Basic Inferential Project - Statistical Inference JHU Data Science Specialisation

Guynemer Cétoute

2025-03-20

The purpose of the this work is to analyze the ToothGrowth data set by comparing the guinea tooth growth by supplement and dose. Firstly, I will do exploratory data analysis on the data set.

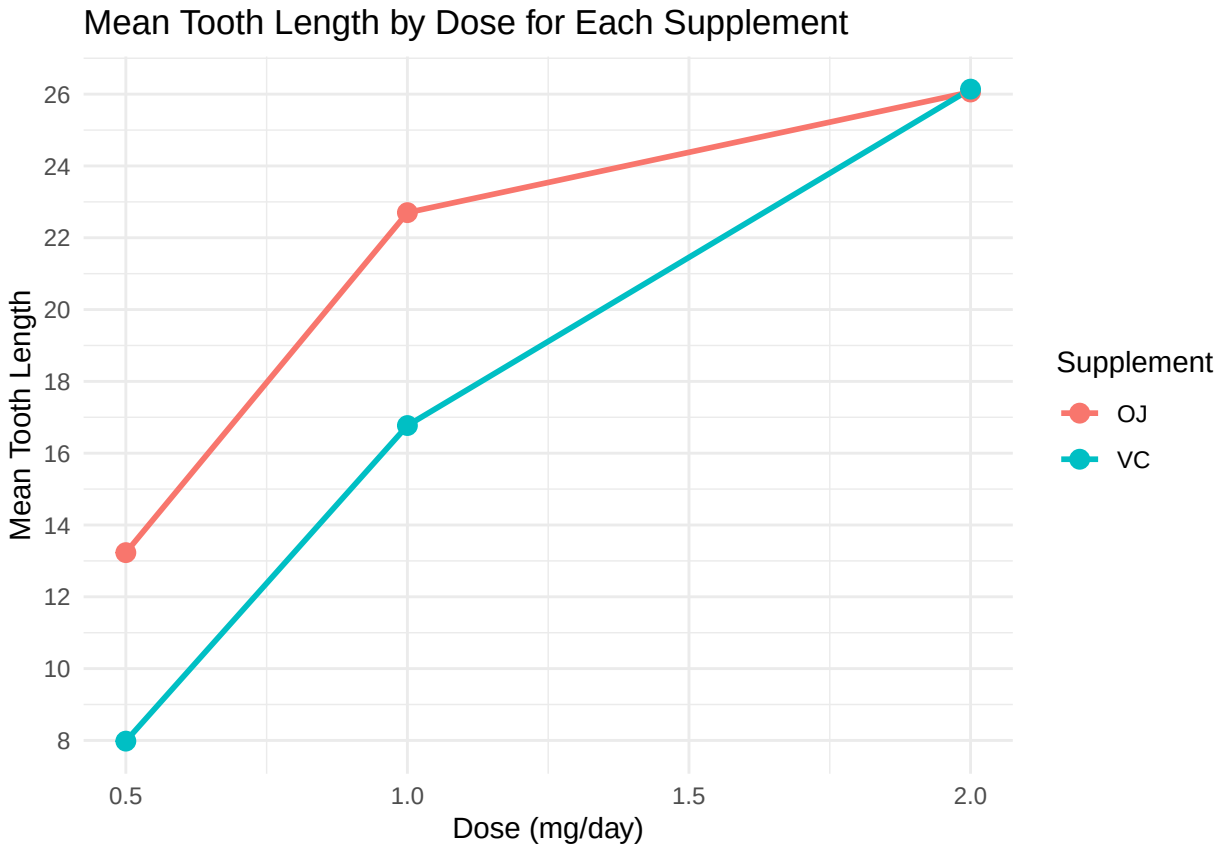
Load the ToothGrowth data and perform some basic exploratory data analyses

```
## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x dplyr::select()    masks MASS::select()
## x dplyr::src()       masks Hmisc::src()
## x dplyr::summarize() masks Hmisc::summarize()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error

##       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
```

Plot

```r
t_summary <- t %>% group_by(supp, dose) %>%
    summarise(mean_len = mean(len), .groups = "drop")

ggplot(t_summary, aes(x = dose, y = mean_len, color = supp)) + geom_line(size = 1) +
    geom_point(size = 3) +
    scale_y_continuous(breaks = seq(0, max(t_summary$mean_len) + 5, by = 2)) +
    labs(title = "Mean Tooth Length by Dose for Each Supplement",
       x = "Dose (mg/day)", y = "Mean Tooth Length", color = "Supplement") +
    theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

Hypothesis 1

Confidence Interval

```r
hypoth1<-t.test(len ~ supp, data = t)
round(hypoth1$conf.int, 3)
```

```
## [1] -0.171  7.571
## attr(,"conf.level")
```

```
## [1] 0.95
```

P-Value

```
round(hypoth1$p.value, 3)
```

```
## [1] 0.061
```

Hypothesis 2

Confidence Interval

```
hypoth2<-t.test(len ~ supp, data = subset(t, dose == 0.5))
round(hypoth2$conf.int, 3)
```

```
## [1] 1.719 8.781
## attr(,"conf.level")
## [1] 0.95
```

P-Value

```
round(hypoth2$p.value, 3)
```

```
## [1] 0.006
```

Hypothesis 3

Confidence Interval

```
hypoth3<-t.test(len ~ supp, data = subset(t, dose == 1))
round(hypoth3$conf.int, 3)
```

```
## [1] 2.802 9.058
## attr(,"conf.level")
## [1] 0.95
```

P-Value

```
round(hypoth3$p.value, 3)
```

```
## [1] 0.001
```

Hypothesis 4

Confidence Interval

```
hypoth4 <-t.test(len ~ supp, data = subset(t, dose == 2))
round(hypoth4$conf.int, 3)
```

```
## [1] -3.798  3.638
## attr(,"conf.level")
## [1] 0.95
```

P-Value

```
round(hypoth4$p.value, 3)
```

```
## [1] 0.964
```

Conclusion

For the Hypothesis 1, the confidence intervals includes 0 and the p-value is greater than the threshold of 0.05. The null hypothesis cannot be rejected.

For the Hypothesis 2, the confidence interval does not include 0 and the p-value is below the 0.05 threshold. The null hypothesis can be rejected. The alternative hypothesis that 0.5 mg/day dosage of orange juice delivers more tooth growth than ascorbic acid is accepted.

For the Hypothesis 3, the confidence interval does not include 0 and the p-value is smaller than the 0.05 threshold. The null hypothesis can be rejected. The alternative hypothesis that 1 mg/day dosage of orange juice delivers more tooth growth than ascorbic acid is accepted.

For the Hypothesis 4, the confidence interval does include 0 and the p-value is larger than the 0.05 threshold. The null hypothesis cannot be rejected.

Finally, we can say that : Orange juice delivers more tooth growth than ascorbic acid for dosages 0.5 & 1.0. Orange juice and ascorbic acid deliver the same amount of tooth growth for dose amount 2.0 mg/day. For the entire data set we cannot conclude orange juice is more effective that ascorbic acid.