Statistical Inference Example

Linda Wang

January 17, 2016

This report presents the analysis on the dataset ToothGrowth from the R datasets package through a summary, exploratory analyses, confidence interval/hypothesis testing, as well as supplementary text, to compare tooth growth by supp and dose.

Conclusions:

1) When having delivery method controlled, guinea pigs with a higher dose level tend to have longer tooth length;

2) When keeping the dose level unchanged, the type of delivery methods received makes a difference in guinea pigs' tooth growth.

First, we need to load the package and the data:

```
library(datasets)
data(ToothGrowth); head(ToothGrowth,3)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
```

Then we make a summary of the dataframe:

```
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
```

After reading the help file for this dataset, we know the "dose" column is actually a factor with 3 levels, but R treats it as numeric at the moment. So we will do a little of a conversion before the exploratory plots:

```
TG_explore <- ToothGrowth; TG_explore$dose <- factor(TG_explore$dose);
summary(TG_explore)
```

```
##       len        supp     dose   
##  Min.   : 4.20   OJ:30   0.5:20  
##  1st Qu.:13.07   VC:30   1  :20
```

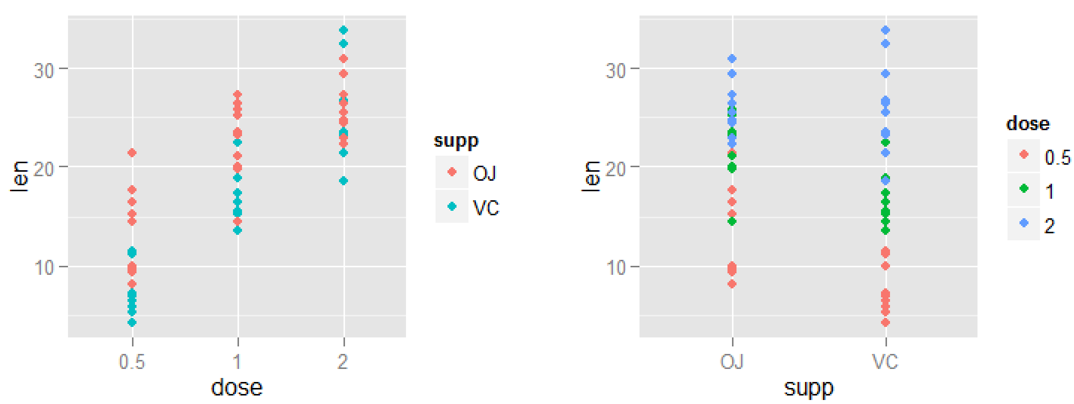
```
##   Median :19.25           2  :20
##   Mean   :18.81
##   3rd Qu.:25.27
##   Max.   :33.90
```

Now the dataset only contains one numeric column, and the other two are factors.

So to perform some very basic exploratory analyses to have a general idea of the dataset, notice that because "supp" and "dose" are the two categorical variables, we can make exploratory plots by looking at the two "supp" groups and/or the three "dose" groups.

Also, because the help file states that the data is "the length of teeth in each of 10 guinea pigs at each of three dose levels with each of two delivery methods", it makes sense to assume the response is paired.

```
library(ggplot2)
library(gridExtra)
p1 <- qplot(dose,len,data=TG_explore,color=supp)
p2 <- qplot(supp,len,data=TG_explore,color=dose)
grid.arrange(p1,p2,ncol=2)
```



And it is especially easy to tell from the second plot that when controlling delivery methods, higher dose levels seems to be related to longer tooth length. The relation between delivery methods and length with dose levels controlled seems to be a little vague from the first plot. Whether or not these guesses are true would be left to the confidence interval/hypothesis testing part of the report.

We are interested in whether the delivery method and/or the dose levels make a difference in the tooth length, and because the sample size is small, we need a t-test rather than a "normal"" test. Here we first subset the responses by "supp" and "dose":

```
library(data.table)
TGdata <- ToothGrowth; TGdata <- data.table(TGdata)
OJ <- TGdata[supp=="OJ",]; VC <- TGdata[supp=="VC",]
OJ_l <- OJ[dose==0.5,len]; OJ_m <- OJ[dose==1,len]; OJ_h <- OJ[dose==2,len]
VC_l <- VC[dose==0.5,len]; VC_m <- VC[dose==1,len]; VC_h <- VC[dose==2,len]
```

We want to verify if it is true that guinea pigs with a higher dose level have a longer tooth length when keeping the delivery method the same. So we preform two group testing with t.test() as the folloiwng: (although they are equivalent when doing inference, we choose to show the p-values for the "OJ" group and confidence intervals for "VC")

```
t.test(OJ_m,OJ_l,alternative="greater",paired=TRUE,conf.level=0.95)$p.value

## [1] 0.00121757

t.test(OJ_h,OJ_m,alternative="greater",paired=TRUE,conf.level=0.95)$p.value

## [1] 0.04191956

t.test(OJ_h,OJ_l,alternative="greater",paired=TRUE,conf.level=0.95)$p.value

## [1] 1.862051e-05

t.test(VC_m,VC_l,alternative="greater",paired=TRUE,conf.level=0.95)$conf.int

## [1] 6.16418     Inf
## attr(,"conf.level")
## [1] 0.95

t.test(VC_h,VC_m,alternative="greater",paired=TRUE,conf.level=0.95)$conf.int

## [1] 6.157075      Inf
## attr(,"conf.level")
## [1] 0.95

t.test(VC_h,VC_l,alternative="greater",paired=TRUE,conf.level=0.95)$conf.int

## [1] 14.76006      Inf
## attr(,"conf.level")
## [1] 0.95
```

The p-values are significant and confidence intervals are systematically above zero. (Conclusions are at the end of the report.)

```
t.test(OJ_l,VC_l,alternative="two.sided",paired=TRUE,conf.level=0.95)$p.value

## [1] 0.01547205

t.test(OJ_m,VC_m,alternative="two.sided",paired=TRUE,conf.level=0.95)$p.value

## [1] 0.008229248

t.test(OJ_h,VC_h,alternative="two.sided",paired=TRUE,conf.level=0.95)$p.value

## [1] 0.9669567
```

We can see the first 2 p-values are less than 0.05, and the last one is much higher, which is consistent with our exploratory plot at the beginning: at dose level equal to 2, there seems to be a lot of "overlapping" of the two delivery groups, and at the other two dose levels the "boundary" looks relatively "clear".

So our conclusions are:
1. When having delivery method controlled, guinea pigs with a higher dose level tend to have longer tooth length;
2. When keeping the dose level unchanged, the type of delivery methods received makes a difference in guinea pigs' tooth growth.

As a reminder, we made the assumptions from the help file that the data is paired, and because the sample size is small, it follows a t-distribution (hence we use t.test() for two group testing). Also, although the difference may be samll, we assume the variances in each group are unequal, and the underlying true population distribution is not heavily skewed.