ToothGrowth Data Analysis

Statistical Inference – Class Project – Part 2

Author: Josh Jensen

Overview:

Using the 'ToothGrowth' data set in R, this will provide a practical demonstration of using confidence intervals as a tool for statistical inference. The `len` is a observed fact continuous variable, whereas `supp` and `dose` are the relevant dimensions.

Load the ToothGrowth data and perform some basic exploratory data analyses:

In the initial exploration, we see that there are 10 observations for each `dose` and each `supp` combination.

```
> # Load dataset
> toothgrowth <- data.frame(ToothGrowth)
>
> # View data
> View(toothgrowth)
> # Table of Number of Observations by supp and dose
> table(toothgrowth$supp,toothgrowth$dose)
     0.5  1  2
  OJ  10 10 10
  VC  10 10 10
> # Summary of len
> summary(toothgrowth$len)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.20   13.08   19.25   18.81   25.28   33.90
```

Provide a basic summary of the data:

From the initial summary statistics we see that on average `len` increases with `dose` and 'OJ' is the higher `supp`. However, we can also see that many the standard deviations are quite large so these differences may be random occurrences.

```
> # Table of len means by supp and dose
> means <- tapply(toothgrowth$len, list(toothgrowth$supp, toothgrowth$dose),
FUN = mean)
> round(means,2)
     0.5     1     2
OJ 13.23 22.70 26.06
VC  7.98 16.77 26.14
> # Table of len standard deviations by supp and dose
> stdevs <- tapply(toothgrowth$len, list(toothgrowth$supp, toothgrowth$dose),
FUN = sd)
> round(stdevs,2)
    0.5    1    2
OJ 4.46 3.91 2.66
VC 2.75 2.52 4.80
```

Use confidence intervals to compare tooth growth by supp and dose:
As the number of observations for all cells are 10 and relatively low, it is most appropriate to use the Students T distribution with 9 degrees of freedom to calculate the confidence intervals.

```
> # Compute 95% confidence interval using Student T distribution with 9 d.f.
> ul_95 <- means + qt(0.975,9)*(stdevs/sqrt(10))
> ll_95 <- means - qt(0.975,9)*(stdevs/sqrt(10))
>
> # Lower limits
> round(ll_95,2)
     0.5     1     2
OJ 10.04 19.90 24.16
VC  6.02 14.97 22.71
> # Upper limits
> round(ul_95,2)
     0.5     1     2
OJ 16.42 25.50 27.96
VC  9.94 18.57 29.57
```

State conclusions and assumptions:

- Conclusions:
    - As dose increases from 0.5 to 1 the mean of `len` is significantly larger with 95% confidence. This is the case for both `supp` values (16.42 vs. 19.90 for OJ; 9.94 vs. 14.97 for VC). However, as `dose` increases from 1 to 2 the mean of `len` is only significantly larger with 95% confidence for the VC `supp`. Note that this is the case as the OJ `supp` the upper limit `len` of `dose` 1 is 25.50, which is greater than the lower limit `len` of `dose` 2 (24.16).
    - For `doses` of 0.5 and 1, the OJ `supp` mean of `len` is significantly larger than that of the VC `supp` with 95% confidence. This can be seen as the upper limits of VC are less than the lower limits of OJ for `doses` 0.5 and 1. However, no meaningful conclusion can be drawn for the `dose` of 2.
- Assumptions:
    - The observed data are i.i.d. normally distributed.
    - The data is roughly symmetric & mound shaped.

Appendix

Full R Script:

```
#
### 1. Load the ToothGrowth data and perform some basic exploratory
data analyses
#

# Load dataset
toothgrowth <- data.frame(ToothGrowth)

# View data
View(toothgrowth)
# Table of Number of Observations by supp and dose
table(toothgrowth$supp,toothgrowth$dose)
# Summary of len
summary(toothgrowth$len)



#
### 2. Provide a basic summary of the data
#

# Table of len means by supp and dose
means <- tapply(toothgrowth$len, list(toothgrowth$supp,
toothgrowth$dose), FUN = mean)
round(means,2)
# Table of len standard deviations by supp and dose
stdevs <- tapply(toothgrowth$len, list(toothgrowth$supp,
toothgrowth$dose), FUN = sd)
round(stdevs,2)



#
### 3. Compare tooth growth by supp and dose
#

# Compute 95% confidence interval using Student T distribution with 9
d.f.
ul_95 <- means + qt(0.975,9)*(stdevs/sqrt(10))
ll_95 <- means - qt(0.975,9)*(stdevs/sqrt(10))

# Lower limits
round(ll_95,2)
# Upper limits
round(ul_95,2)
```