ToothGrowth

"Stone Cold" Tomaz Rodrigues

2025-02-22

Statistical Inference Course Project

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package. Load the ToothGrowth data and perform some basic exploratory data analyses Provide a basic summary of the data. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering) State your conclusions and the assumptions needed for your conclusions.

About the data

Taken from https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/ToothGrowth The response 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).

Format A data frame with 60 observations on 3 variables.

[,1] len numeric Tooth length [,2] supp factor Supplement type (VC or OJ). [,3] dose numeric Dose in milligrams/day

```r
#You can see a bit of the raw data below to get a better idea of the format
print(head(ToothGrowth))
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```r
#getting an idea of how len responds to different doses of the 2 supplements
# Using data.table
ToothGrowth_dt <- as.data.table(ToothGrowth)
mean_min_max_dt <- ToothGrowth_dt[, .(mean_len = mean(len),
                                      min_len = min(len),
                                      max_len = max(len)),
                                  by = .(supp, dose)] %>%
  setorder(supp)
print(mean_min_max_dt)
```

```
##      supp  dose mean_len min_len max_len
##    <fctr> <num>    <num>   <num>   <num>
## 1:     OJ   0.5    13.23     8.2    21.5
## 2:     OJ   1.0    22.70    14.5    27.3
```

```
## 3:     OJ   2.0    26.06    22.4    30.9
## 4:     VC   0.5     7.98     4.2    11.5
## 5:     VC   1.0    16.77    13.6    22.5
## 6:     VC   2.0    26.14    18.5    33.9
```

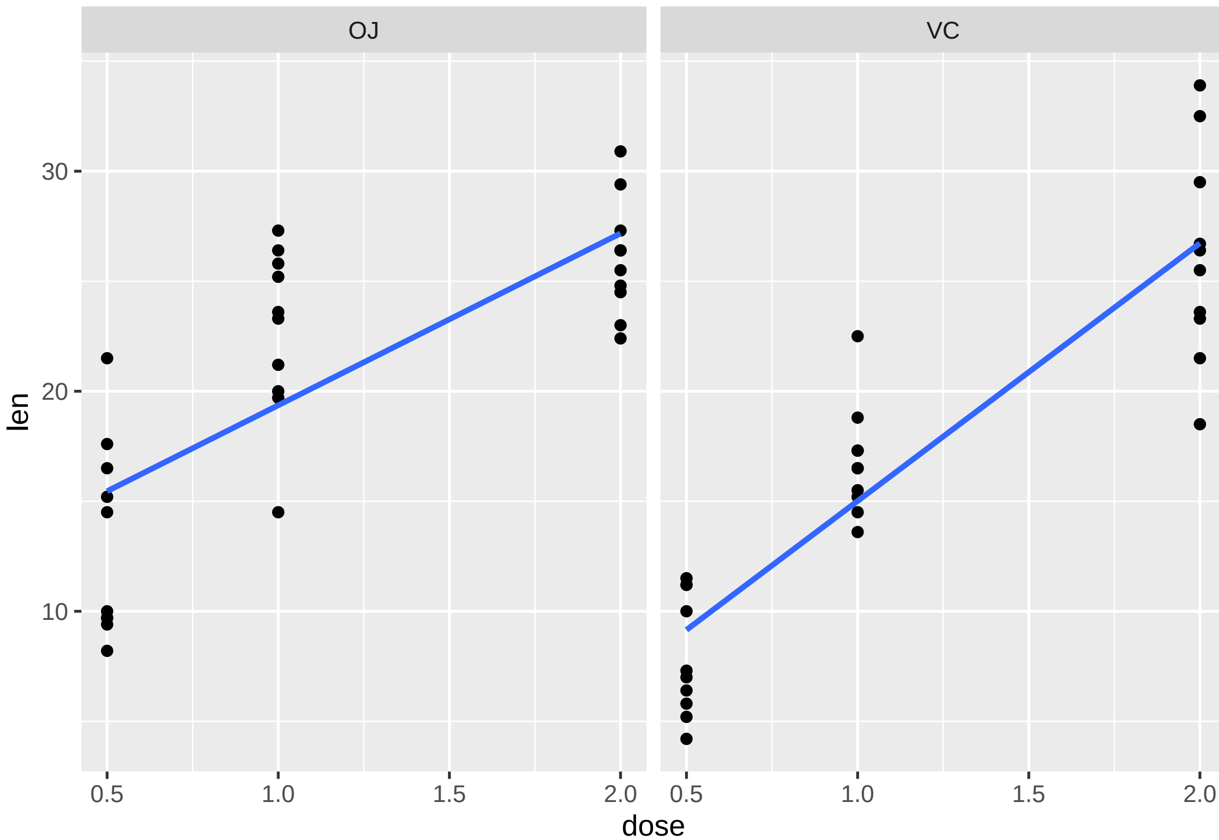
Plotting the data and establishing hypotheses

```
ToothGrowth_dt <- as.data.table(ToothGrowth)
mean_min_max_dt <- ToothGrowth_dt[, .(mean_len = mean(len),
                                      min_len = min(len),
                                      max_len = max(len)),
                                  by = .(supp, dose)] %>%
  setorder(supp)
print(mean_min_max_dt)
```

```
##      supp  dose mean_len min_len max_len
##    <fctr> <num>    <num>   <num>   <num>
## 1:     OJ   0.5    13.23     8.2    21.5
## 2:     OJ   1.0    22.70    14.5    27.3
## 3:     OJ   2.0    26.06    22.4    30.9
## 4:     VC   0.5     7.98     4.2    11.5
## 5:     VC   1.0    16.77    13.6    22.5
## 6:     VC   2.0    26.14    18.5    33.9
```

```
#Getting a basic idea of how len responds to dose
##plotting the data to establish a relationship between dose and len for each supp
ggplot(ToothGrowth, aes(x = dose, y = len)) +
  geom_point() +  # Add points
  facet_wrap(~ supp) + # Create panels based on "sup"
  geom_smooth(method = "lm", se = FALSE) + # Add linear trend lines
  scale_y_continuous(limits = c(min(ToothGrowth$len), max(ToothGrowth$len))) # Set same y-axis scale
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#As you can see from the plotted data, the higher doses appear to result in more tooth growth for the r
# It appears that that at a dose of 2.0 this difference is muddied and both supps produce a similar eff
```

Testing hypotheses formally with t.test

```
#now we will use T inteval testing to establish confidence intervals and try to more conclusively
#establish this relationship

lowDose <- subset(ToothGrowth, dose==0.5)
t.test(len ~ supp, var.equal = TRUE, data = lowDose)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  1.770262 8.729738
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
#this t test establishes a clear difference between the two groups. since the p value is
# 0.005 we accept the hypothesis that OJ has more impact than VC

mediumDose <-  subset(ToothGrowth, dose==1.0)
```

```r
t.test(len ~ supp, var.equal = TRUE, data = lowDose)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 18, p-value = 0.005304
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  1.770262 8.729738
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```r
#this t test establishes a clear difference between the two groups. since the p value is
# 0.005 we accept the hypothesis that OJ has more impact than VC

highestDose <-  subset(ToothGrowth, dose==2.0)
t.test(len ~ supp, var.equal = TRUE, data = highestDose)
```

```
##
##  Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 18, p-value = 0.9637
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -3.722999  3.562999
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

```r
#this t test does not establish a clear difference between the two supplements at the dose of 2.0
#thus we are unable to reject the null hypothesis that the true means are similar
#this keeps with our previous hypotheses created from the chart
```

Conclusions

The hypotheses we established after plotting the data were proven formally to be mostly correct by the T testing section with confidence intervals. At doses of 0.5 and 1.0 the orange juice group has a 99.5% chance of responding more than the VC group. However, at dose 2.0 both groups seemed to have remarkably similar results. To more conclusively establish these results additional data would be needed.