Statistical Inference - Project 1

Baskaran Viswanathan

September 22, 2015

Project for the "Statistical Inference" course (Coursera, September 2015)

The `ToothGrowth` data set contains "… *The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid). …*" [Ref.; C. I. Bliss (1952) The Statistics of Bioassay. Academic Press]

Quick exploration of the data set

For the purposes of this analysis, we will convert the dose variable into a factor.

```
data(ToothGrowth)
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

The distributions of tooth length by delivery method and dose levels (see graph in the next page), seem to indicate that for 0.5 and 1 mg, the mean tooth length is greater for those where orange juice was used. For the 2 mg level, no noticeable difference in the mean is observed, but the dispersion for the OJ group seems to be smaller.

Multiple hypothesis testing

To ascertain whether there is a real difference between the groups by dose level and delivery method, we will employ a series of two-sided unpaired t-tests to obtain the confidence intervals and p-values.

The comparison will be done at each dose level, between delivery methods, asumming unequal variance (supported by the distribution graph). The p-values will be adjusted using the (conservative) Bonferroni correction, and the comparative results shown in the table below.

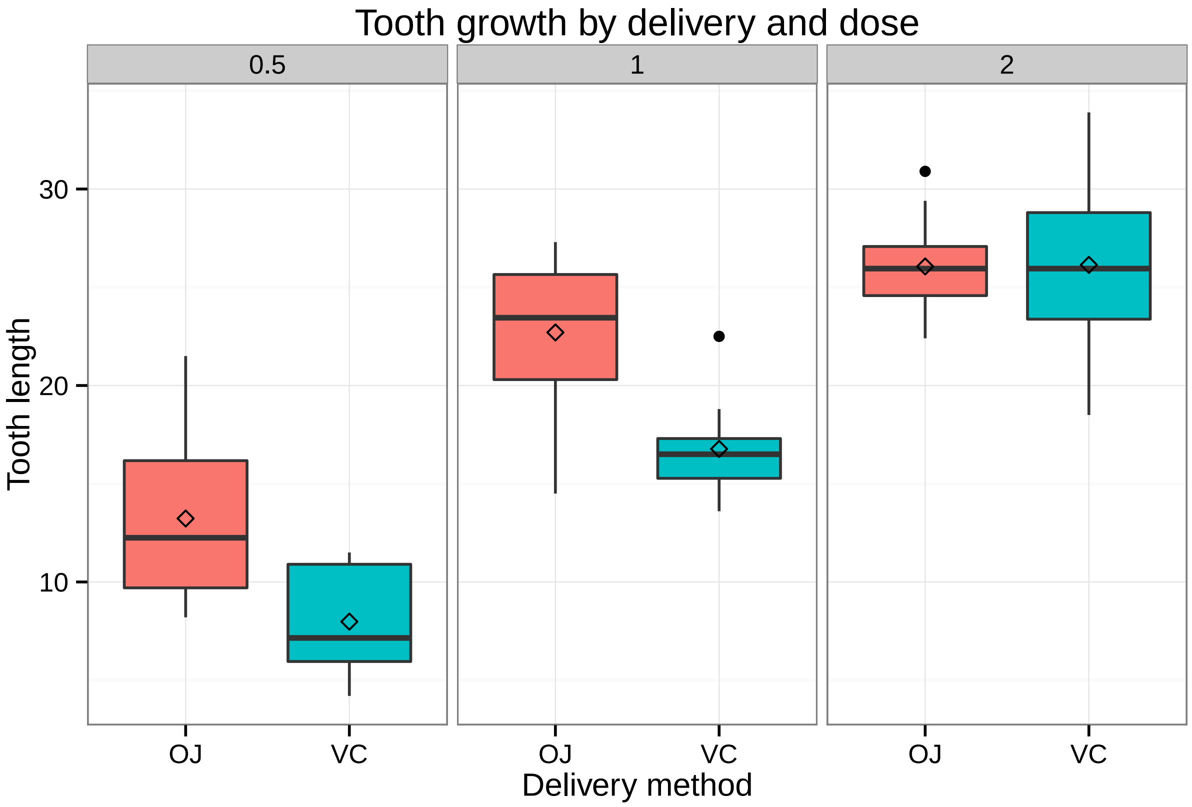
Our null hypothesis in all cases is that there is no difference in the means between the two groups (two-sided test).

```
library(pander)
ts <- lapply(c(.5, 1, 2), function(x) {
    t.test(len ~ supp, data=subset(ToothGrowth, dose==x), paired=FALSE, var.equal=FALSE)
    })
pvals <- c(ts[[1]]$p.value, ts[[2]]$p.value, ts[[3]]$p.value)
stats <- c(ts[[1]]$statistic, ts[[2]]$statistic, ts[[3]]$statistic)
adjp <- p.adjust(pvals, method = "bonferroni")
lls <- sapply(c(ts[[1]]$conf.int[1], ts[[2]]$conf.int[1], ts[[3]]$conf.int[1]), round, 3)
uls <- sapply(c(ts[[1]]$conf.int[2], ts[[2]]$conf.int[2], ts[[3]]$conf.int[2]), round, 3)
df <- data.frame(dose=c(0.5, 1, 2), t=stats, p=pvals, adj=adjp,
                 ci=paste0("[",paste(lls, uls, sep=", "), "]"))
colnames(df) <- c("Dose", "t", "p-value", "adj. p-value", "conf. int.")
pander(df, round=3, split.tables=120,
       caption="*Two-sided comparison of delivery methods by dose*")
```

Table 1: *Two-sided comparison of delivery methods by dose*

Dose	t	p-value	adj. p-value	conf. int.
0.5	3.17	0.006	0.019	[1.719, 8.781]
1	4.033	0.001	0.003	[2.802, 9.058]
2	-0.046	0.964	1	[-3.798, 3.638]

```r
library(ggplot2)
ggplot(data=ToothGrowth, aes(y=len, x=supp, fill=supp)) + geom_boxplot() +
    facet_wrap(~ dose, ncol=3) + ylab("Tooth length") + xlab("Delivery method") +
    ggtitle("Tooth growth by delivery and dose") +
    stat_summary(fun.y=mean, geom="point", shape=5, size=2) +
    theme_bw()+ theme(legend.position = "none")
```



Conclusions

- At the 0.5 and 1 mg dose levels, there is a statistically significant difference between the means of the OJ and VC groups. The adjusted p-values are significant at the $\alpha = 0.05$ level, and the 95% confidence intervals do not include zero.
- For the 2 mg dose level, we fail to reject the null hypothesis, the adjusted p-value is much greater than 0.5, and the 95% confidence interval includes zero. So, it seems that at that Vitamin C level, there is no significative influence of the delivery method on tooth growth in guinea pigs.
- Because the effect size is very small for the 2 mg level, to be able to detect a significative difference, we would need a much bigger sample size (n=10 does not give the test enough power)

```
s <- subset(ToothGrowth, dose==2)
d <- split(s, s$supp)
n1 <- 10; m1 <- mean(d$OJ$len); s1 <- sd(d$OJ$len)
n2 <- 10; m2 <- mean(d$VC$len); s2 <- sd(d$VC$len)
pooled_sd <-  sqrt( ((n1 - 1) * s1^2 + (n2-1) * s1^2) / (n1 + n2-2))
cat("effect size:", round((m2 - m1)/pooled_sd, 3), "\nestimated sample size:",
    round(power.t.test(power=0.9, delta = m1 - m2, sd=pooled_sd)$n,0))
```

```
## effect size: 0.03
## estimated sample size: 23148
```