Statistical Inference Part 2

THE DATA

The data used for this assignment is the 'ToothGrowth' data: The Effect of Vitamin C on Tooth Growth in Guinea Pigs.

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

The data frame has 60 observations on 3 variables.

1) len: numeric Tooth length
2) supp: factor Supplement type (VC or OJ).
3) dose: numeric Dose in milligrams/day

```
# load data
data(ToothGrowth) #Dataset - The Effect of Vitamin C on Tooth Growth in Guinea Pigs
head(ToothGrowth)
```

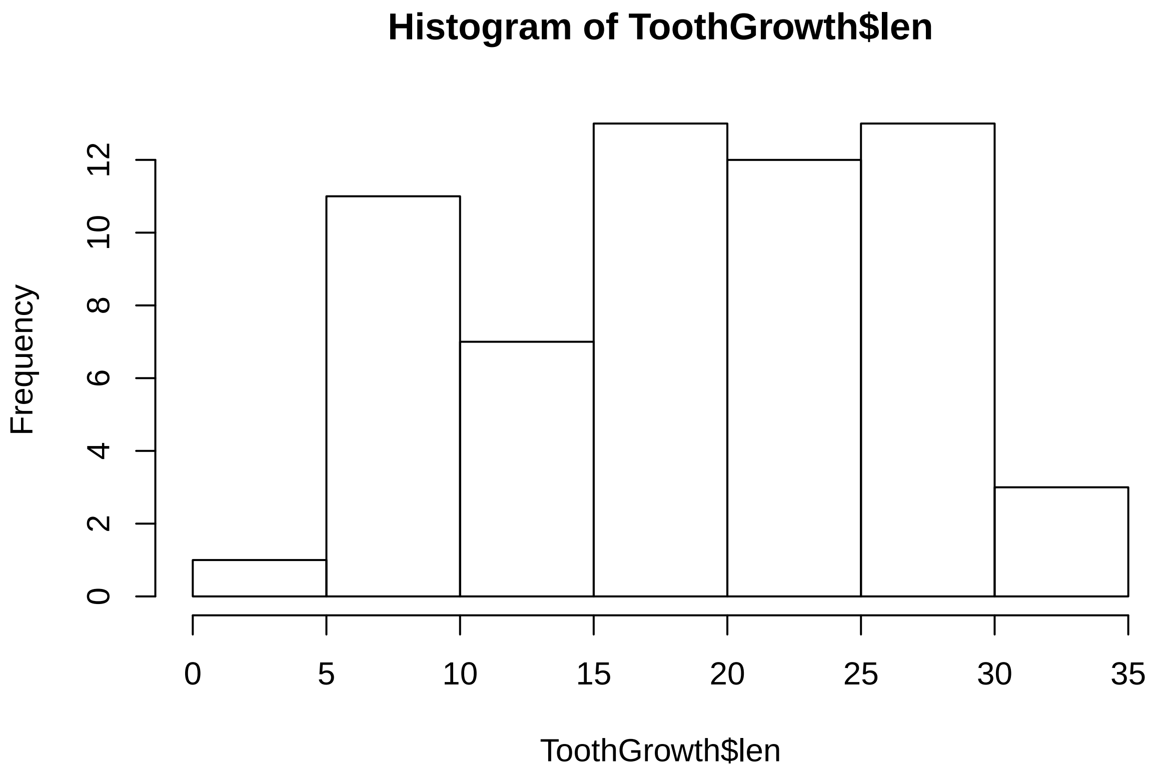
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

The mean toothlength is 18.8 - we'll explore whether this is different by supplement type and dose.

```
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  
```

```
hist(ToothGrowth$len)
```



TOOTH LENGTH BY SUPPLEMENT TYPE

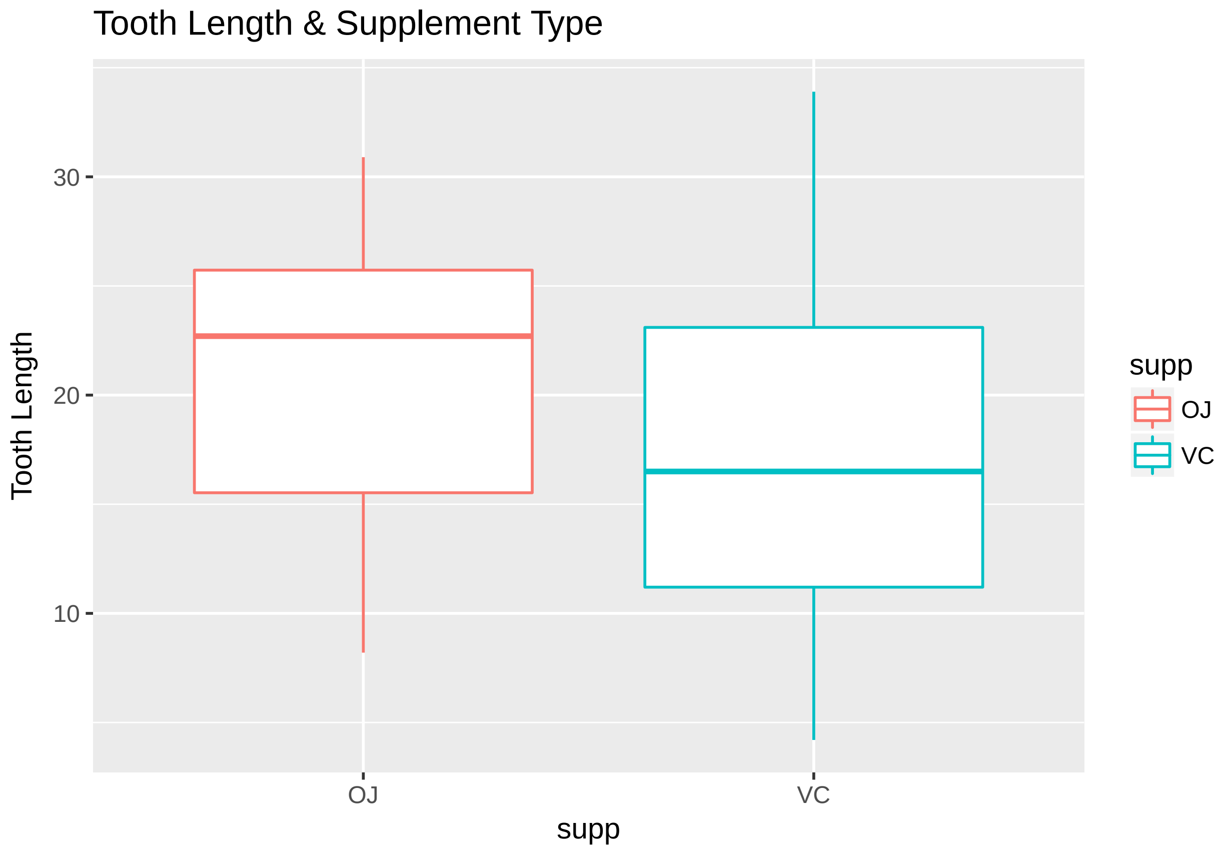
In this sample, the mean tooth length with OJ (20.7) is higher than through VC (17.0). However, the box plot shows more variability in tooth length through the VC deliverability method.

```
tapply(ToothGrowth$len, ToothGrowth$supp, mean)
```

```
##       OJ       VC 
## 20.66333 16.96333
```

```
library(ggplot2)

ggplot(ToothGrowth, aes(x=supp, y=len, colour = supp)) +
  geom_boxplot() +
  labs(x="supp", y="Tooth Length", title="Tooth Length & Supplement Type")
```



** Unpaired Two Sample T-Test: Comparing Tooth Length Mean by Supplement Type **

Let's compute an unpaired two sample t-test to see if there's a significant difference in tooth length between the two delivery methods.

The p value of the test is 0.06, which is higher than the significant alpha = 0.05, so we can't conclude that they are significantly different.

```
t.test(len ~ supp, data = ToothGrowth, var.equal=TRUE)
```

```
## 
##  Two Sample t-test
## 
## data:  len by supp
## t = 1.9153, df = 58, p-value = 0.06039
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1670064  7.5670064
## sample estimates:
## mean in group OJ mean in group VC 
##         20.66333         16.96333
```

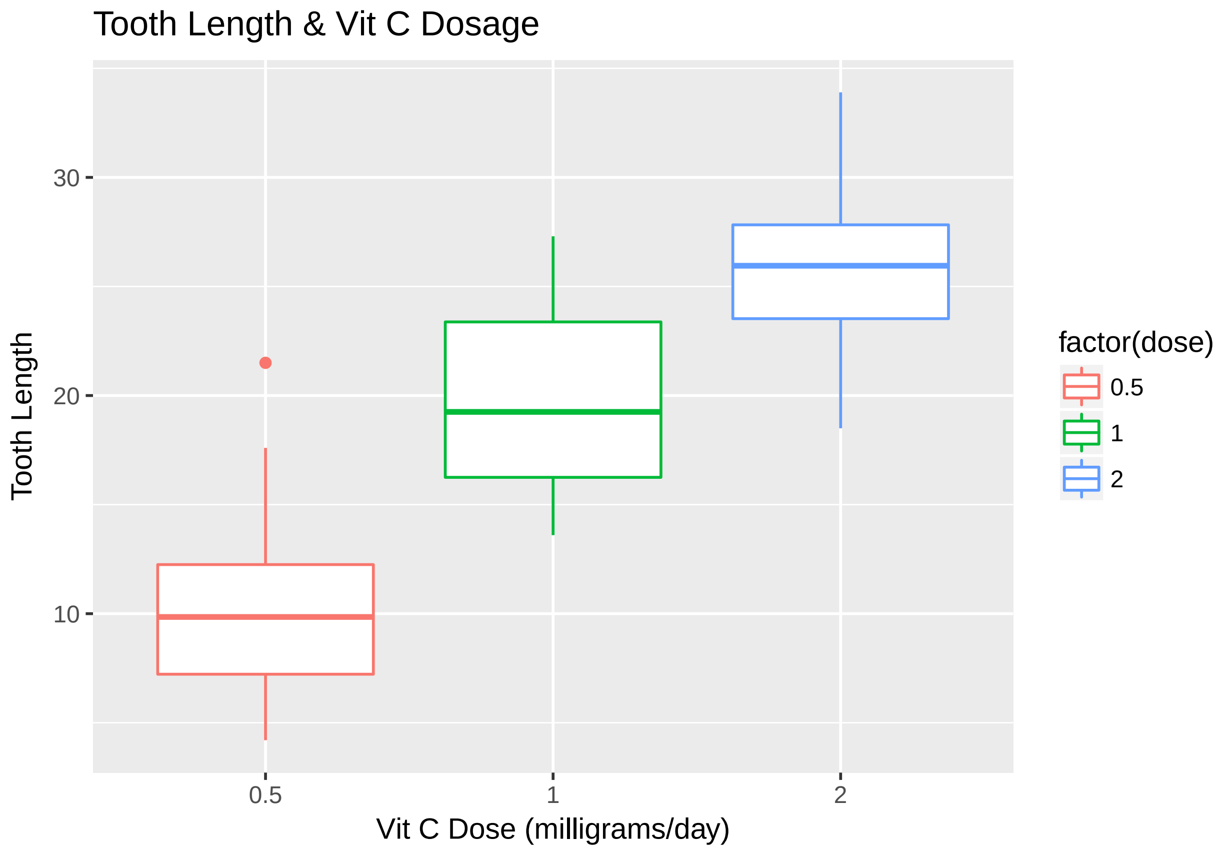
TOOTH LENGTH BY VITAMNIN C DOSE

The box plot shows some very clear differences in tooth length by dose levels.

```
tapply(ToothGrowth$len, ToothGrowth$dose, mean)
```

```
##    0.5      1      2
## 10.605 19.735 26.100
```

```
ggplot(ToothGrowth, aes(x=factor(dose), y=len, colour = factor(dose))) +
  geom_boxplot() +
  labs(x="Vit C Dose (milligrams/day)", y="Tooth Length", title="Tooth Length & Vit C Dosage")
```



** Unpaired Two Sample T-Test: Comparing Tooth Length Mean by Dose Level **

Let's compute an unpaired two sample t-test to see if there's a significant difference in tooth length between the 0.5 and 2 milligrams/day Vitamin C dose levels.

The p value of the test is close to 0, which is below the significant alpha = 0.05, so we can conclude that they are significantly different.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dose <- ToothGrowth %>%
  filter(dose == 0.5 | dose == 2) %>%
  mutate(dose = factor(dose))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
t.test(len ~ dose, data = dose, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 38, p-value = 2.838e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15352 -12.83648
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

** Permutation Test **

Let's also compare these two dose groups doing permutation tests.

First, we'll subset the data for the two groups we're inerested in:

```
sub <- ToothGrowth %>%
  filter(dose == 0.5 | dose == 2) %>%
  select(-supp) %>%
  mutate(dose = as.factor(dose))
```

Then, we will discard the dose group labels and save them in 'group'. We then evaluate the probablity of getting that difference in mean tooth length between the two doses.

It is not probable to see the differences observed, with the difference in the mean = 1 in the permuteations, as opposed to an observed difference of -15.5.

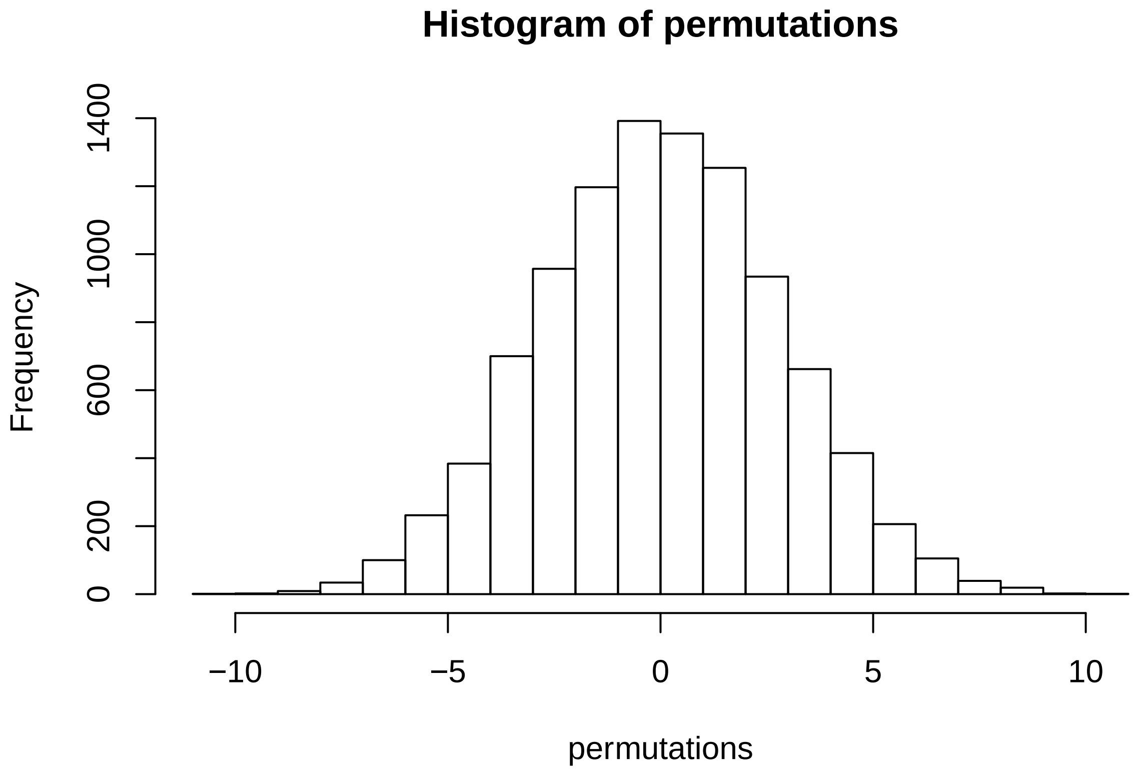
```
y <- sub$len
group <- sub$dose
testStat <- function(w, g) mean(w[g == '0.5']) - mean(w[g == '2'])
observedStat <- testStat(y, group)
permutations <- sapply(1:10000, function(i) testStat(y, sample(group)))
observedStat
```

```
## [1] -15.495
```

```
mean(permutations > observedStat)
```

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

```
hist(permutations)
```


Histogram of permutations
Frequency
0
200
600
1000
1400
−10
−5
0
5
10
permutations