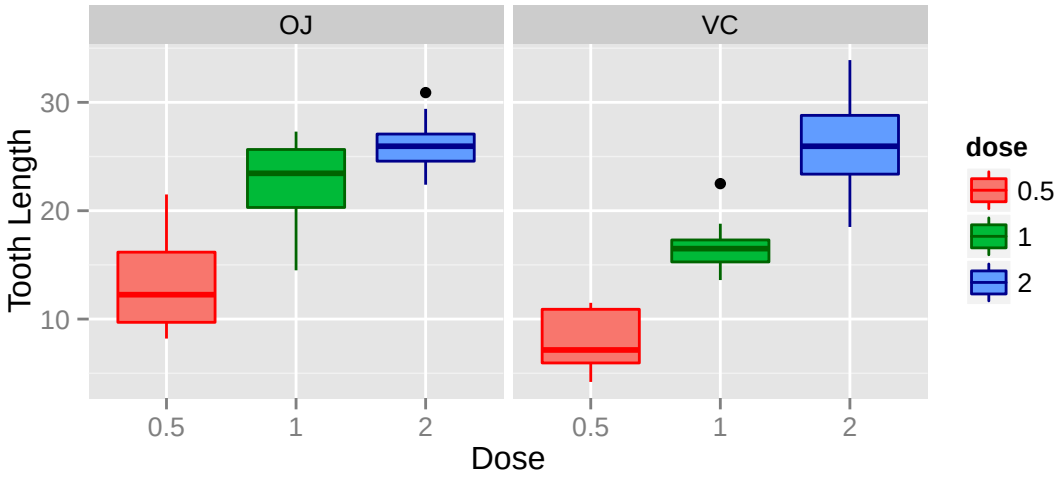
Statistical Inference Course Project

T-StrawClown

Overview

We're going to analyze the ToothGrowth data in the R datasets package to compare tooth growth by supplement type and dosage.
The response of the ToothGrowth dataset is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC). First let's take a look at data in the dataset.

Picture 1: Tooth Growth Data

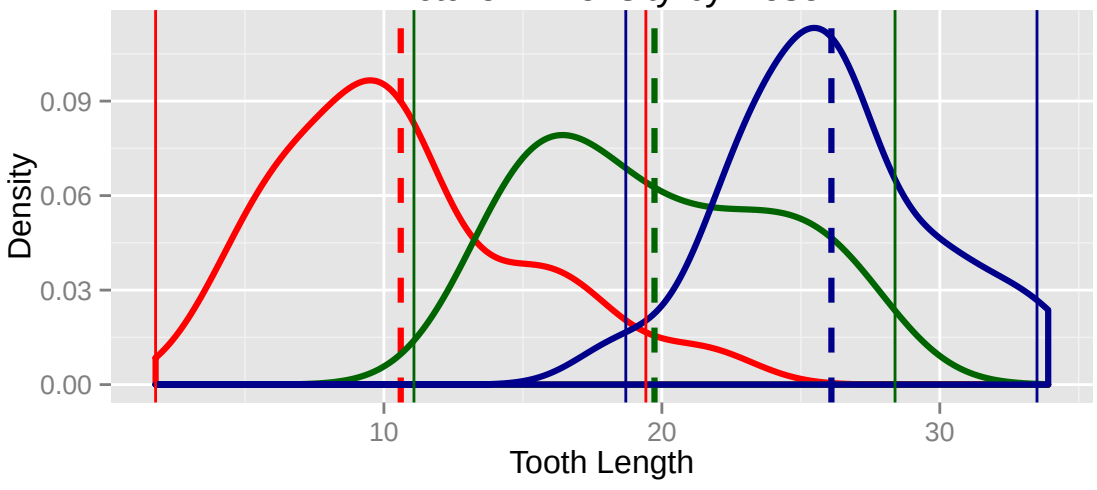


The first impression is that tooth length depends on the dose. So we're going to investigate it.

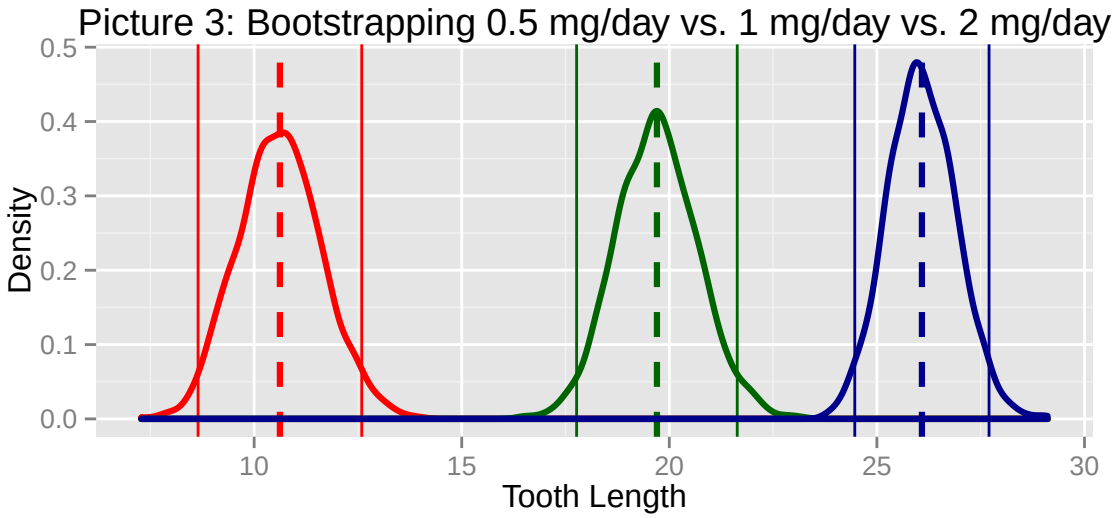
Effect of Dose

Let's take a look at density chart. Red color represents 0.5 mg/day dose, green - 1 mg/day and blue - 2 mg/day.

Picture 2: Density by Dose



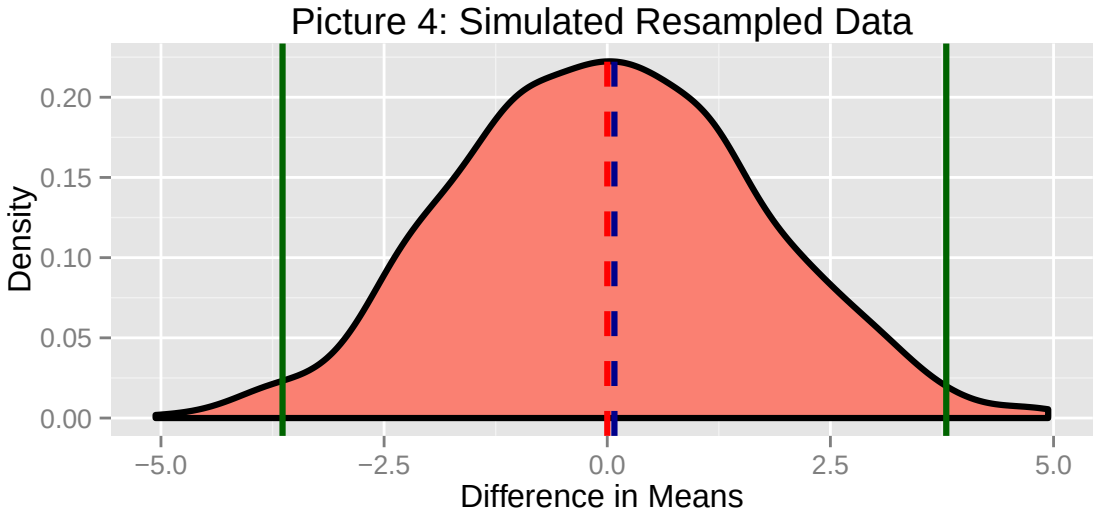
Vertical dashed lines represent means of respective dosage and vertical solid lines represent 0.025 and 0.975 quantiles of normal distribution with mean and variance exactly the same as in the ToothGrowth. Notice that 0.025 quantile of green and blue distributions are in between 0.025 and 0.975 quantiles of red distribution, so we can't rule out the possibility that actual means of population with respect to dosage of vitamin C are actually the same. In other words there is a reasonable doubt that dosage has no influence on length of tooth.
We're going to use the bootstrap principle to evaluate true means of each dosage group and construct confidence intervals for these means. We're going to simulate 1000 samples of size 20 with replacement from given data for each of the doses. This is how simulated data looks like. Again color represent dose (red for 0.5 mg/day, green - 1 mg/day, blue - 2 mg/day), vertical dashed lines represent mean of each group and vertical solid lines represent confidence intervals for means of normally distributed data.



Looks good. Confidence intervals for means don't overlap anymore so now we can be sure that true means of population are dependent on the dose - 1 md/day is better than 0.5 md/day and 2 mg/day is better than 1 mg/day.

Effect of Supplement Type

So now that we know that dose does matter - the more vitamin C is fed to guinea pigs the better are results. Let's verify if delivery methods (orange juice vs. ascorbic acid) has any effect. Since we already know that 2 mg/day is the best option, we're going to use only data of those guinea pigs where 2 mg/day dose was applied.
Again we're going to do 1000 simulations of size 20 (20 guinea pigs have been delivered 2 mg/day dose of vitamin C), but this time we're going to use resampling technique by mixing delivery method (supplement type) and then validating if difference in means of 2 groups in the ToothGrowth dataset is different compared to simulated data. We're going to define our null hypothesis as "delivery method has no influence on the length of tooth, thus difference in means of initial and simulated data is 0" and will try to reject it in favor of alternative hypothesis, saying that difference in means is significant. Let's take a look at distribution of simulated data.



The red dashed line represents difference in means of simulated data and the blue dashed line represents difference in means of initial data. Green vertical lines represent confidence interval of t.test, testing our null hypothesis. Here are results of t-test:

```
##
##  Welch Two Sample t-test
##
## data:  subset(dose2, supp == "VC")$len and subset(dose2, supp == "OJ")$len
## 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:
##  -3.63807  3.79807
## sample estimates:
## mean of x mean of y
##     26.14     26.06
```

And this means that we fail to reject null hypothesis, thus delivery method (orange juice vs. ascorbic acid) doesn't really make any significant difference in respect to the length of tooth.
It is interesting to note that for 1 mg/day dose the story is completely different and if we would repeat the test for that group of guinea pigs, we would reject null hypothesis in favor of alternative. For that group delivery method actually matters and giving orange juice to guinea pigs is a better way to stimulate tooth growth, which isn't surprising - orange juice is always better than some acid :)

```
##
##  Welch Two Sample t-test
##
## data:  subset(dose1, supp == "VC")$len and subset(dose1, supp == "OJ")$len
## 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:
##  -9.057852 -2.802148
## sample estimates:
## mean of x mean of y
##     16.77     22.70
```

Let's also calculate the power of rejecting null hypothesis of one sided t-test with two groups using pooled standard deviation when it actually is false for this group of guinea pigs.

```
##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 5.93
##              sd = 3.288034
##       sig.level = 0.05
##           power = 0.9871305
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

Pretty good, 0.9871 probability of rejecting null hypothesis when it is false.

Conclusions

In real life it would be difficult to rely on results of such a small dataset, but from what we have we can conclude: in case you need your guinea pigs to have long teeth, feed them with 2 mg/day of vitamin C whichever way they prefer (or is cheaper) - orange juice or ascorbic acid.

Supplement 1: R code

I'm sorry I couldn't squeeze into 3 pages, but I want it to be as readable as possible.

```
library(ggplot2)
data("ToothGrowth")
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
g1 <- ggplot(data = ToothGrowth, aes(x = dose, y = len, fill = dose)) +
        geom_boxplot(colour = c("red", "darkgreen", "darkblue")) +
        facet_grid(. ~ supp) +
        ggtitle("Picture 1: Tooth Growth Data by Type of Supplement") +
        ylab("Tooth Length") +
        xlab("Dose")

#bootstrapping
dose05 <- subset(ToothGrowth, dose == "0.5")
dose1 <- subset(ToothGrowth, dose == "1")
dose2 <- subset(ToothGrowth, dose == "2")

n05 <- dim(dose05)[1]
n1 <- dim(dose1)[1]
n2 <- dim(dose2)[1]
sims <- 1000
set.seed(42)
sim_data05 <- matrix(data = sample(dose05$len,
                                   n05 * sims,
                                   replace = TRUE),
                     nrow = sims,
                     ncol = n05)
means05 <- apply(sim_data05, 1, mean)

set.seed(42)
sim_data1 <- matrix(data = sample(dose1$len,
                                  n1 * sims,
                                  replace = TRUE),
                    nrow = sims,
                    ncol = n1)
means1 <- apply(sim_data1, 1, mean)

set.seed(42)
sim_data2 <- matrix(data = sample(dose2$len,
                                  n2 * sims,
                                  replace = TRUE),
                    nrow = sims,
                    ncol = n2)
means2 <- apply(sim_data2, 1, mean)

g2 <- ggplot() +
        ggtitle("Picture 2: Density by Dose") +
        xlab("Tooth Length") +
        ylab("Density") +
        geom_density(data = data.frame(x = dose05$len),
                     aes(x = x),
                     color = "red",
                     size = 1.1) +
        geom_vline(data = data.frame(x = dose05$len),
                   xintercept = mean(dose05$len),
                   linetype = 2,
```

```
                    color = "red",
                    size = 1.1) +
        geom_density(data = data.frame(x = dose1$len),
                     aes(x = x),
                     color = "darkgreen",
                     size = 1.1) +
        geom_vline(data = data.frame(x = dose1$len),
                    xintercept = mean(dose1$len),
                    linetype = 2,
                    color = "darkgreen",
                    size = 1.1) +
        geom_density(data = data.frame(x = dose2$len),
                     aes(x = x),
                     color = "darkblue",
                     size = 1.1) +
        geom_vline(data = data.frame(x = dose2$len),
                    xintercept = mean(dose2$len),
                    linetype = 2,
                    color = "darkblue",
                    size = 1.1) +
        geom_vline(xintercept = mean(dose05$len) + c(-1, 1) *
                                qnorm(.975) * sd(dose05$len),
                    #xintercept = t.test(dose05$len)$conf,
                    linetype = 1,
                    color = "red",
                    width = 1.1) +
        geom_vline(xintercept = mean(dose1$len) + c(-1, 1) *
                                qnorm(.975) * sd(dose1$len),
                    #xintercept = t.test(dose1$len)$conf,
                    linetype = 1,
                    color = "darkgreen",
                    width = 1.1) +
        geom_vline(xintercept = mean(dose2$len) + c(-1, 1) *
                                qnorm(.975) * sd(dose2$len),
                    #xintercept = t.test(dose2$len)$conf,
                    linetype = 1,
                    color = "darkblue",
                    width = 1.1)
g3 <- ggplot() +
        ggtitle("Picture 2: Bootstraping 0.5 mg/day dose vs. 2 mg/day dose") +
        xlab("Mean of Sample Data") +
        ylab("Density") +
        geom_density(data = data.frame(x = means05),
                       aes(x = x),
                     color = "red",
                     size = 1.1) +
        geom_vline(data = data.frame(x = means05),
                    xintercept = mean(means05),
                    linetype = 2,
                    color = "red",
                    size = 1.1) +
        geom_vline(xintercept = mean(means05) + c(-1, 1) *
                                qnorm(.975) * sd(means05),
                    linetype = 1,
                    color = "red",
                    width = 1.1) +
        geom_density(data = data.frame(x = means1),
                     aes(x = x),
```

```
                    color = "darkgreen",
                    size = 1.1) +
        geom_vline(data = data.frame(x = means1),
                   xintercept = mean(means1),
                   linetype = 2,
                   color = "darkgreen",
                   size = 1.1) +
        geom_vline(xintercept = mean(means1) + c(-1, 1) *
                               qnorm(.975) * sd(means1),
                   linetype = 1,
                   color = "darkgreen",
                   width = 1.1) +
        geom_density(data = data.frame(x = means2),
                     aes(x = x),
                     color = "darkblue",
                     size = 1.1) +
        geom_vline(data = data.frame(x = means2),
                   xintercept = mean(means2),
                   linetype = 2,
                   color = "darkblue",
                   size = 1.1) +
        geom_vline(xintercept = mean(means2) + c(-1, 1) *
                               qnorm(.975) * sd(means2),
                   linetype = 1,
                   color = "darkblue",
                   width = 1.1)

# resampling
dose2$supp <- as.character(dose2$supp)
means_diff <- function(v_len, v_supp)
                mean(v_len[v_supp == "VC"]) -
                mean(v_len[v_supp == "OJ"])
resampled_means <- sapply(1:sims,
                          function(i) means_diff(
                                  v_len = dose2$len,
                                  v_supp = sample(dose2$supp)))

g5 <- ggplot(data = data.frame(x = resampled_means),
             aes(x = x)) +
        ggtitle("Picture 4: Simulated Resampled Data") +
        xlab("Difference in Means") +
        ylab("Density") +
        geom_density(color = "black",
                     size = 1.1,
                     fill = "salmon") +
        geom_vline(xintercept = mean(resampled_means),
                   colour = "red",
                   linetype = 2,
                   size = 1.1) +
        geom_vline(xintercept = means_diff(v_len = dose2$len,
                                           v_supp = dose2$supp),
                   color = "darkblue",
                   linetype = 2,
                   size = 1.1) +
        geom_vline(xintercept = t.test(subset(dose2, supp == "VC")$len,
                                       subset(dose2, supp == "OJ")$len,
                                       paired = FALSE,
                                       var.equal = FALSE)$conf.int,
```

```
                    linetype = 1,
                    colour = "darkgreen",
                    size = 1.1)

ttest_diff2 <- t.test(subset(dose2, supp == "VC")$len,
                      subset(dose2, supp == "OJ")$len,
                      paired = FALSE,
                      var.equal = FALSE)

dose1$supp <- as.character(dose1$supp)
ttest_diff1 <- t.test(subset(dose1, supp == "VC")$len,
                      subset(dose1, supp == "OJ")$len,
                      paired = FALSE,
                      var.equal = FALSE)

n1 <- dim(subset(dose1, supp == "VC"))[1]
n2 <- dim(subset(dose1, supp == "OJ"))[1]
sp <- sqrt( ((n1 - 1) * sd(subset(dose1, supp == "VC")$len)^2 +
             (n2 - 1) * sd(subset(dose1, supp == "OJ")$len)^2) / (n1 + n2-2))
md <- mean(subset(dose1, supp == "OJ")$len) -
      mean(subset(dose1, supp == "VC")$len)
semd <- sp * sqrt(1 / n1 + 1/n2)
pttest_diff1 <- power.t.test(n = 10,
                             delta = md,
                             type = "two.sample",
                             alt = "one.sided",
                             sd = sp)
```