

Statistical Inference Course Project Part 2

Author: Norman Benbrahim

Overview In the second portion of the class, we're going to analyze the ToothGrowth data in the R datasets package.

Question 1

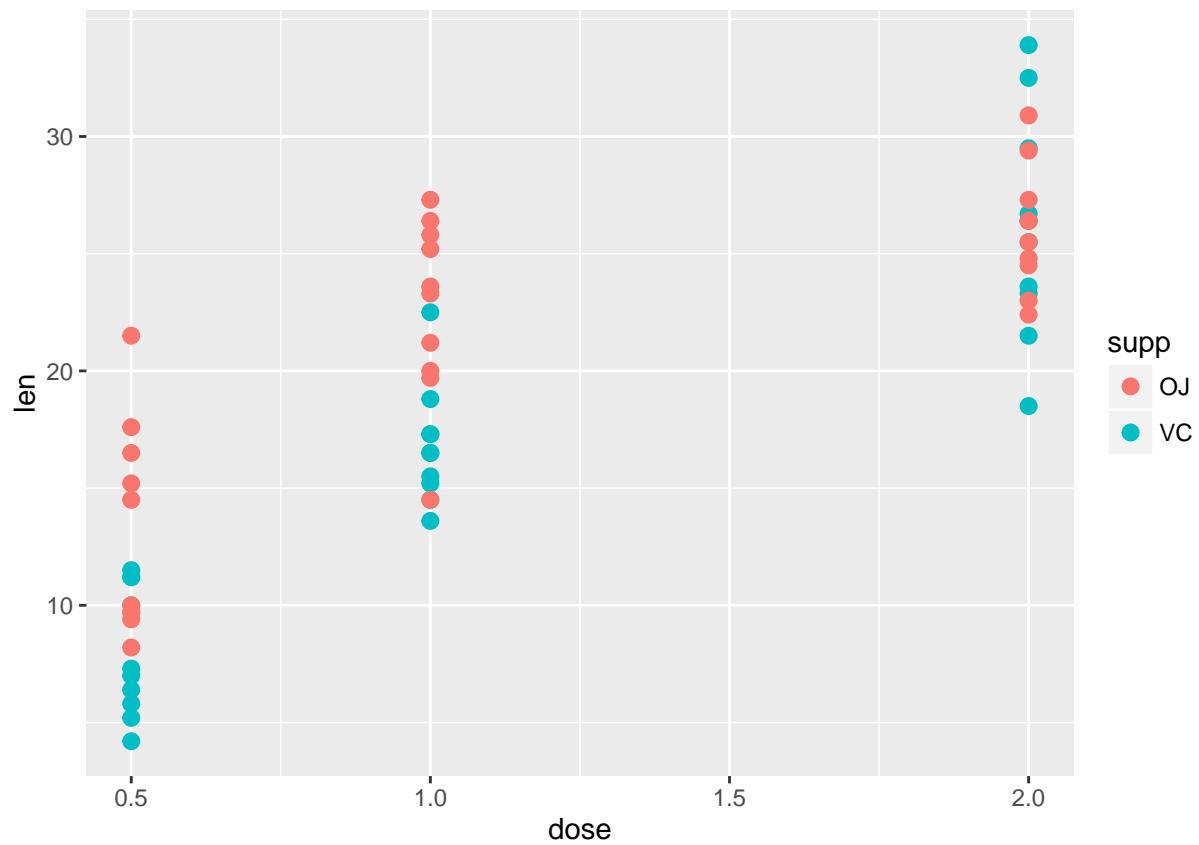
Load the ToothGrowth data and perform some basic exploratory data analyses

```
library(datasets)
library(ggplot2)
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 ...
```

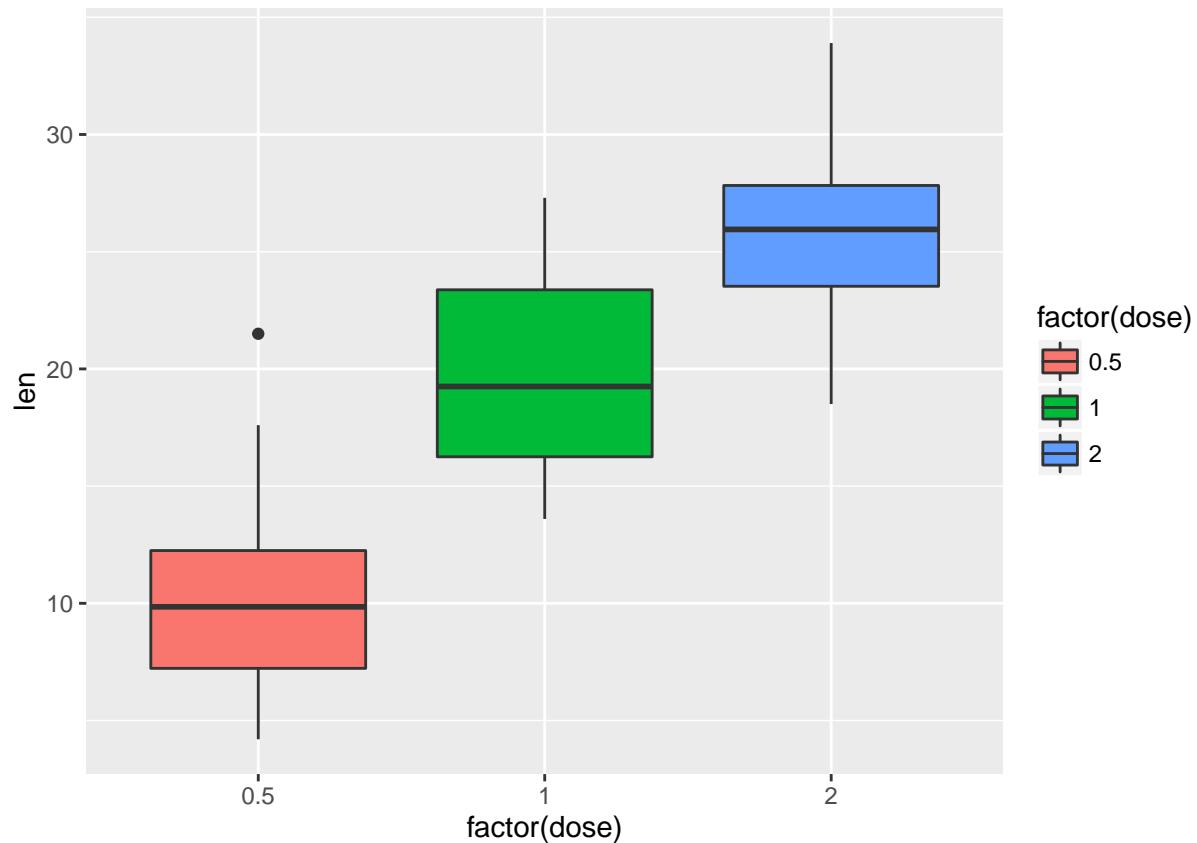
We see the data can be split into two categories based on **supp**, the supplement type. Let's plot the dose in milligrams (dose) vs the tooth length (len) color coded by the supplement type

```
g <- ggplot(ToothGrowth, aes(dose, len))
g <- g + geom_point(aes(colour = supp), size = 2.5)
g
```



Now in a boxplot

```
g <- ggplot(ToothGrowth, aes(factor(dose), len))
g <- g + geom_boxplot(aes(fill = factor(dose)))
g
```



Question 2

Provide a basic summary of the data

R has a summary function that can do this

```
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
```

Question 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)

Let's do a simple t-test

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

```
##
```

```
## Welch Two Sample t-test
##
## data: len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.1710156 7.5710156
## sample estimates:
## mean in group OJ mean in group VC
## 20.66333 16.96333
```

In this case we see the null cannot be rejected because the p-value is higher than an alpha level of 0.05:

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

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

So we can not say with any confidence that supplement type have any impact on tooth growth. Let's run 3 separate t tests now for each dose level. Since we're running multiple tests, we need to correct for multiple comparisons. I will simply do a **Bonferonni correction**, and set the new significance level to α/m where m is the number of tests conducted. In this case $m = 3$, so we get a new alpha $\alpha_b = 0.016$.

```
library(dplyr)

d0.5to1 <- filter(ToothGrowth, dose %in% c(0.5, 1.0))
d1to2 <- filter(ToothGrowth, dose %in% c(1.0, 2.0))
d0.5to2 <- filter(ToothGrowth, dose %in% c(0.5, 2.0))
```

First we check the p values for dose levels 0.5 to 1 and also display their confidence intervals

```
t.test(len ~ dose, data = d0.5to1)$p.value
```

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

```
t.test(len ~ dose, data = d0.5to1)$conf.int[1:2]
```

```
## [1] -11.983781 -6.276219
```

Much less than 0.016, the null can be rejected

Now check the levels 1 to 2

```
t.test(len ~ dose, data = d1to2)$p.value
```

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

```
t.test(len ~ dose, data = d1to2)$conf.int[1:2]
```

```
## [1] -8.996481 -3.733519
```

Also much less than 0.016, the null can be rejected.

Finally, check levels 0.5 to 2

```
t.test(len ~ dose, data = d0.5to2)$p.value
```

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

```
t.test(len ~ dose, data = d0.5to2)$conf.int[1:2]
```

```
## [1] -18.15617 -12.83383
```

Once again much less than 0.016. All the intervals do not contain 0, so our conclusions are firm.

Question 4

State your conclusions and the assumptions needed for your conclusions

The conclusions are that the supplement type have no apparent impact on growth, and increasing the dose levels in all cases leads to increased tooth growth. The assumptions are that the variance in the population are equal (default setting for `t.test`), and the design of the experiment tells us that the choice of guinea pig was randomized by dosage and supplement type. We also assumed multiple comparisons and corrected via Bonferonni correction.