John Hopkins University – Data Science Specialization – Statistical Inference Course – Solution to Course Project – Part 2/2

Dr. Guy Cohen

July 22, 2015

Simulations

PLEASE NOTE: Some of the code and plots are shown in the attached appendix

Task 1: Load the ToothGrowth data and perform some basic exploratory data analyses.
Solution: I load the data into R.

```
library(datasets)
data(ToothGrowth)
```

I do some basic exploratory analysis.

```
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 ...
```

```
summary(ToothGrowth)
```

```
##       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  
```

```
table(ToothGrowth$dose)
```

```
## 
## 0.5   1   2 
##  20  20  20 
```

```
print(paste("Number of complete cases is: ", sum(complete.cases(ToothGrowth))))
```

```
## [1] "Number of complete cases is:  60"
```

I see no data are missing. I plot various box plots to compare. The code for plotting the boxplots and the boxplots appear in the **appendix**.

From the graphs I see that for a given dose of 0.5mg or 1mg, orange juice seems to be related to greater tooth length than vitamin C. We also see that the relation of "dose" to "len" seems significant. An increase in dose seems to increase the tooth length. An exception to the rule is seen when comparing a dose of 1mg to 2mg when orange juice is given. In this case, the difference is not clear-cut and additional investigation is needed.

I now examine regression lines for tooth length vs. dose. The code and the plots appear in the **appendix**. Indeed, I see that even with the confidence intervals, the positive correlation between dose and tooth length for the vitamin C supplement is statistically significant, and that this correlation is more robust than it is with orange juice.

Task 2: Provide a basic summary of the data.
Solution: The dataset is titled "The Effect of Vitamin C on Tooth Growth in Guinea Pigs." The dataset is given as 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

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

Using the summary() function I see

```
summary(ToothGrowth)
```

```
##       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  
```

As described above, there is a clear positive correlation between dose and tooth length, and a stronger relation of tooth length to orange juice than to vitamin C when dose is fixed as equal to 0.5mg or 1mg. These observations will be studied quantitatively below.

Tasks 3 and 4: 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.
Solution: The hypotheses I wish to test are the following:
1. Is length different when supplement type is orange juice vs. vitamin C?
2. Is length greater when supplement type is orange juice vs. vitamin C and dose is 0.5 mg?
3. Is length greater when supplement type is orange juice vs. vitamin C and dose is 1 mg?
4. Is length greater when supplement type is orange juice vs. vitamin C and dose is 2 mg?
5. Is length lower when dose is 0.5 mg vs. when dose is 1 mg ?
6. Is length lower when dose is 1 mg vs. when dose is 2 mg ?
7. Is length lower when dose is 0.5 mg vs. when dose is 1 mg with orange juice?
8. Is length lower when dose is 0.5 mg vs. when dose is 1 mg with vitamin C?
9. Is length lower when dose is 1 mg vs. when dose is 2 mg with orange juice?
10. Is length lower when dose is 1 mg vs. when dose is 2 mg with vitamin C?

Assumptions for all hypotheses: The samples parent population distributions are normal. The observations in each sample are independent and identically-distributed (IID) random variables (RV).

Code for testing the hypotheses and its output are given in the **appendix**. All hypotheses are tested with the *t*-test. Conclusions from running the code:

1. Answer is **No**. I fail to reject null hypothesis of equal length. 95% confidence interval for length difference is (-0.1710156 7.5710156) .
2. Answer is **Yes**. I reject null hypothesis of equal length. 95% confidence interval for length difference is (2.34604, ∞) .
3. Answer is **Yes**. I reject null hypothesis of equal length. 95% confidence interval for length difference is (3.356158, ∞) .
4. Answer is **No**. I fail to reject null hypothesis of equal length. 95% confidence interval for length difference is (-3.1335, ∞) .
5. Answer is **Yes**. I reject null hypothesis of equal length. 95% confidence interval for length difference is (-∞, -6.753323) .
6. Answer is **Yes**. I reject null hypothesis of equal length. 95% confidence interval for length difference is (-∞, -4.17387) .
7. Answer is **Yes**. I reject null hypothesis of equal length. 95% confidence interval for length difference is (-∞, -6.214316) .
8. Answer is **Yes**. I reject null hypothesis of equal length. 95% confidence interval for length difference is (-∞, -6.746867) .
9. Answer is **Yes**. I reject null hypothesis of equal length. 95% confidence interval for length difference is (-∞, -0.7486236) .
10. Answer is **Yes**. I reject null hypothesis of equal length. 95% confidence interval for length difference is (-∞, -6.346525) .