.:2.00  
##  Max.   :33.9           Max.   :2.00
```

In simple terms, it is a small dataset of two supplements at different dosages and documents tooth growth over these two dimensions. Exploratory data analysis suggests that dosage *could very much* be influential and the supplement *may* be influential.

Let us create a null hypothesis H0: Dosage has no influence on tooth growth. We then use the t.test() function to evaluate length versus dose:

```
t.test(tg$len, tg$dose)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  tg$len and tg$dose
```

```
## t = 17.81, df = 59.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  15.66 19.63
## sample estimates:
## mean of x mean of y
##    18.813     1.167
```

With such a low p-value, the null hypothesis must go, that means we must accept the alternative hypothesis Ha: Dosage does have an influence on tooth growth.

Let us create another null hypothesis H0: The supplement has no influence on tooth growth.

```r
t.test(len ~ supp, data=tg)
```

```
##
##  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
```

The p-value is higher than 0.05 - suggesting this hypothesis holds; just! That means the alternative hypothesis is rejected and we retain the null hypothesis: the supplement has no impact on tooth growth.

Conclusions

The following assumptions were made: 1. We assume data is *roughly* symmetric and mound shaped - something exploratory data analysis suggests. 2. We *assume* the groups are independent - and *know* the sample sizes are equal, therefore we can use Gosset's T-test.

The conclusions is clear: if you want to improve tooth growth, it doesn't matter whether you take "OJ" or "VC" as your supplement, just make sure the dose is large!