Statistical Inference Final Project: Demonstrating the Central Limit Theorem and using Confidence Intervals - Part 2

James Wright

October 2016

Overview

This report covers the Statistical Inference course project, which is made up of two parts. The first part of the project is a demonstration of the Central Limit Theorem using the exponential distribution in R - how sampling works with the Central Limit Theorem. The second part of the project was an analysis of the ToothGrowth dataset available in the R datasets package.The aim here is to use confidence intervals to compare tooth growth by supp and dose.

Part 2: Confindence Intervals

The ToothGrowth Data Set

The ToothGrowth data set is contained in the R library datasets. The dataset contains data relating to a study which investigated the effect of vitamin C on tooth growth in guinea pigs.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).

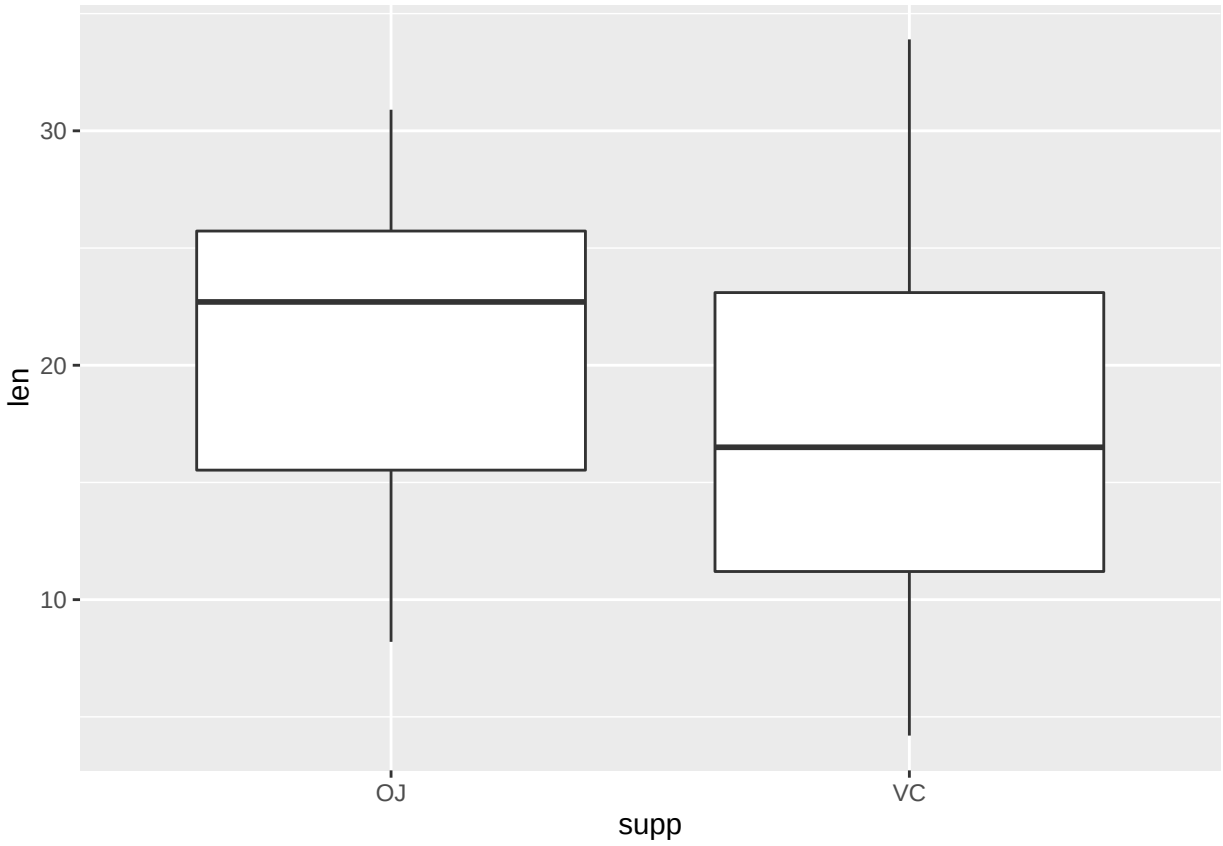
We start off by loading the dataset, and then create a brief summary of the data to see what variables we're working with.

```
##       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 summary (and the documentation associated with the data set) tells us that we have 3 variables in the data frame:

- len : a numeric variable of the tooth length (mm)
- supp: a factor variable of the type of vitamin C supplement (absorbic acid or orange juice)
- dose: a numeric variable of the dose of the supplement (mg/day)

We can now plot some exploratory graphs to get a better feel for what the data is showing us.

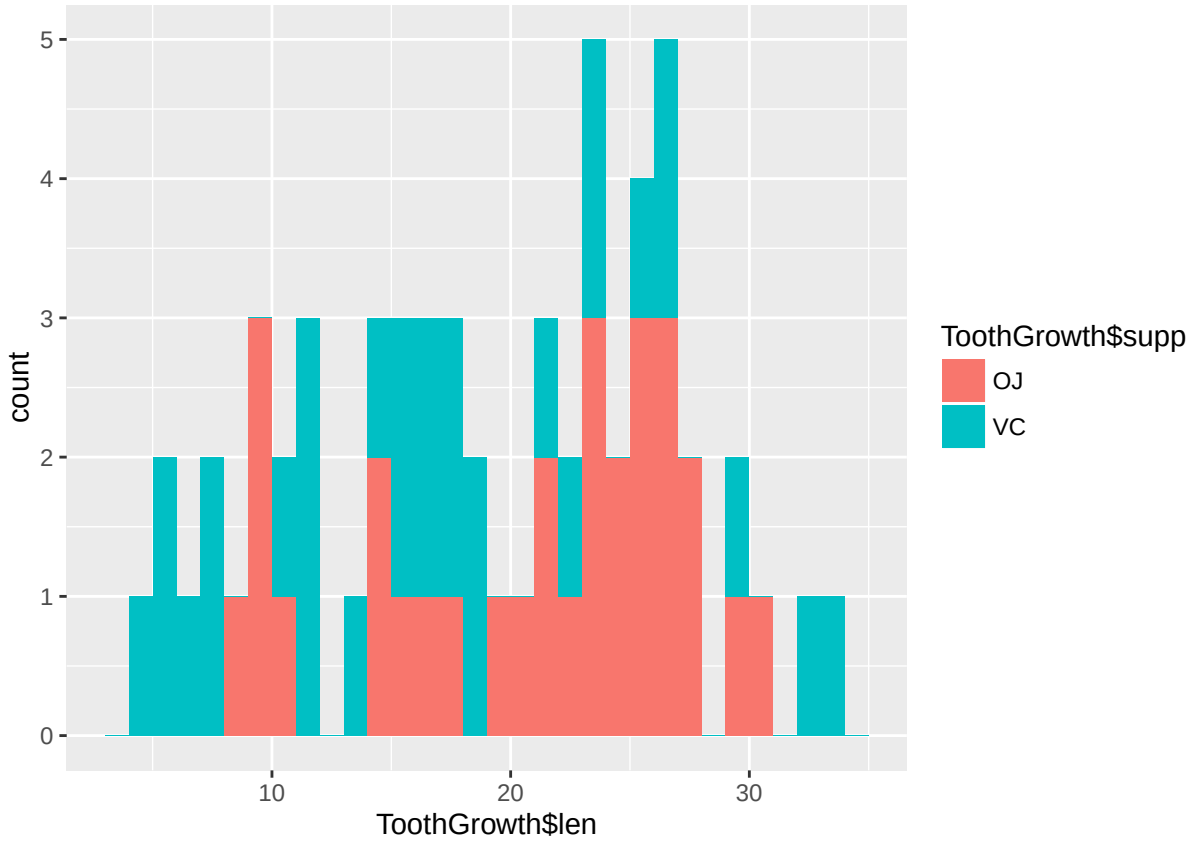


A boxplot shows us the variability, or spread, of the data. From the boxplot above, we can see that there is a larger range in odontoblast lengths of the guinea pigs that were given the asorbic acid supplement compared to those given the orange juice supplement.

We can also see that there is a larger spread of data between the first quartile (Q1) and the median (Q2) for the orange juice supplement (larger spread than between Q2 and Q3), indicating that the data in the middle 50% of the data set are skewed left. The distances from the minimum value to Q1 and Q2 to the maximum value are about equal, indicating that there is no tail on either side of the data set for the orange juice supplement.

The asorbic acid data are slightly more variable between the median and Q3, indicating a slight right skew of the data contained within the middle 50% of the data set. Again, there is no particularly strong indication of a tail on either side.

The described characteristics of the boxplots given above are confirmed by a histogram of each data set, shown below:



The boxplots don't give us any indication of the mean of the data sets, and it is difficult to see from the histogram where each distribution is centred. We can see what the mean odontoblast length is, grouped by each supplement, with the following code:

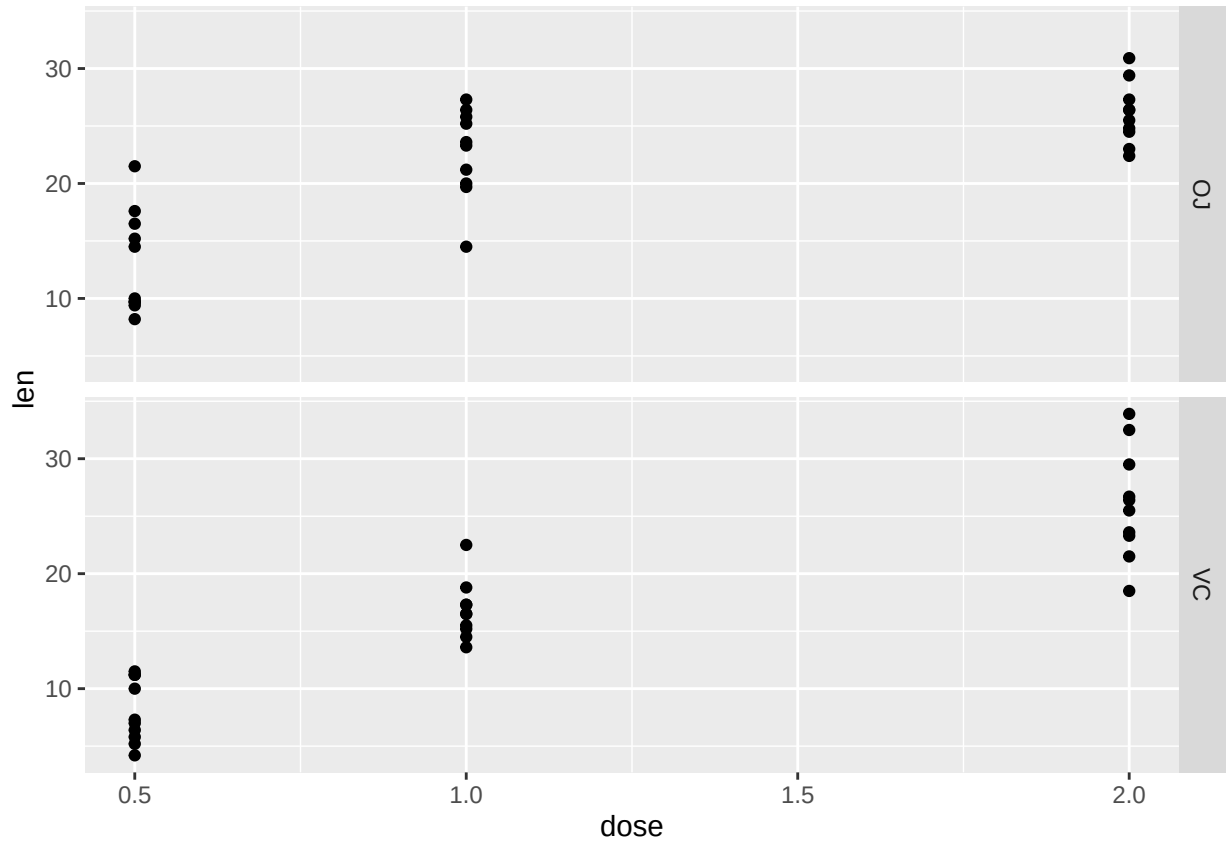
```
library(dplyr)
```

```
grouped = group_by(ToothGrowth,supp)
summarise_each(grouped,funs(mean(.)),len)
```

```
## Source: local data frame [2 x 2]
##
##     supp      len
##   (fctr)    (dbl)
## 1     OJ 20.66333
## 2     VC 16.96333
```

It appears as though the mean odontoblast length of the sample of the guinea pigs that were given the orange juice supplement was greater than the mean odontoblast length of those that were given the ascorbic acid. This assertion will form the hypothesis to be tested a little later on.

We can also see how the length of the odontoblasts were affected by dose:



We can summarise the effects of dose on the length of odontoblast with:

```
grouped = group_by(ToothGrowth,dose)
summarise_each(grouped,funs(mean(.)),len)
```

```
## Source: local data frame [3 x 2]
##
##    dose    len
##   (dbl)  (dbl)
## 1   0.5 10.605
## 2   1.0 19.735
## 3   2.0 26.100
```

This summary shows that the mean length of odontoblasts increases with dose of vitamin C for this particular sample.

Confidence Intervals

We want to look at the assertion made earlier, that the orange juice supplement was more effective at increasing the length of the odontoblasts than the ascorbic acid supplement. Taking the mean of each sample suggested that this was the case, and now we want to estimate what the difference in means is with 95% confidence. We will do this by performing a two-sided t-test.We will assume that both samples have different variances, and we will assume that different guinea pigs were used for each trial - i.e. the data samples are indpendent and not paired.

```r
OJ = ToothGrowth[ToothGrowth$supp=="OJ",]
VC = ToothGrowth[ToothGrowth$supp=="VC",]

#Assume unequal variances
T = t.test(OJ$len,VC$len,paired = FALSE,var.equal=FALSE)
print(T)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ$len and VC$len
## 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
```

```r
#Mean difference
Xbar = mean(OJ$len)
Ybar= mean(VC$len)

md = Xbar-Ybar
print(md)
```

```
## [1] 3.7
```

```r
MoE = md - T$conf.int[1]
print(MoE)
```

```
## [1] 3.871016
```

```r
MoE = T$conf.int[2] - md
print(MoE)
```

```
## [1] 3.871016
```

The confidence interval helps us decide whether or not we can use the difference in our sample means to describe the difference in population means. We have just calculated our t-interval, which is the difference in means of the two samples plus or minus our margin of error. We can see that the margin of error is larger than the difference in our sample means. So the results of the t-test are that we can say, with 95% confidence, that there is no difference in odontoblast length between those that were given the orange juice supplement and those that were given the ascorbic acid supplement.

This result was obtained by looking at the averages of all doses. We can gain a bit more insight by looking at the effect of dose on the impact of the supplements on odontoblast lengths. Doing this (see following code) shows us that in actual fact, at doses of 0.5 and 1 mg/day, the odontoblast length is on average greater for the guinea pigs that were given the orange juice supplement (with 95% confidence). At doses of 2 mg/day, the sample difference is smaller than our margin of error and so we can't say, with 95% confidence, that there is a difference in length of odontoblasts between the two delivery methods.

```r
OJ_dhalf = ToothGrowth[(ToothGrowth$supp=="OJ" & ToothGrowth$dose==0.5),]
VC_dhalf = ToothGrowth[(ToothGrowth$supp=="VC" & ToothGrowth$dose==0.5),]

#Assume unequal variances
T = t.test(OJ_dhalf$len,VC_dhalf$len,paired = FALSE,var.equal=FALSE)

#Mean difference
Xbar = mean(OJ_dhalf$len)
Ybar= mean(VC_dhalf$len)

md = Xbar-Ybar
MoE = md - T$conf.int[1]
print(c(md,MoE))
```

```
## [1] 5.250000 3.530943
```

```r
OJ_d1 = ToothGrowth[(ToothGrowth$supp=="OJ" & ToothGrowth$dose==1),]
VC_d1 = ToothGrowth[(ToothGrowth$supp=="VC" & ToothGrowth$dose==1),]

#Assume unequal variances
T = t.test(OJ_d1$len,VC_d1$len,paired = FALSE,var.equal=FALSE)

#Mean difference
Xbar = mean(OJ_d1$len)
Ybar= mean(VC_d1$len)

md = Xbar-Ybar
MoE = md - T$conf.int[1]
print(c(md,MoE))
```

```
## [1] 5.930000 3.127852
```

```r
OJ_d2 = ToothGrowth[(ToothGrowth$supp=="OJ" & ToothGrowth$dose==2),]
VC_d2 = ToothGrowth[(ToothGrowth$supp=="VC" & ToothGrowth$dose==2),]

#Assume unequal variances
T = t.test(OJ_d2$len,VC_d2$len,paired = FALSE,var.equal=FALSE)

#Mean difference
Xbar = mean(OJ_d2$len)
Ybar= mean(VC_d2$len)

md = Xbar-Ybar
MoE = md - T$conf.int[1]
print(c(md,MoE))
```

```
## [1] -0.08000  3.71807
```