Statistical Inference Course Project: Part 2

Technophobe01

2015-04-25

Contents

1 Introduction

The purpose of this document is to develop and answer the Coursera Peer Assessment Part 2 of the Coursera Statistical Inference Course Project. Our goal is to analyze the ToothGrowth data in the R datasets package. This reports source code is available via **GitHub**, if you wish to reveiw the markdown formatting in more detail.

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
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)
4. State your conclusions and the assumptions needed for your conclusions.

1.1 Data Set Description

The ToothGroth dataset referenced by this document is derived from an original study *The Journal of Nutrition - Crampton, E.W. (1947)* The study investigated vitamin C intake of natutal (orange juice) and artificial (ascorbic acid) sources by means of a bioassay which measured the teeth growth in guinea pigs. Per the linked study, the length of odontoblast cells in 60 guinea pigs divided in 6 groups of 10 animals which were fed 3 dose levels of Vitamin C (0.5, 1 and 2 mg) with each of two delivery methods (orange juice or ascorbic acid.). The data set contains 60 obeservations on 3 variables.

Variable Name	**'R' Class**	**Variable Description**
len	numeric	Length of Odontoblast (Tooth) in Microns.
supp	factor	Supplement type - Orange Juice or Ascorbic Acid.
dose	numeric	Dose in Milligrams - 0.5mg, 1mg and 2mg.

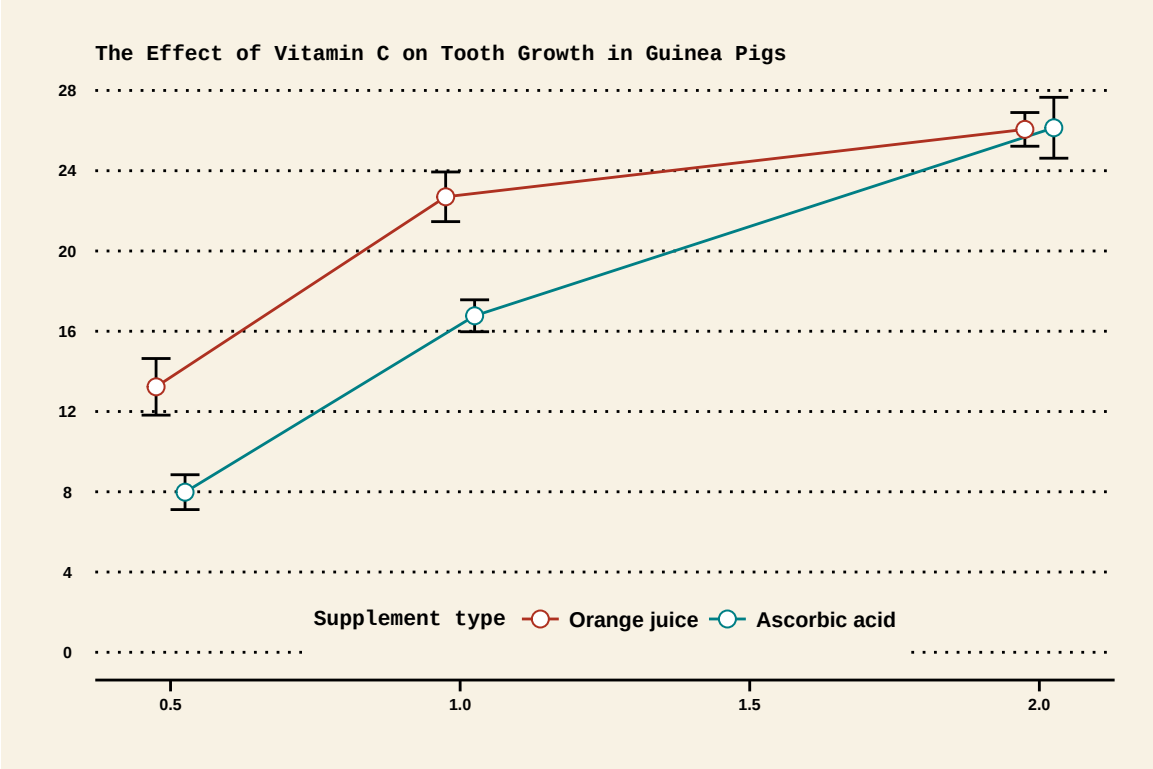

Figure 1: Interaction plot comparing dose and delivery mechanism on tooth growth.

1.1.1 Load the ToothGrowth data and perform some basic exploratory data analyses

The mean plot (Fig. 1) provides a simple yet clear indication that tooth growth increases with the dose of ascorbic acid for both orange juice and vitamin C. For the 0.5mg and 1mg doses, orange juice produced more growth than Vitamin C. For 2mg of ascorbic acid, both delivery methods produced near identical growth.

1.1.2 Provide a basic summary of the data.

A basic statistical summary of the ToothGrowth dataset is provided in Table 2. It details the mean μ, standard deviation σ, varience σ^2, as well as the mediun, min and max values of the dataset organised by supplement and dose.

Table 2: ToothGrowth Summary Statistics

Supplement	dose	N	μ	σ	σ^2	median	min	max
OJ	0.5	10	13.23	4.460	19.889	12.25	8.2	21.5
	1	10	22.70	3.911	15.296	23.45	14.5	27.3
	2	10	26.06	2.655	7.049	25.95	22.4	30.9
	All	30	20.66	6.606	43.633	22.70	8.2	30.9
VC	0.5	10	7.98	2.747	7.544	7.15	4.2	11.5
	1	10	16.77	2.515	6.327	16.50	13.6	22.5
	2	10	26.14	4.798	23.018	25.95	18.5	33.9
	All	30	16.96	8.266	68.327	16.50	4.2	33.9

1.1.3 Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

1.1.3.1 Confidence Interval Tests We wish to assess the significance of the mean differences between the different vitamin C sources. The results found the 0.5 and 1 doses of the 2 different sources were significant with **P Values** of **0.0063586** (.5mg) and **0.0010384** (1mg) respectively. However there was no significant differences at the 2.0mg dose **0.9638516** (2mg).

1.1.3.2 Hypothesis Tests What is our hypotheis? Well, judging from the data above, we would like to confirm that:

1. There is a difference in tooth growth when different **suppliments** are used
2. There is a difference in tooth growth when different **dose** amounts are used?

1.1.3.2.1 Supliment Hypothesis Test If we review the answer below, we can see that the p-value result 0.0025498 is significant, thus we can reject the the null hypothesis. i.e. By rejecting the Null hypothesis we are concluding there are grounds to say that there *is* a relationship. i.e. Orange juice appears to be a better supplement for tooth growth than vitamin C.

	Test.statistic	df	P.Value	Alternate hypothesis
Supplement	3.303	29	0.00255	two.sided

1.1.3.2.2 Dose Hypothesis Test To discover the truth of the second hypothesis, We really need to compare the three doses against each other, two at a time. We can use a paired two-sided t-test for this, creating three groups G1=(0.5, 1), G2=(0.5, 2), G3=(1, 2). i.e. Our intent is to determine if there is a difference in tooth growth when different **dose** amounts are used?

Test.statistic	df	P.Value	Conf.A	Conf.B	Alternate hypothesis
-6.967	19	1.225e-06	-6.387	-11.87	two.sided
-11.29	19	7.19e-10	-12.62	-18.37	two.sided
-4.605	19	0.0001934	-3.472	-9.258	two.sided

Here, we observe an increase in growth with increasing dose (due to small P Value). The 95% confidence interval ranges however are work reviewing in light of figure 1. from **(-12.62, -18.37)** comparing dose 2.0 to 0.5 dose, to **(-3.47, -9.26)** comparing dose 2.0 to 1.0. We should however note that the above test values assume there is no interaction between supplement and dose, which **Figure 1** demonstrates is not true.

1.1.4 State your conclusions and the assumptions needed for your conclusions.

The data and graphs indicate that tooth growth increases with the dose of ascorbic acid for both orange juice and Vitamin C. For the 0.5 and 1mg doses, orange juice produced more tooth growth than Vitamin C. For 2mg of ascorbic acid, both delivery methods produced identical growth.

Thus, we observe that vitamin c intake from a natural sourse such as orange juice as oposed to a synthetic ascorbic acid was shown to be significant at lower doses but no difference was observed at higher doses (2mg) dose. Here we assume that all the animals tested were equivalent between groups. Factors such as health, age and gender are not available, if they were these factors might modify our conclusions.

2 Appendix A: Code and Environment

2.1 Supporting Code

2.1.1 Code Setup

This reports source code is available via **GitHub**, if you wish to understand the markdown formatting in more detail.

2.1.2 Load the ToothGrowth data and perform some basic exploratory data analyses

```
# We generate a mean plot to quickly derive whether their is a material impact
# of one or the other of the supplements. It helped that I read the original
# referenced research. Thanks go to Robert Kabacoff and his "R in Action" worth
# reading.
#
# The errorbars overlapped, so we use position_dodge to move them horizontally
# (Minor Clean up)
pd <- position_dodge(0.1) # move them .05 to the left and right

# summarySE provides the standard deviation, standard error of the mean, and a
# (default 95%) confidence interval
tgc <- summarySE(ToothGrowth, measurevar="len",
                 groupvars=c("supp","dose"))
gp <- ggplot(tgc, aes(x=dose, y=len, colour=supp, group=supp))
gp <- gp + theme_wsj()
gp <- gp + geom_errorbar(aes(ymin=len-se, ymax=len+se),
                         colour="black",
                         width=.1, position=pd)
gp <- gp + geom_line(position=pd)
gp <- gp + geom_point(position=pd, size=3, shape=21,
                      fill="white") # 21 is a filled circle
gp <- gp + xlab("Dose (mg)")
gp <- gp + ylab("Tooth length")
# Legend label, use darker colors
gp <- gp + scale_colour_hue(name="Supplement type",
                    breaks=c("OJ", "VC"),
                    labels=c("Orange juice", "Ascorbic acid"),
                    l=40)  # Use darker colors, lightness=40
gp <- gp + ggtitle("The Effect of Vitamin C on Tooth Growth in Guinea Pigs")
gp <- gp + expand_limits(y=0)                          # Expand y range
gp <- gp + scale_y_continuous(breaks=0:20*4)           # Set tick every 4
gp <- gp + theme(legend.position=c(.5, .1),
                 plot.title = element_text(size=8,
                                           lineheight=.8, face="bold"))
gp <- gp + theme(legend.title = element_text(size = 8, face = 'bold'))
gp <- gp + theme(legend.text = element_text(size = 8, face = 'bold'))
gp <- gp + theme(axis.text = element_text(size = 6, face = 'bold'))

print(gp)
```

2.1.3 Provide a basic summary of the data.

```
# There is a lot going on here. I strongly recommend you read the tables
# documentation and vignette. The package was created by Duncan Murdoch.
# Red[http://cran.r-project.org/web/packages/tables/]
#
# The first two line are important as they set the stage, first we extract
# the suplement data, and then dose (based on the factor supp), from the
# ToothGrowth data frame. We then calaculate the lenth of the vectors we are
# working with (30 for OJ, 30 for VC) i.e. (~ (N = n=1)).
#
# Now we define the columns... A subtle point here is that we use 'len' but we
# ommit a heading for len by specing heading() as empty. We then apply mean to
# values of len to create the mean values column, next sd, next var, next
# mediumm, min, max...

tabular(data=ToothGrowth,
        (Supplement=(supp))*(dose=(as.factor(dose)+1)) ~ (N = n=1) +
                (Format(digits=2)*Heading()*len)*((Heading("$\\mu$")*mean)+
                                                  (Heading("$\\sigma$")*sd)+
                                                  (Heading("$\\sigma^2$")*var)+
                                                  median+min+max))

# Note: You can pass tabular output through the Hmisc::latex() function (Harrell,
# 2011, Harrell et~al., 2011) to produce LATEX output, which when processed by
# pdflatex will produce nice tables in pdf. I prefer to use pandoc for printing
# as this works well across pdf and html output.

pandoc.table(as.matrix(tb), caption = "ToothGrowth Summary Statistics",
             style = "multiline",
             round=justify, split.tables = 100)
```

2.1.4 Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

```
# Perform suplement t.test where we compare length of tooth growth to the supplement used

t1 <- t.test(ToothGrowth$len ~ ToothGrowth$supp, paired=T)
displayTable <- data.frame("Test statistic"=c(t1$statistic),
                           "df"=c(t1$parameter),
                           "P Value"=c(t1$p.value),
                           row.names=c("Supplement"))
displayTable <- cbind(displayTable,"Alternate hypothesis"=c(t1$alternative))
pandoc.table(displayTable,
             style = "multiline",
             split.tables = 100)
```

```
# Perform dose t.test group test where we compare length of tooth growth by group
ToothGrowth.doses_0.5_1.0 <- ToothGrowth %>% filter(dose==0.5 | dose==1)
ToothGrowth.doses_0.5_2.0 <- ToothGrowth %>% filter(dose==0.5 | dose==2)
ToothGrowth.doses_1.0_2.0 <- ToothGrowth %>% filter(dose==1 | dose==2)
```

```
t1 <- t.test(len ~ dose, data=ToothGrowth.doses_0.5_1.0, paired=T)
t2 <- t.test(len ~ dose, data=ToothGrowth.doses_0.5_2.0, paired=T)
t3 <- t.test(len ~ dose, data=ToothGrowth.doses_1.0_2.0, paired=T)

rm(displayTable)
displayTable <- data.frame("Test statistic"=c(t1$statistic, t2$statistic, t3$statistic),
                       "df"=c(t1$parameter,t2$parameter,t3$parameter),
                       "P Value"=c(t1$p.value,t2$p.value,t3$p.value),
                       "Conf A"=c(t1$conf.int[2],t2$conf.int[2],t3$conf.int[2]),
                       "Conf B"=c(t1$conf.int[1],t2$conf.int[1],t3$conf.int[1]))
displayTable <- cbind(displayTable,"Alternate hypothesis"=c(t1$alternative,
                                                            t2$alternative, t3$alternative))
pandoc.table(displayTable,
             style = "multiline",
             split.tables = 100)
```

2.2 References

- *R in Action*
  - By: Robert Kabacoff Publisher: Manning Publications Pub. Date: August 24, 2011, ISBN-10: 1-935182-39-0
- *Mathematical Statistics with Resampling and R*
  - By: Laura Chihara; Tim Hesterberg Publisher: John Wiley & Sons Pub. Date: September 6, 2011 Print ISBN: 978-1-11-02985-5
- *Think Stats, 2nd Edition*
  - By: Allen B. Downey Publisher: O'Reilly Media, Inc. Pub. Date: October 28, 2014 Print ISBN-13: 978-1-4919-0733-7

2.3 Environment

```
##                _
## platform       x86_64-apple-darwin13.4.0
## arch           x86_64
## os             darwin13.4.0
## system         x86_64, darwin13.4.0
## status
## major          3
## minor          2.0
## year           2015
## month          04
## day            16
## svn rev        68180
## language       R
## version.string R version 3.2.0 (2015-04-16)
## nickname       Full of Ingredients
```