Course Project1 Part2

doyougnu

July 25, 2015

Overview

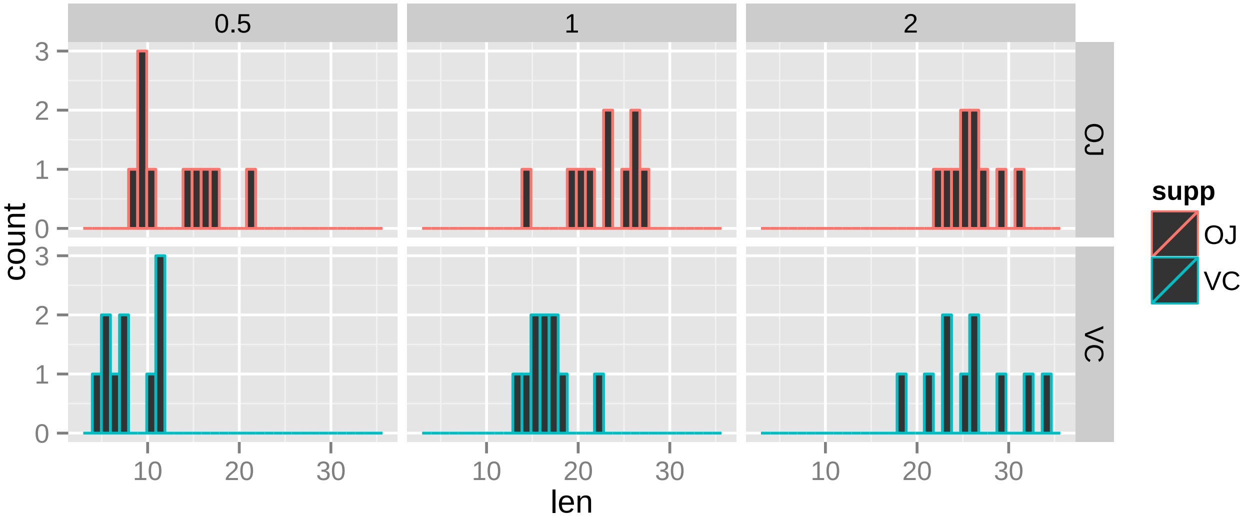
This brief report will explore the ToothGrowth base dataset with R, see C. I. Bliss (1952) The Statistics of Bioassay. Academic Press for more information. It will investigate the central questions of what variable/factors are statistically significant given the data through the use of two-sample student t-testing. This report finds that dose and dose administration method are both statistically significant factors affects the length of guinea pig teeth.

```
#load data
library(datasets); library(ggplot2)
data("ToothGrowth"); df <- ToothGrowth

#lets look at the data
str(df)
```

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

```
#do some quick plots
qplot(data = df, x = len, facets = supp ~ dose, color = supp)
```



Interesting it looks like data under label “OJ” is right shifted compared to data under “VC” for all levels of dose.

Lets look at the confidence intervals for each the data grouped by the supp variable, and by the dose variable

```r
#perform unpaired t.test in R
tResults <- t.test(len ~ supp, data = df, paired = FALSE)
```

One would generally use a paired t-test if the experimental samples are identical through the course of the treatment. Or to relate it to the ToothGrowth dataset: The experiment uses 10 guinea pigs, and administers 3 different Vitamin C dose levels **but splits** the guinea pigs into two groups of Vitamin C application treatmenst. That is one group is "VC" and the other is "OJ", hence it is unpaired because **the same guinea pig is not undergoing both treament OJ and VC**. Looking at the results of the t.test the 95% confidence intervals are -0.1710156, 7.5710156 and the p-value was 0.0606345 which is greater than 0.05, hence we reject the alternate hypothesis (that the means of the two samples is different), and accept the Null hypothesis (that the means of the two samples is the same). Now lets investigate the statistical difference between the dose groups. As we are limited to students t tests I will perform a t.test for each dose group combination, and assume that the data is normally distributed, I will not assume equal variances as t.test defaults to a welch's t-test.

```r
#generate child datasets for VC
dose.0.5_VC <- df[df$dose == 0.5 & df$supp == "VC", ]
dose.1.0_VC <- df[df$dose == 1 & df$supp == "VC", ]
dose.2.0_VC <- df[df$dose == 2 & df$supp == "VC", ]

#generate child datasets for OJ
dose.0.5_OJ <- df[df$dose == 0.5 & df$supp == "OJ", ]
dose.1.0_OJ <- df[df$dose == 1 & df$supp == "OJ", ]
dose.2.0_OJ <- df[df$dose == 2 & df$supp == "OJ", ]

#t.test between 0.5 and 1.0
ttest_helper <- function(df1, df2, compVar = "len", paired = FALSE) {
    a <- t.test(df1[, compVar], df2[, compVar], paired = paired)
    a
}

#perform within supp t.tests
ttest_0.5_1.0_VC <- ttest_helper(dose.0.5_VC, dose.1.0_VC)
ttest_1.0_2.0_VC <- ttest_helper(dose.1.0_VC, dose.2.0_VC)
ttest_2.0_0.5_VC <- ttest_helper(dose.0.5_VC, dose.2.0_VC)

ttest_0.5_1.0_OJ <- ttest_helper(dose.0.5_OJ, dose.1.0_OJ)
ttest_1.0_2.0_OJ <- ttest_helper(dose.1.0_OJ, dose.2.0_OJ)
ttest_2.0_0.5_OJ <- ttest_helper(dose.0.5_OJ, dose.2.0_OJ)

#perform between supp groups t.tests
ttest_0.5_VC_OJ <- ttest_helper(dose.0.5_VC, dose.0.5_OJ)
ttest_1.0_VC_OJ <- ttest_helper(dose.1.0_VC, dose.1.0_OJ)
ttest_2.0_VC_OJ <- ttest_helper(dose.2.0_VC, dose.2.0_OJ)

ttest_0.5_1.0_VC_OJ <- ttest_helper(dose.0.5_VC, dose.1.0_OJ)
ttest_1.0_2.0_VC_OJ <- ttest_helper(dose.1.0_VC, dose.2.0_OJ)
ttest_2.0_0.5_VC_OJ <- ttest_helper(dose.2.0_VC, dose.0.5_OJ)
```

Results

```r
#pack results in a container
ttest_list <- list(ttest_0.5_1.0_VC, ttest_1.0_2.0_VC, ttest_2.0_0.5_VC,
        ttest_0.5_1.0_OJ, ttest_1.0_2.0_OJ, ttest_2.0_0.5_OJ,
        ttest_0.5_VC_OJ, ttest_1.0_VC_OJ, ttest_2.0_VC_OJ,
        ttest_0.5_1.0_VC_OJ, ttest_1.0_2.0_VC_OJ, ttest_2.0_0.5_VC_OJ)
#I couldnt figure how to do this in R, sometimes I wish R was Clojure
ttest_list_names <- c("ttest_0.5_1.0_VC", "ttest_1.0_2.0_VC", "ttest_2.0_0.5_VC",
   "ttest_0.5_1.0_OJ", "ttest_1.0_2.0_OJ", "ttest_2.0_0.5_OJ",
   "ttest_0.5_VC_OJ", "ttest_1.0_VC_OJ", "ttest_2.0_VC_OJ",
   "ttest_0.5_1.0_VC_OJ", "ttest_1.0_2.0_VC_OJ", "ttest_2.0_0.5_VC_OJ")

#extract pvalues and confidence intervals
ttest_list_pvals <- sort(unlist(lapply(ttest_list, function(x) {x[3][1]})))
ttest_list_confs <- lapply(ttest_list, function(x) {x[4]})

#Make data table
library(data.table)
dt <- data.table(T_Test = 0, P_Value = rep(0, 12), Conf_Int1 = 0, Conf_Int2 = 0)
dt[,1] <- ttest_list_names; dt[,2] <- ttest_list_pvals;
dt[,3] <- ttest_list_confs[[1]]; dt[,4] <- ttest_list_confs[[2]]

#convert to real table
library(knitr)
kable(dt)
```

T_Test	P_Value	Conf_Int1	Conf_Int2
ttest_0.5_1.0_VC	0.0000000	-11.265712	-13.054267
ttest_1.0_2.0_VC	0.0000000	-6.314288	-5.685733
ttest_2.0_0.5_VC	0.0000002	-11.265712	-13.054267
ttest_0.5_1.0_OJ	0.0000007	-6.314288	-5.685733
ttest_1.0_2.0_OJ	0.0000013	-11.265712	-13.054267
ttest_2.0_0.5_OJ	0.0000072	-6.314288	-5.685733
ttest_0.5_VC_OJ	0.0000878	-11.265712	-13.054267
ttest_1.0_VC_OJ	0.0000916	-6.314288	-5.685733
ttest_2.0_VC_OJ	0.0010384	-11.265712	-13.054267
ttest_0.5_1.0_VC_OJ	0.0063586	-6.314288	-5.685733
ttest_1.0_2.0_VC_OJ	0.0391951	-11.265712	-13.054267
ttest_2.0_0.5_VC_OJ	0.9638516	-6.314288	-5.685733

Conclusions

Starting from the most significant p-value, we can see that every level of dose, when compared to each other level is statistically significant. In sum, Dose is a statistically significant variable in this experiment when convolved with the “supp” variable, and when deconvloved with the “supp” variable. The only time dose is not statistically significant is in the 1.0 dose VC, 2.0 dose OJ treatment condition and 2.0_VC, 0.5 OJ treatment condtion. It is important to note that when OJ and VC were compared globally they were not statistically different. When compared within each dose treatment group **they are** statistically different.