Statistical Inference Course Project Part 2: Basic Inferential Data Analysis

David Kochar

2017-11-05

Overview

This report will perform exploratory data analysis from R's "ToothGrowth" data set, and also use hypothesis testing to compare tooth growth by supplement type and dose.

Setup

We will first prepare the workspace environment by setting global options, and installing and loading required libraries.

Set Global Options

Establish global options for the report output

```
list.of.packages <- c("knitr")
new.packages <-
list.of.packages[!(list.of.packages %in% installed.packages()[, "Package"])]
if (length(new.packages))
install.packages(new.packages, repos = "http://cran.us.r-project.org")
suppressWarnings (suppressMessages (library (knitr)))
knitr::opts_chunk$set(
fig.width = 8,
fig.height = 4,
fig.path = 'figures/StatisticalInferenceCourseProject/StatisticalInferenceCourseProjectPart1_',
echo = TRUE,
warning = FALSE,
message = FALSE
)
```

Prepare Workspace and Load Libraries

Clear any existing variables from the workspace, set the working directory, and install required libraries if neccessary

```
#Clear variables
rm (list = ls (all = TRUE))
#Get and set working directory
setwd (getwd ())
#Check installed status of requried packages, and install if necessary
list.of.packages <- c("ggplot2", "dplyr", "kableExtra", "statsr")
new.packages <-
list.of.packages[!(list.of.packages %in% installed.packages()[, "Package"])]
if (length(new.packages))
```

```
install.packages(new.packages, repos = "http://cran.us.r-project.org")
suppressWarnings (suppressMessages (library (ggplot2)))
suppressWarnings (suppressMessages (library (dplyr)))
suppressWarnings (suppressMessages (library (kableExtra)))
suppressWarnings (suppressMessages (library (statsr)))
```

Load Data

Load the ToothGrowth data set.

```
data ( ToothGrowth ) #Load the ToothGrowth data set
```

Exploratory Data Analysis

Let's get a feel for our data set by determining the data types and any factor levels.

```
str ( ToothGrowth ) #Describe the ToothGrowth data set
```

```
## '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 ...
```

It appears that "dose" is actually a categorical variable in this context. Let's determine if the values are discrete.

```
unique ( ToothGrowth$dose ) #find unique does values
```

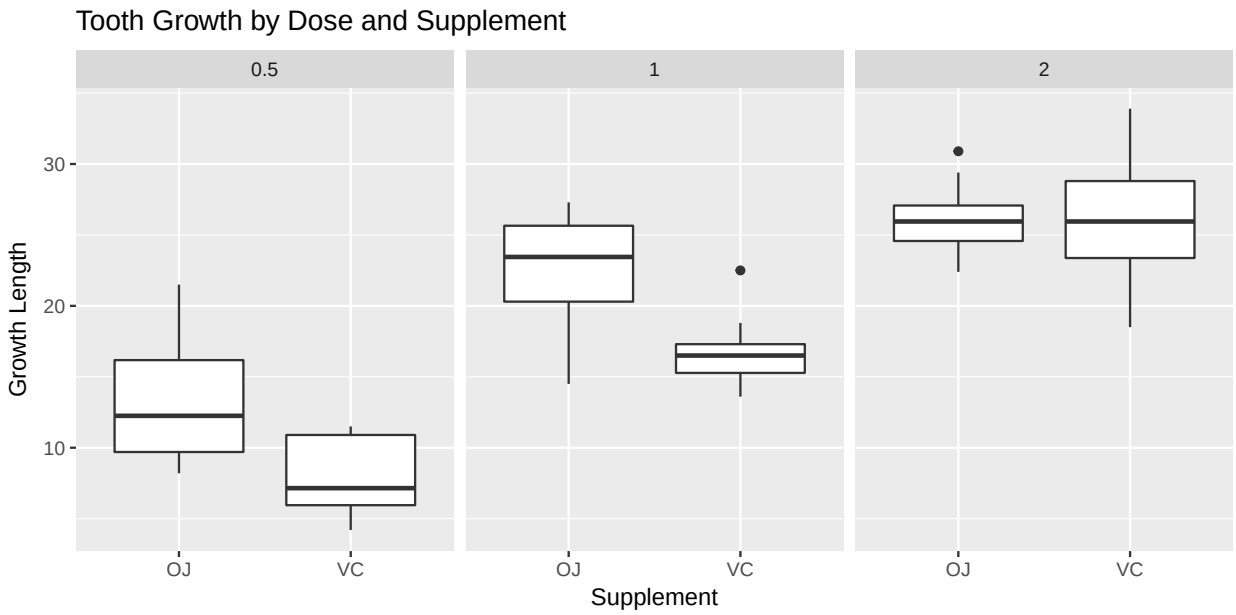
```
## [1] 0.5 1.0 2.0
```

The variable "dose" is indeed categorical in this context, as it has three distinct values. Let's create a new version of ToothGrowth, ToothGrowth2, where we will coerce dose to be a factor variable.

```
ToothGrowth2 <-
  as.data.frame (
  ToothGrowth,
  supp = ToothGrowth$supp,
  dose = as.factor(ToothGrowth$dose),
  len = ToothGrowth$len
  ) #select the columns from ToothGrowth, but coerce dose to a factor variable
```

Let's compare the distributions growth by supplement and dose.

```
ggplot(aes(x = supp, y = len), data = ToothGrowth2) +
  geom_boxplot() +
  facet_wrap( ~ dose) +
  labs(title = "Tooth Growth by Dose and Supplement", x = "Supplement", y = "Growth Length")
```



Growth appears to increase as the dosage increases, and the OJ supplement appears more effective in doeses of 0.5 and 1. The variability in the OJ supplement appears to decrease as the dosage increase, but variability for the VC supplement is reduced only for doses of 1.

Let's summarise the data by supplement and dose.

```
ToothGrowthSummary <- ToothGrowth2 %>%
  group_by (supp, dose) %>%
  summarise (
  MinGrowth = min(len),
  MaxGrowth = max(len),
  AverageGrowth = mean(len),
  MedianGrowth = median(len),
  GrowthIQR = IQR(len)
  ) %>%
  arrange (dose, supp)

 ToothGrowthSummary
```

```
## # A tibble: 6 x 7
## # Groups:   supp [2]
##     supp  dose MinGrowth MaxGrowth AverageGrowth MedianGrowth GrowthIQR
##   <fctr> <dbl>     <dbl>     <dbl>         <dbl>        <dbl>     <dbl>
## 1     OJ   0.5       8.2      21.5         13.23        12.25     6.475
## 2     VC   0.5       4.2      11.5          7.98         7.15     4.950
## 3     OJ   1.0      14.5      27.3         22.70        23.45     5.350
## 4     VC   1.0      13.6      22.5         16.77        16.50     2.025
## 5     OJ   2.0      22.4      30.9         26.06        25.95     2.500
## 6     VC   2.0      18.5      33.9         26.14        25.95     5.425
```

```
 #suppressWarnings (suppressMessages (library (kableExtra)))
 #ToothGrowthSummary %>%
 #kable("html") %>%
 #kable_styling()
```

As we visually suspected, our summary shows growth increases as dosage increases, and the OJ supplement

shows apparently more significant mean growth in the 0.5 and 1.0 doses.

Now let's perform hypothesis testing to confirm what we've observed visually for tooth growth as it relates to supplement and dose.

Hypothesis Testing

We will first test significance for supplement as it relates to tooth growth

Supplement Significance

We want to know if there is a statistically significant difference in tooth growth between the two supplements, OJ and VC. We will perform a two-sided t-test to compare independent means. Our null hypothesis is that there is no difference in the tooth growth length means when comparing supplements, and our alternative hypothesis is that there is a difference.

We will test for significance at each dose level, so we will need to split the data into three sets, one for each dose level.

```
HalfDose <- ToothGrowth2 %>% filter (dose == "0.5")
FullDose <- ToothGrowth2 %>% filter (dose == "1")
DoubleDose <- ToothGrowth2 %>% filter (dose == "2")
```

Dose of 0.5 Significance Test

For our supplement testing, we first test our dose level of 0.5 for significance.

```
t.test(
  len ~ supp,
  data = HalfDose,
  alternative = "two.sided",
  mu = 0,
  paired = FALSE,
  var.equal = FALSE,
  conf.level = 0.95
  )
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

Using a 95% confidence interval, we reject the null hypothesis as our p-value is less than 0.05. Thus, we conclude there is statisically significant difference between the tooth growth means of each supplement at the 0.5 dose level.

Dose of 1 Significance Test

For our supplement testing, we next test our dose level of 1 for significance.

```
t.test(
  len ~ supp,
  data = FullDose,
  alternative = "two.sided",
  mu = 0,
  paired = FALSE,
  var.equal = FALSE,
  conf.level = 0.95
  )
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

Using a 95% confidence interval, we reject the null hypothesis as our p-value is less than 0.05. Thus, we conclude there is statisically significant difference between the tooth growth means of each supplement at the 1 dose level.

Dose of 2 Significance Test

For our supplement testing, we finally test our dose level of 2 for significance.

```
t.test(
  len ~ supp,
  data = DoubleDose,
  alternative = "two.sided",
  mu = 0,
  paired = FALSE,
  var.equal = FALSE,
  conf.level = 0.95
  )
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

Using a 95% confidence interval, we fail to reject the null hypothesis as our p-value is not less than 0.05.

Thus, we conclude there is no statisically significant difference between the tooth growth means of each supplement at the 2 dose level.

Dose Significance

We want to know if there is a statistically significant difference in tooth growth between dose levels. We will perform a two-sided t-test to compare independent means. Our null hypothesis is that there is no difference in the tooth growth length means when comparing dose levels, and our alternative hypothesis is that there is a difference.

Since we have three different dose levels, we will have to compare these dose levels in paired permutations. So, we will compare 0.5 & 1, 0.5 & 2, and 1 & 2. We will need to split our data into three sets for each paired permutation.

```
HalfandFullDose <- ToothGrowth2 %>% filter (dose %in% c("0.5", "1"))
HalfandDoubleDose <-ToothGrowth2 %>% filter (dose %in% c("0.5", "2"))
FullandDoubleDose <- ToothGrowth2 %>% filter (dose %in% c("1", "2"))
```

Compare a Dose of 0.5 to a Dose of 1

Let's first test for sigificance between a dose of 0.5 and a Dose of 1.

```
t.test(
  len ~ dose,
  data = HalfandFullDose,
  alternative = "two.sided",
  mu = 0,
  paired = FALSE,
  var.equal = FALSE,
  conf.level = 0.95
  )
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

Using a 95% confidence interval when comparing a dose of 0.5 to a Dose of 1, we reject the null hypothesis as the p-value is less than 0.05. Thus, there is a statiscally significant difference between tooth growth when comparing these doses.

Compare a Dose of 0.5 to a Dose of 2

Next, let's test for sigificance between a dose of 0.5 and a Dose of 2.

```
t.test(
  len ~ dose,
  data = HalfandDoubleDose,
```

```
  alternative = "two.sided",
  mu = 0,
  paired = FALSE,
  var.equal = FALSE,
  conf.level = 0.95
  )
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

Using a 95% confidence interval when comparing a dose of 0.5 to a Dose of 2, we reject the null hypothesis as the p-value is less than 0.05. Thus, there is a statiscally significant difference between tooth growth when comparing these doses.

Compare a Dose of 1 to a Dose of 2

Finally, let's test for sigificance between a dose of 1 and a Dose of 2.

```
t.test(
  len ~ dose,
  data = FullandDoubleDose,
  alternative = "two.sided",
  mu = 0,
  paired = FALSE,
  var.equal = FALSE,
  conf.level = 0.95
  )
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

Using a 95% confidence interval when comparing a dose of 1 to a Dose of 2, we reject the null hypothesis as the p-value is less than 0.05. Thus, there is a statiscally significant difference between tooth growth when comparing these doses.

Conclusions and Assumptions

Conclusions

When considering the multi-variate effect of dose size when comparing the independent means of supplement type, we see a significantly higher tooth growth effect with the OJ supplement at doses of 0.5 and 1. This is a counter-intuitive effect as one would expect significantly more tooth growth at higher doses, including a dose of 2.

As for dose levels, there is a statisically significant tooth growth effect when comparing each dose level to the other dose levels.

Assumptions

Several assumptions were made for the stated conclusions. These assumptions are:

- The tooth growth variable has a continuous scale
- The tooth growth observations are independent of each other
- The tooth growth variable follows a normal distribution
- The variance between supplements is unequal
- The variance between dose levels is unequal
- The data controls for confounding variables