Inferential_Tooth_Growth

JeremiShane

11/24/2017

Overview

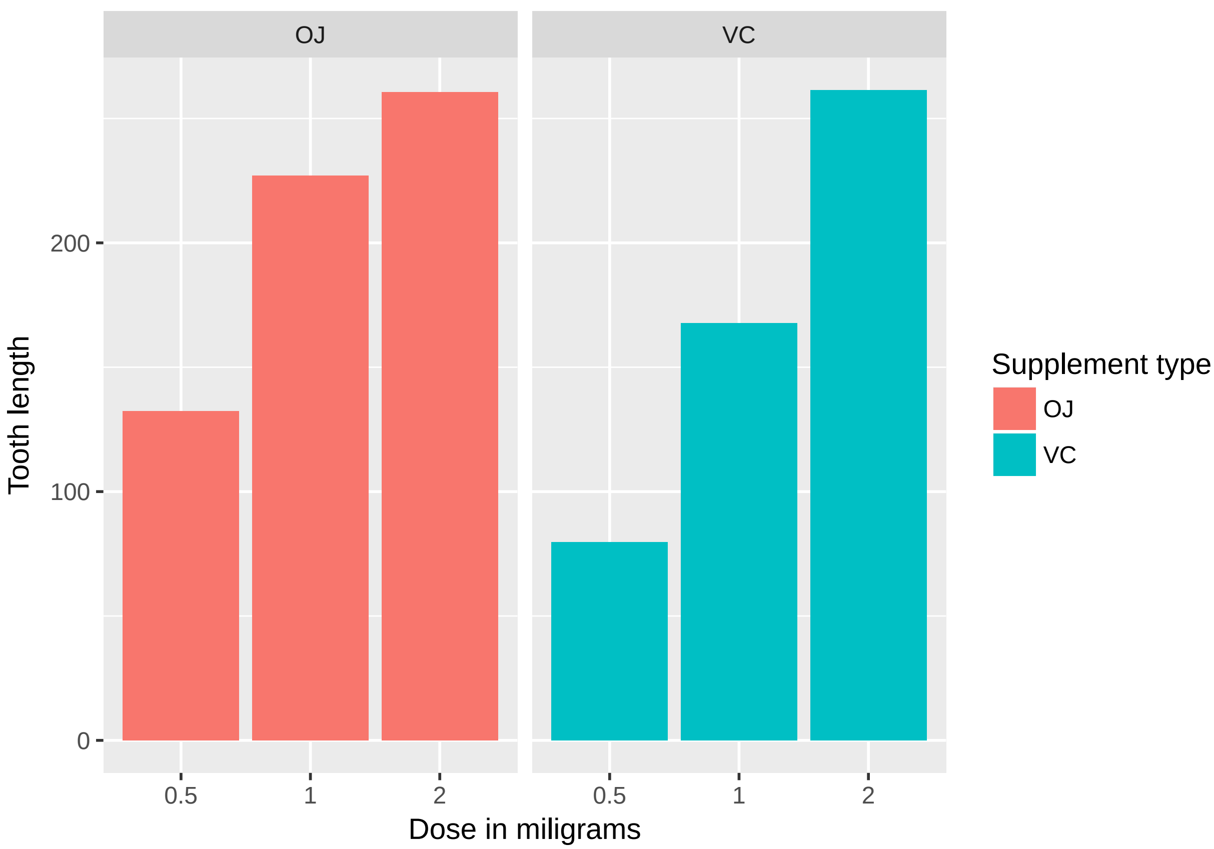
We will analyze the ToothGrowth data in the R datasets package to compare tooth growth by supp and dose.

- len : Tooth length
- supp : Supplement type (VC or OJ)
- dose : Dose in milligrams

The data is set of observations of teeth from 60 guinea pigs. There are 3 dose levels of Vitamin C (0.5, 1 and 2 mg), and there are 2 delivery methods (orange juice or ascorbic acid supplement).

source: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html

```
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
        geom_bar(stat="identity",) +
        facet_grid(. ~ supp) +
        xlab("Dose in miligrams") +
        ylab("Tooth length") +
        guides(fill=guide_legend(title="Supplement type"))
```



From the chart above there appears to be a positive correlation between the tooth length and the dose levels of Vitamin C, for both delivery methods. **It does appear that the lower dosages (<=1mg) of Orange**

Juice have a higher impact than the lower dosages of the Vitamin C supplement. Let's test this hypothsis.

We will use the t.test() in R. The assumption for the student's t test is that both groups are sampled from normal distributions with equal variances. There is also a widely used modification of the t-test, known as Welch's t-test that adjusts the number of degrees of freedom when the variances are thought not to be equal to each other.

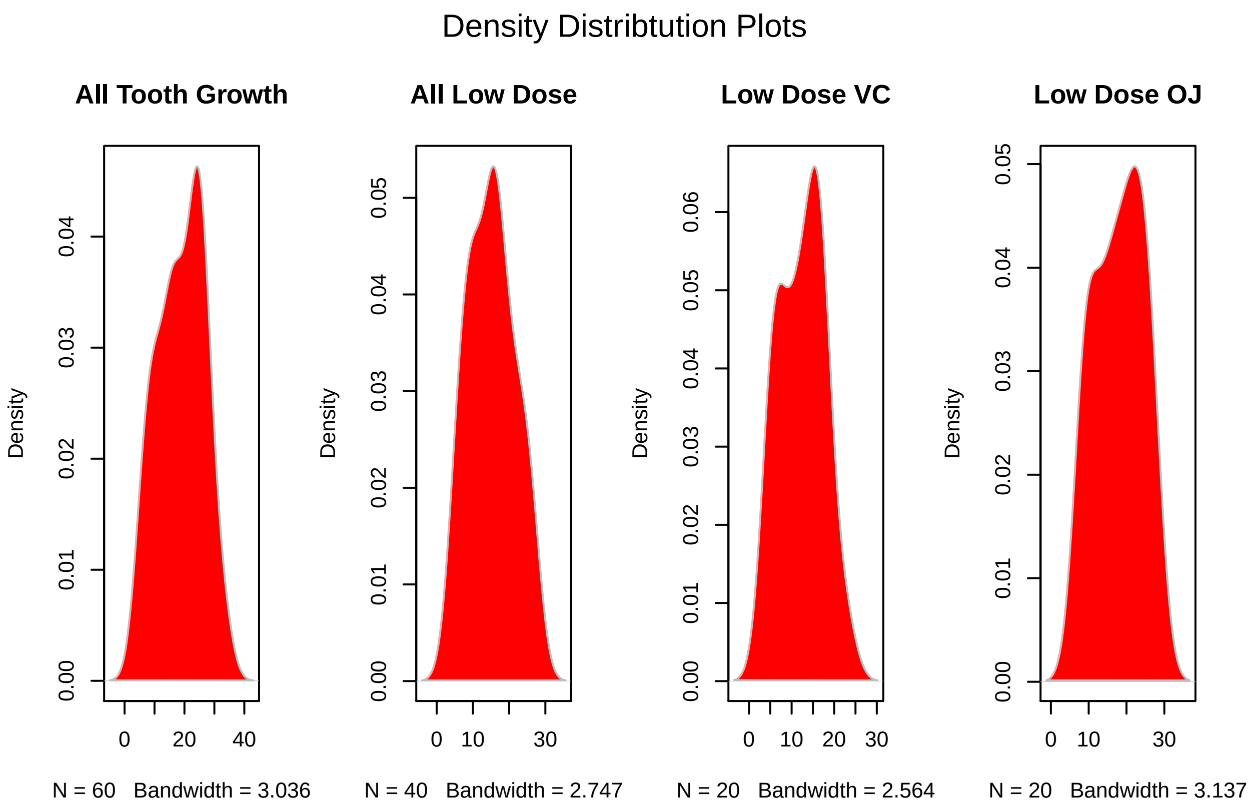
Evaluate Variance and Distribution

```
# Compare low dosage OJ and VC
lowdoseOJ <- ToothGrowth[ToothGrowth$dose<=1 & ToothGrowth$supp=="OJ", ]
lowdoseVC <- ToothGrowth[ToothGrowth$dose<=1 & ToothGrowth$supp=="VC", ]
var(lowdoseVC$len); var(lowdoseOJ$len)  ## compare variance
```

```
## [1] 26.90303
```

```
## [1] 40.26661
```

```
## Density plots.  Do the data look normal?
par(mfrow=c(1,4), oma=c(0,0,3,0)); dtg <- density(ToothGrowth$len)
plot(dtg, type="n", main="All Tooth Growth"); polygon(dtg, col="red", border="grey")
lowdose <- ToothGrowth[ToothGrowth$dose<=1, ]; dtg <- density(lowdose$len)
plot(dtg, type="n", main="All Low Dose"); polygon(dtg, col="red", border="grey")
dtg <- density(lowdoseVC$len); plot(dtg, type="n", main="Low Dose VC")
polygon(dtg, col="red", border="grey"); dtg <- density(lowdoseOJ$len)
plot(dtg, type="n", main="Low Dose OJ"); polygon(dtg, col="red", border="grey")
mtext(text="Density Distribtution Plots", side=3, outer=TRUE )
```



We have tested the variances of the two samples and they don't appear to be equal. So we will check the values of the t-test with var.equal = FALSE, and also with var.equal=TRUE. The density plots do show the data to look somewhat normal. So we will assume the data to be normal.

Perform T.Test

```
t.test(len ~ supp, paired=FALSE, var.equal=FALSE, data=lowdose)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.0503, df = 36.553, p-value = 0.004239
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.875234 9.304766
## sample estimates:
## mean in group OJ mean in group VC
##           17.965           12.375
```

```
t.test(len ~ supp, paired=FALSE, var.equal=TRUE, data=lowdose)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 3.0503, df = 38, p-value = 0.004153
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.880063 9.299937
## sample estimates:
## mean in group OJ mean in group VC
##           17.965           12.375
```

The results of our test show that Vitamin C at dose levels of .5 and 1 mg are more effective through Orange Juice at contributing to tooth length. This is statistically supported since our confidence interval does not include zero, and we have a very low p-value. So we reject the Null Hypothesis and accept our Alternative.

Additional Notes

[sample mean] +- 2(sd)/sqrt(n) is the 95% confidence interval

creating a confidence interval using the Central Limit Theorem (CLT) estimate +- (quantile from standard normal distribution) X (Estimated standard error of the esitmate)

T Confidence intervals: estimate +- (quantile from t distribution) X (estimated standard error the estimate)

tails of t-distribution are a little heavier than normal distribution so the confidence interval is going to be a little wider. As you collect more data the t-interval becomes more and more like the z-interval

With a t-test, The t-value measures the size of the difference relative to the variation in your sample data

the P-value answers the question "How likely would it be to get a statistic this large or larger if the null was actually true?". If the answer to that question is "very unlikely", in other words the P-value is very small, then it sheds doubt on the null being true, since you actually observed a statistic that extreme.