Inference Analysis

Andrew Hope

April 6, 2018

Inference Analysis Comparing Partitioned Data

Overview

This analysis aims to determine if there is a significant difference in mean tooth length based on two different variables. The analysis will first describe the dataset, and will then use an independent two-sample t-test to look for significant differences between subset means. The anlysis will use an alpha value of 0.5, and it assumes that the samples were taken randomly from a population with an approximately normal distribution.

Load Data

Load in the ToothGrowth dataset from the R dataset library.

```
library(datasets)
tg <- ToothGrowth
```

Exploratory Analysis and Transformations

Use some basic exploratory techniques to understand this small dataset.

```
dim(tg)
```

```
## [1] 60  3
```

```
summary(tg)
```

```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000  
```

The quartiles for the dose column are a bit unusual in the summary. They are multiples of 0.5. Explore this further.

```
unique(tg$dose)
```

```
## [1] 0.5 1.0 2.0
```

This variable should be a factor rather than numeric. Perform this transformation.

```
    tg$dose <- as.factor(tg$dose)
    summary(tg)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

Calculate Statistics

Split the data by supp and dose.

```
    supp <- split(tg$len, tg$supp)
    dose <- split(tg$len, tg$dose)
```

Calculate the mean, variance, and standard deviation for each value of supp.

```
    ms <- tapply(tg$len, tg$supp, mean)
    vs <- tapply(tg$len, tg$supp, var)
    ms
```

```
##       OJ       VC
## 20.66333 16.96333
```

```
    vs
```

```
##       OJ       VC
## 43.63344 68.32723
```

```
    sqrt(vs)
```

```
##       OJ       VC
## 6.605561 8.266029
```

Do the same with dose.

```
    md <- tapply(tg$len, tg$dose, mean)
    vd <- tapply(tg$len, tg$dose, var)
    md
```

```
##    0.5      1      2 
## 10.605 19.735 26.100
```

```
    vd
```

```
##      0.5        1        2 
## 20.24787 19.49608 14.24421
```

```
    sqrt(vd)
```

```
##      0.5        1        2 
## 4.499763 4.415436 3.774150
```

Compare Mean of Partitions

Compare the means (grouped by supp) using a 95% confidence interval and an independent two-tailed t-test.

```
supp.t <- t.test(supp$OJ, supp$VC, conf.level = 0.95)
supp.t
```

```
## 
##  Welch Two Sample t-test
## 
## data:  supp$OJ and supp$VC
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y 
##  20.66333  16.96333
```

For dose, execute an independent two-sample t-test between each factor level. This will result in three comparisons: 0.5 and 1, 0.5 and 2, and 1 and 2.

First comparison – factor level 0.5 to 1.

```
dose.t12 <- t.test(dose$`0.5`, dose$`1`, conf.level = 0.95)
dose.t12
```

```
##
##  Welch Two Sample t-test
##
## data:  dose$`0.5` and dose$`1`
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Second comparison – factor level 0.5 to 2.

```
dose.t13 <- t.test(dose$`0.5`, dose$`2`, conf.level = 0.95)
dose.t13
```

```
##
##  Welch Two Sample t-test
##
## data:  dose$`0.5` and dose$`2`
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

Third comparison – factor level 1 to 2.

```
dose.t23 <- t.test(dose$`1`, dose$`2`, conf.level = 0.95)
dose.t23
```

```
##
##  Welch Two Sample t-test
##
## data:  dose$`1` and dose$`2`
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

Conclusions

For the supp variable, using a 95% confidence interval does not allow us to reject a null hypothesis. The two groups do not have significantly different means. This is indicated by the confidence interval containing zero, and also the p-value being greater than our accepted alpha of 0.5.

```
    supp.t$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

```
    supp.t$p.value
```

```
## [1] 0.06063451
```

For the dose variable, each comparison gave a confidence interval that did not include zero. This indicates that the groups have significantly different means. For all three comparisons, the p-value is less than 0.5, allowing us to comfortably reject the null hypothesis.

```
    dose.t12$p.value
```

```
## [1] 1.268301e-07
```

```
    dose.t13$p.value
```

```
## [1] 4.397525e-14
```

```
    dose.t23$p.value
```

```
## [1] 1.90643e-05
```