Project: Using the t distribution in hypothesis testing

Vineet W. Singh

19 February 2018

Introduction

The t distribution is used to test statistics on sets of data when the sample sizes are small.
This brief project demonstrates the use of the t distriution in finding confidence intervals and conducting hypothesis tests on a set of data with 60 observations.

Proposal

In the current case, we will be using the t distribution to conduct tests on the "ToothGrowth" dataframe that is a part of the datasets package in R.
The ToothGrowth data set contains data regarding Vitamin C given as a supplement (in two forms) and the resulting growth of teeth in 60 guinea pigs.
More information about the dataset can be found at: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html

We hope to answer the following questions after analysing the data:

1) Do the teeth grow the same, for all dosage of supplements or do different doses of supplements result in different growth.
2) Do both the supplements result in the same rate of growth of teeth or is the growth rate different for each supplement.

Analysis

To begin the analysis the dplyr and datasets packages are loaded using the library command. This automatically loads the tooth growth data into R.

```
library(datasets)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

The structure of the ToothGrowth dataset can be found using the str function.

```
str(ToothGrowth)
```

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

We need to partition this set according to the supplements concerned, either "VC" for Ascorbic Acid or "OJ" for orange juice. The resultant datasets are stored in different frames.

We also need to add a new variable "rate" which will measure the rate at which the teeth have grown according to the amount of supplement given. In effect, we are normalising the length with respect to the dose.

```
tg1 <- ToothGrowth %>% filter(supp=="VC")
tg1<-tg1 %>% mutate(rate=len/dose)
tg2 <- ToothGrowth %>% filter(supp=="OJ")
tg2<-tg2 %>% mutate(rate=len/dose)
```

We can also generate some statistics for each supplement group.

Ascorbic Acid : Summary

```
#calculate summaries for sub group
tg1 %>% group_by(dose) %>% summarise(mean=mean(rate),sd=sd(rate),n=n())
```

```
## # A tibble: 3 x 4
##    dose  mean    sd     n
##   <dbl> <dbl> <dbl> <int>
## 1 0.500  16.0  5.49    10
## 2 1.00   16.8  2.52    10
## 3 2.00   13.1  2.40    10
```

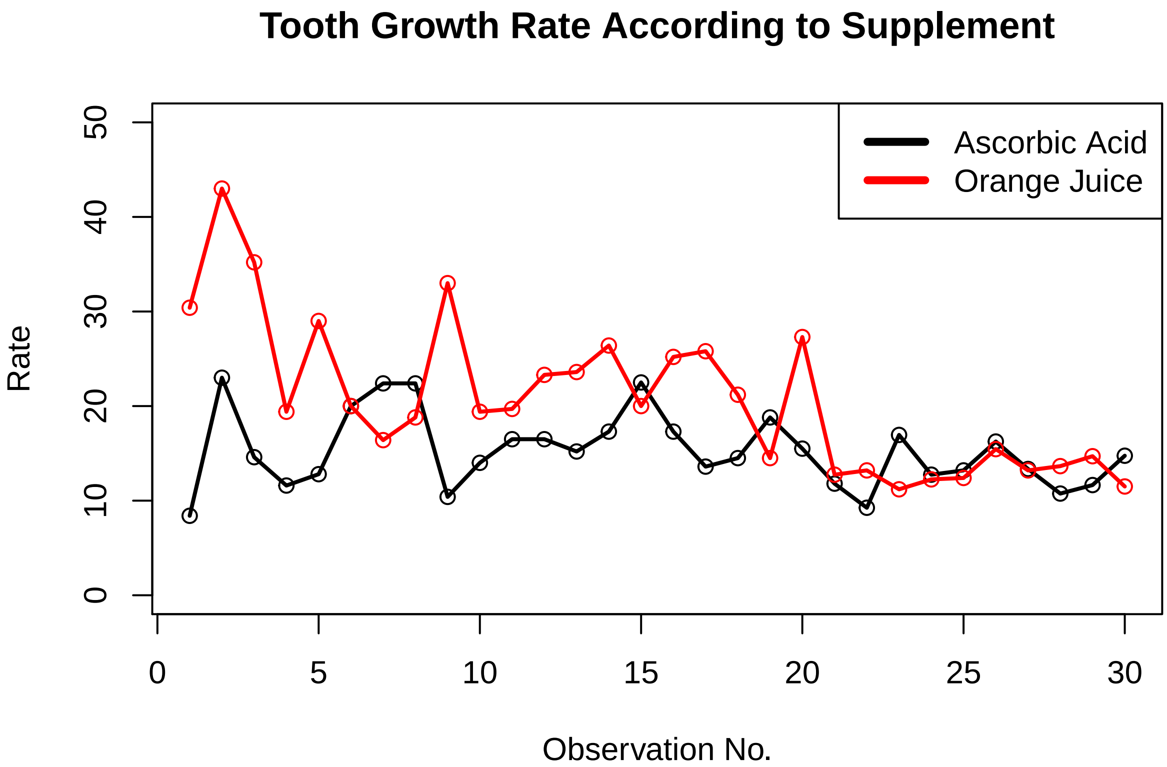
Orange Juice : Summary

```
#calculate summaries for sub group
tg2 %>% group_by(dose) %>% summarise(mean=mean(rate),sd=sd(rate),n=n())
```

```
## # A tibble: 3 x 4
##    dose  mean    sd     n
##   <dbl> <dbl> <dbl> <int>
## 1 0.500  26.5  8.92    10
## 2 1.00   22.7  3.91    10
## 3 2.00   13.0  1.33    10
```

We can draw a plot of the growth rates according to the supplement types.

```
plot(x=row.names(tg1),y=tg1$rate,type="p",pch=1,ylab="Rate",
     main="Tooth Growth Rate According to Supplement",
     xlab="Observation No.", ylim=c(0,50),col=1)
lines(x=row.names(tg1),y=tg1$rate,type="l",lty=1,lwd=2,col=1)
points(x=row.names(tg2),y=tg2$rate,pch=1,col=2)
lines(x=row.names(tg2),y=tg2$rate,type="l",lty=1,lwd=2,col=2)
legend("topright", c("Ascorbic Acid","Orange Juice"),col=c(1,2),lwd=4)
```



From this preliminary analysis, we can observe that the growth rates for both supplements are different. Within each supplement group the growth rates vary according to dosages of supplement given, with the maximum variabilty in the 0.5 mg group for both supplements.

But to get detailed answers we will need to formulate some hypothesis and carry out some tests.

We will need to make some assumptions about the data before we carry out the tests.

We can assume that the growth rate in each animal in each subgroup is independent since each animal is different. Further more we can assume that the variance within each of the 6 subgroups (supplement & dosage) is the same.

However, it is probably not very wise to assume that the variance accross different subgroups is the same as, clearly the dosage levels are different across each subgroup in one supplement group.

The first null hypothesis we propose is:
Null Hypothesis No. 1
Mean growth rates for the 0.5(mg/day) group and the 1(mg/day) group of the ascorbic acid set are the same.

To assist us in our analysis we divide the data up further into three groups for each supplement type.

```
tg11<- tg1 %>% filter(dose==0.5)
tg12<- tg1 %>% filter(dose==1)
tg13<- tg1 %>% filter(dose==2)

tg21<- tg2 %>% filter(dose==0.5)
tg22<- tg2 %>% filter(dose==1)
tg23<- tg2 %>% filter(dose==2)
```

Since we only have 10 measurements each subgroup, we need to make confidence intervals and do tests using the t distribution. This is done using the following command:

```r
t.test(tg11$rate,tg12$rate,paired=F,var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg11$rate and tg12$rate
## t = -0.42396, df = 12.615, p-value = 0.6787
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.950377  3.330377
## sample estimates:
## mean of x mean of y
##     15.96     16.77
```

The t-value is -0.42396, p-value is above .50 and the confidence interval for the difference of means contains 0.
We cannot, therefore, reject null hypothesis no. 1.
We can assume that the mean growth rates for both 0.5 (mg/day) and 1 (mg/day) dose groups of the ascorbic acid supplement set are the same.

The next null hypothesis we propose is:
Null Hypothesis No. 2.
Mean growth rates for the 1(mg/day) group and the 2(mg/day) group of the ascorbic acid set are the same.

We test using the t.test function as outlined above:

```r
t.test(tg12$rate,tg13$rate,paired=F,var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg12$rate and tg13$rate
## t = 3.3662, df = 17.96, p-value = 0.003448
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.390398 6.009602
## sample estimates:
## mean of x mean of y
##     16.77     13.07
```

Here we come across a very high t-value of 3.3662, a low p value and a confidence interval where the difference is well above 1.
We can, therefore, reject the null hypothesis no. 2
and conclude that the means for the two subgroups are different.

We propose the following hypotheses:
Null hypothesis No. 3:
Mean growth rates for the 0.5(mg/day) group and the 1(mg/day) group of the orange juice set are the same.
Null hypothesis No. 4
Mean growth rates for the 1(mg/day) group and the 2(mg/day) group of the orange juice set are the same.

```r
t.test(tg21$rate,tg22$rate,paired=F,var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg21$rate and tg22$rate
## t = 1.2209, df = 12.337, p-value = 0.245
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.93001 10.45001
## sample estimates:
## mean of x mean of y
##     26.46     22.70
```

```
t.test(tg22$rate,tg23$rate,paired=F,var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg22$rate and tg23$rate
## t = 7.404, df = 11.047, p-value = 1.321e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.796869 12.543131
## sample estimates:
## mean of x mean of y
##     22.70     13.03
```

With the test number three, we find that the t-value is above 1 but not very high, the confidence interval of the difference contains 0 so we cannot rule out that the means might not be the same. Results of test number 4 tell a different story, however, the t-value is very large, the p-value is negligible and the confidence interval of the difference does not contain 0.
Using this argument, we cannot reject null hypothesis no. 3, but we can reject null hypothesis no. 4.

Based on the results of the 4 hypothesis tests above, we should merge the two groups of 0.5(mg/day) and 1(mg/day) for each of the two supplement types. and then do a comparision of the two supplement types based on two groups for each of the supplement type 0.5-1(mg/day) and 2(mg/day).

```
#CI for mean growth rate for tg11 and tg12 includes 0 so we cannot rule out
#that the mean growth rate for both tg21 and tg22 can be the same
#it might make sense to group tg11 and tg12 together and assume a common mean
#growth rate.

#merge tg21 and tg22 into one frame
tg11<-rbind(tg11,tg12)

#reassign tg23 to tg22
tg12<-tg13

#CI for mean growth rate for tg21 and tg22 includes 0 so we cannot rule out
#that the mean growth rate for both tg21 and tg22 can be the same
#it might make sense to group tg21 and tg22 together and assume a common growth
#rate.

#merge tg21 and tg22 into one frame
tg21<-rbind(tg21,tg22)

#reassign tg23 to tg22
tg22<-tg23
```

The new data sets are ready for the final phase of the analysis. Hypothesis tests between the two supplement groups to answer the questions about which supplement type and dose effects growth the most.

Null Hypothesis No. 5
There is no difference in the Mean growth rates between 0.5-1 (mg/day) dosage sub-group of the Ascorbic Acid and Orange Juice supplement groups.

```
#check rates and hypothesis test for groups according to dose & supplement
t.test(tg11$rate,tg21$rate,var.equal=F,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg11$rate and tg21$rate
## t = -4.5183, df = 31.083, p-value = 8.455e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.922742  -4.507258
## sample estimates:
## mean of x mean of y
##    16.365    24.580
```

It is clear from the large negative t-value that the means are not the same. The confidence interval does not contain 0 and is entirely negative implying that, according to this test, the average value of tooth growth in the first group (ascorbic acid) will lie below the average value of the tooth growth in the second group (orange juice).
We can, therefore reject the null hypothesis no. 5

Null Hypothesis No. 6
There is no difference in the Mean growth rates between 2(mg/day) dosage sub-group of the Ascorbic Acid and Orange Juice supplement groups.

```
#check rates and hypothesis test for groups according to dose & supplement
t.test(tg12$rate,tg22$rate,var.equal=F,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg12$rate and tg22$rate
## 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:
##  -1.819035  1.899035
## sample estimates:
## mean of x mean of y
##     13.07     13.03
```

This result is more interesting. As can be seen from the plot, the average tooth growth rate for each supplement type begins to get close to one another once the dosage begins to reach the 2(mg/day) level. The t-value is small and p-value large, The confidence interval of the difference is small and does contain 0. We can conclude that based on evidence provided by this data there is no difference in the mean growth rates of the 2(mg/day) subgroup of each supplement group.
We cannot, therefore, reject null hyposthesis no. 6.

It would be interesting to make one final test using all of the 30 observations of each supplement group.

We finally propose:
Null hyposthesis no. 7.
There is no difference in the Mean growth rates between the Ascorbic Acid and Orange Juice supplement groups.

```r
#check rates and hypothesis test for groups according to supplement
t.test(tg1$rate,tg2$rate,var.equal=F,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  tg1$rate and tg2$rate
## t = -3.3707, df = 42.61, p-value = 0.001603
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.732894 -2.193773
## sample estimates:
## mean of x mean of y
##  15.26667  20.73000
```

This result shows that overall if we take normalised growth rates for all observations in the two groups, the growth rates for each supplement type is different as the t-value is large, p-value is negligible and the confidence interval of the difference does not contain 0. It is safe to conclude that the mean growth rates for each supplement type is different.
We can, therefore reject the null hypothesis no. 7.

Conclusion

After analysing the data and proposing seven hypothesis we can conclude that:

1) Average growth rates of teeth for the animals that were given 0.5 to 1(mg/day) of ascorbic acid as a supplement are similar. The average growth rate of the teeth of animals given 2(mg/day) ascorbic acid as a supplement is not similar to the average growth rates of the group of animals given 0.5-1(mg/day) ascorbic acid as a supplement.

2) Average growth rates of teeth for the animals that were given 0.5 to 1(mg/day) of orange juice as a supplement are similar. The average growth rate of the teeth of animals given 2(mg/day) orange juice as a supplement is not similar to the average growth rates of the group of animals given 0.5-1(mg/day) orange juice as a supplement.

3) The average growth rate of teeth of animals given 0.5-1(mg/day) ascorbic acid supplement is not similar to the average growth rate of teeth of animals given 0.5-1(mg/day) orange juice supplement.

4) The average growth rate of teeth of animals given 2(mg/day) ascorbic acid supplement is similar to the average growth rate of teeth of animals given 0.5-1(mg/day) orange juice supplement.

5) Overall, if all data points are taken together (thereby ignoring conlcusion 4 above), the average growth rate of teeth of animals given ascorbic acid supplement is not similar to the average growth rate of teeth of animals given orange juice supplement.