Statistical Inference Course Project

Author: Dale Hunscher

Creation Date: Friday, 27 April 2018

Part 1: Simulation Exercise

Overview

> In this part of the project, you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

We'll compare sample mean and variance with their theorical values. Then we will use histograms to do a visual comparison of the sample distribution with the narmal paradigm.

Setup

First we set our working directory and load libraries:

Next we set the random seed so our project results will be reproducible.

Set the variables we'll need based on the project instructions:

Explorations Part One: Comparing Sample and Theory

Compute the exponential sample data set, and then the sample and theoretical means and standard deviations.

Comparing means: first the sample and theoretical means…

```
## [1] 5.032989
```

```
## [1] 5
```

…and the sample and theoretical variances…

```
## [1] 0.6276361
```

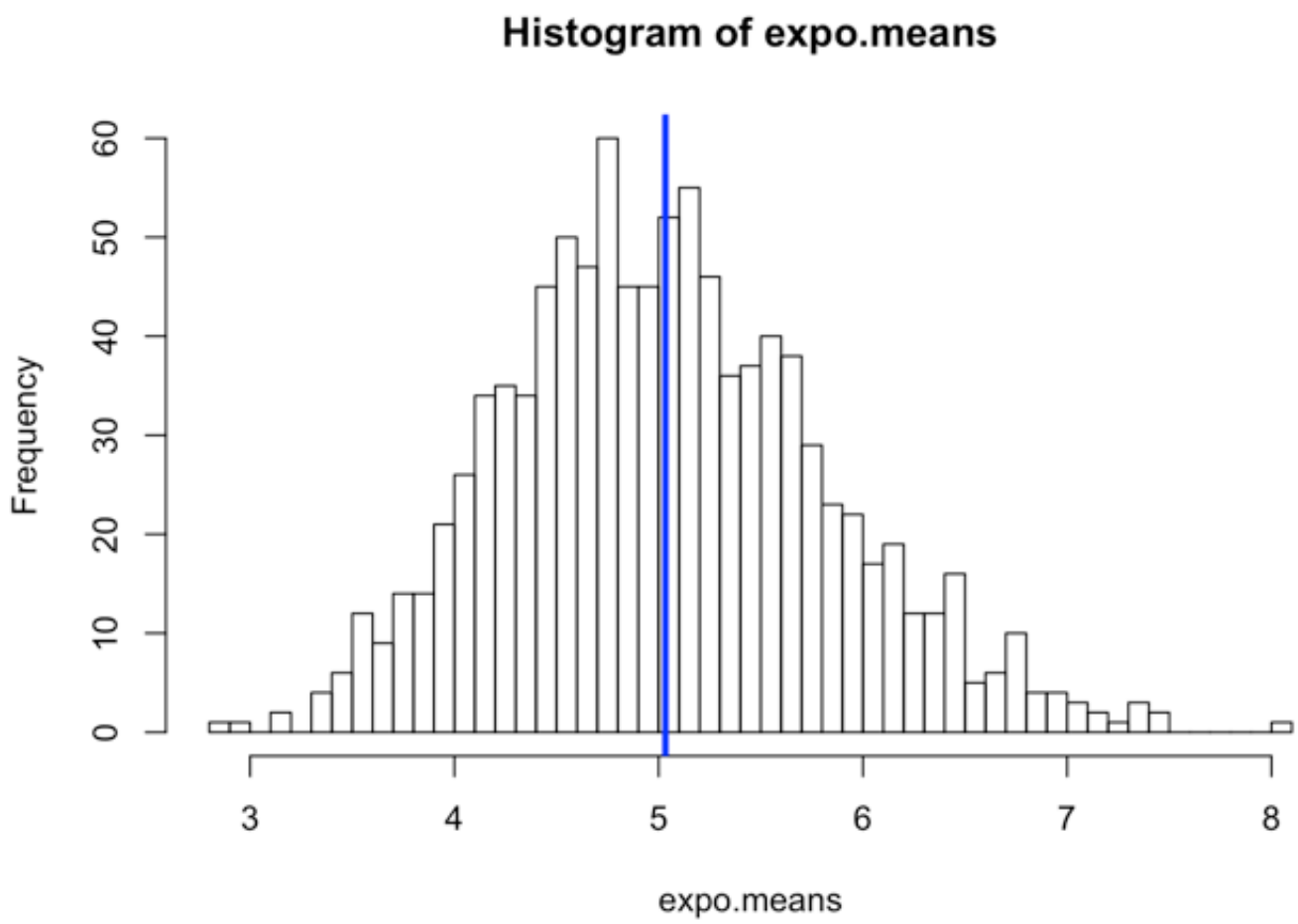
```
## [1] 0.625
```

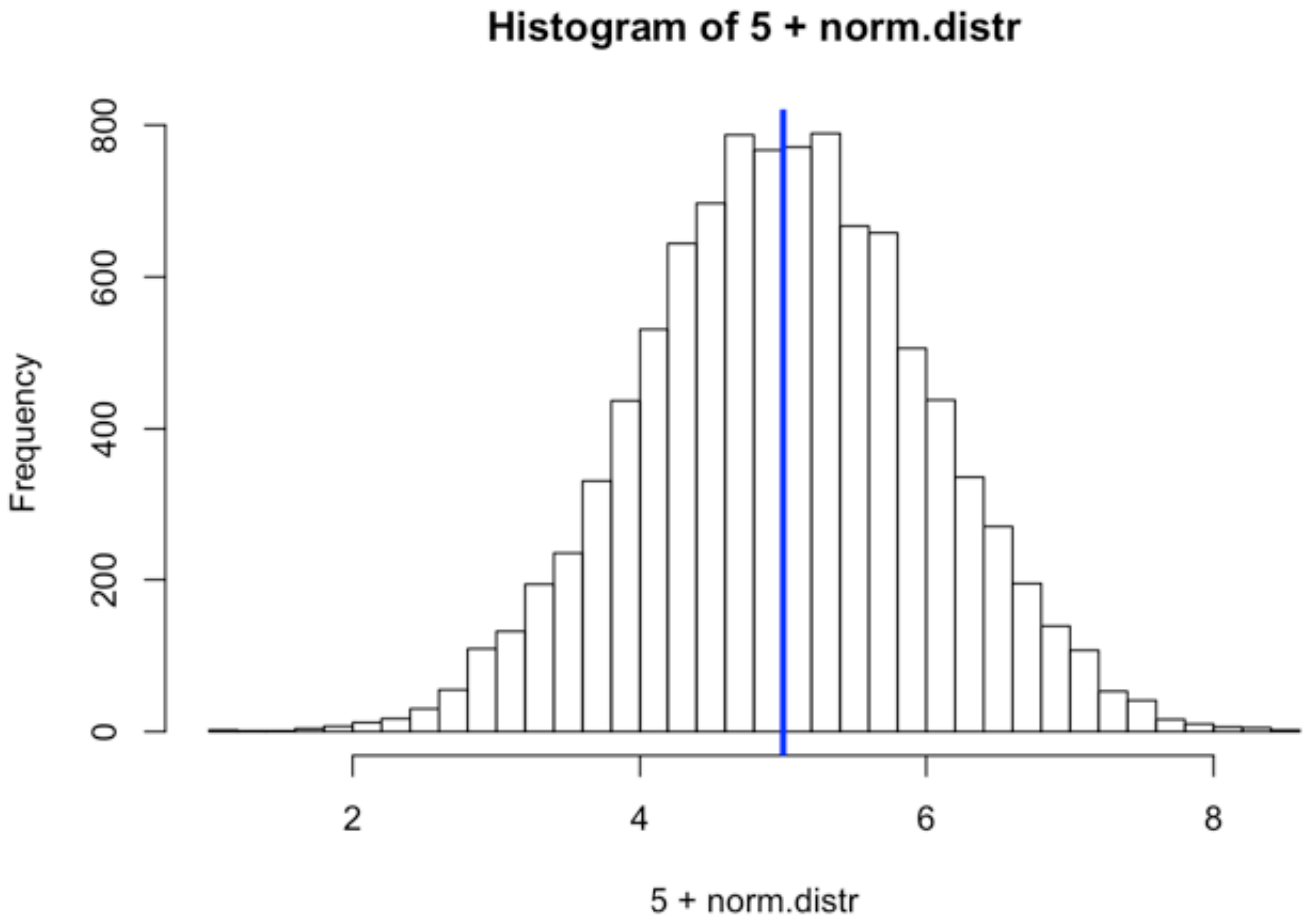
The sample and theoretical values are quite close, as we can see.

Explorations Part Two: A Visual Comparison

The first histogram below histogram shows the distribution of the sample means. The vertical bar marks the mean of the distribution of sample means.

For comparison purposes, we'll compute a random sample drawn from the normal distribution and show a histogram. We'll up the sample "n" to 10,000 to ensure it's close to the paradigm for a normal distribution. We'll also add 5 (theoretical mean of the sample exponential distribution) so the histograms' scales are comparable.





We can see that the distribution of even a small set of means of random sample data sets drawn from the exponential distribution is close to the normal paradigm.

Part 2: Tooth Growth Analysis

Overview

We will be investigating the ToothGrowth data set from the R datasets package.

The official description says:

> *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).*

Dietary supplements cost money. They also carry the risk of side effects, the likelihood of the occurrence of which typically increases as dosage increases.

We'll start by getting familiar with the data set, using exploratory statistical techniques. Then we posit some hypotheses with respect to the effects of choice of dietary supplement and dosage level.

Finally we'll summarize the conclusions we can draw from our investigation.

Setup

Let's intialize some variables we'll be using later on.

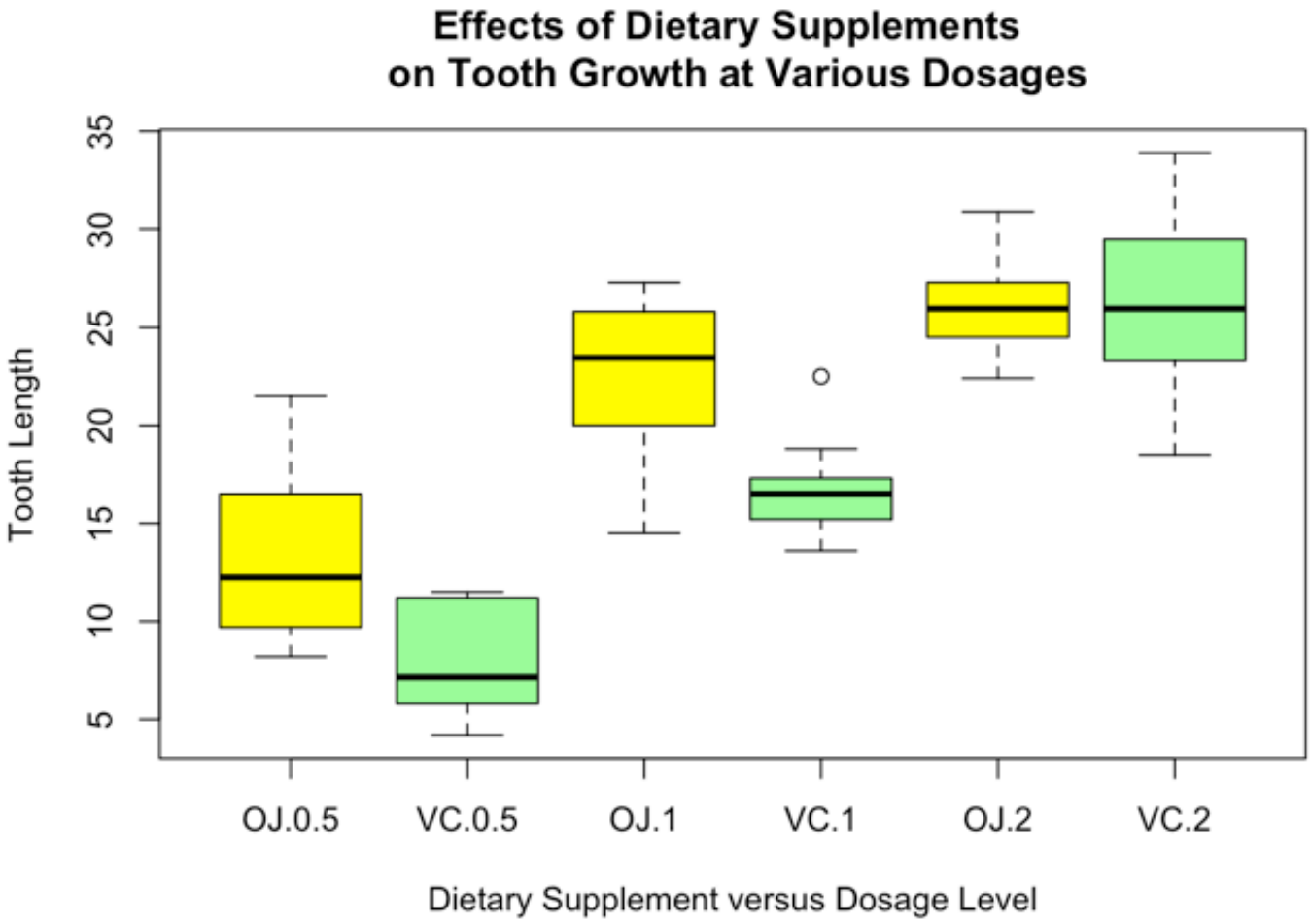
Exploratory Analysis

Now we'll produce a summary of the data set and see what we can learn from it.

```
##       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
```

We have 60 data points. 30 are associated with the supp (dietary supplement) factor "OJ" (referring to ascorbic acid in orange juice form), and the remaining 30 with the supp factor "VC" (referring to ascorbic acid - Vitamin C - in pure chemical form). Each row represents a tooth length measurement (len column), a dietary supplement (supp column), and a dosage (dose column)

Now let's do a box plot of the data set and see what we can learn from that.



In a box plot, the dotted-line "whiskers" represent the range of data values; the colored boxes represent the two quartiles between 25% and 75%, the middle 50% of the values; and the solid line dividing the colored box shows the median point in the value range.

From this chart, we see that orange juice (OJ) appears to be more effective at stimulating tooth growth at 0.5 and 1.0 mg/day, though in both cases it has wider value ranges than ascorbic acid (VC). At the 2.0 dosage level, the two supplements are much closer in their effects.

Another observation of interest is that the orange juice boxes for dosages of 1.0 and 2.0 mg/day overlap. We will want to look closely at these to find out whether or not the higher dose is significantly more effective. If not, the data may suggest that the lower dose is preferable, whether in terms of lower cost or reduced probability of side effects from the supplement.

We will need to do a deeper analysis to compare the two supplements at the three dosage levels, to provide more information to support any decision-making with respect to supplement and dosage level selection. To do this analysis, we will test some hypotheses against the data set.

Hypothesis Testing: Does Type of Supplement Matter?

We choose to use the Student's T-test statistic because the sample groups are small and the statistic is relatively robust with respect to the normalcy of the data. However, we are NOT making an assumption that the variances are approximately equal. The confidence intervals in the box plot imply considerable differences in variance

Our null hypothesis is that the two groups, OJ and VC, are the same at any given dosage level. Because the Tooth Growth chart above appears to show that OJ is more effective than VC, our alternative hypothesis is that the mean of the OJ group is significantly greater than the mean of the VC group with a confidence of 95%.

We will do three separate analyses to test the significance of the difference between the two treatments at each dosage level.

Dosage = 0.5 mg/day

Difference beween the group means (positive value shows that OJ mean is greater than VC mean):

```
## [1] 5.25
```

Orange juice appears to have a significant advantage over pure ascorbic acid, based both on the large difference in means and eyeballing the Tooth Growth chart.

T-Test of the alternative hypothesis that OJ mean significantly greater than VC mean:

P-value is 0.0031793
Upper limit of rejection region:
2.3460403

The mean of 5.25 is greater than the upper limit of the rejection region at 95% confidence level, so we can reject the null hypothesis. At a dose of 0.5 mg/day, ascorbic acid via orange juice is significantly more effective than pure ascorbic acid alone.

Dosage = 1.0 mg/day

Difference beween the group means (positive value shows that OJ mean is greater than VC mean):

```
## [1] 5.93
```

Orange juice still appears to have a significant advantage over pure ascorbic acid, based both on the large difference in means and eyeballing the Tooth Growth chart.

T-Test of the alternative hypothesis that OJ mean significantly greater than VC mean:

P-value is 5.191879410^{-4}
Upper limit of rejection region:
3.3561576

The mean of 5.93 is greater than the upper limit of the rejection region at 95% confidence level, so we can reject the null hypothesis. At a dose of 1.0 mg/day, ascorbic acid via orange juice is significantly more effective than pure ascorbic acid alone.

Dosage = 2.0 mg/day

Difference beween the group means (positive value shows that OJ mean is greater than VC mean):

```
## [1] -0.08
```

The difference is now close to zero (and actually leans in favor of pure ascorbic acid, in constrast to the smaller dosages). We can reasonably assume we will fail to reject the null hypothesis in this case, but let's do the test to make sure.

This time we will use a two-tail test with the null hypothesis that the difference in means is insignificant. This will test the more general case of the two means being significantly different. If the difference turned out to be significant, we would need to investigate more deeply to determine which treatment was the greater.

T-Test of the alternative hypothesis that OJ mean significantly greater than VC mean:

P-value is 0.9638516
Upper limit of rejection region:
-3.7980705 : 3.6380705

The p-value is much greater than the 95% cutoff of 0.05, so we can accept the null hypothesis, that the difference in means is insignificant. The null hypothesis's posited mean of 0 and the actual mean of the differences, {r mean(oj.group$len[which(oj.group$dose == 2.0)])-mean(vc.group$len[which(vc.group$dose == 2.0)]) }, are both well within the rejection region. At a dose of 2.0 mg/day, ascorbic acid via orange juice is NOT significantly more effective than pure ascorbic acid alone.

Hypothesis Testing: Does Dosage Matter?

Our analysis would be incomplete if we did not also look at the significance of the different dosage levels under each supplement.

Why do we care? We have seen thus far that a dietary supplement of ascorbic acid in orange juice is superior to pure ascorbic acid for lower dosage levels (0.5 and 1.0 mg/day), but the effectiveness of the two supplements is not significantly different at the higher dosage of 2.0 mg/day.

Let's assume that pure ascorbic acid is less expensive per milligram than its equivalent in orange juice. The most cost-effective approach would then be pure ascorbic acid at 2.0 mg/day – but only if the gain in growth rates between, for example, 1.0 and 2.0 mg/day were significant enough to justify the increased cost.

A deeper analysis would compare the cost and benefit of the different combinations of supplement and dosage in terms of growth rate improvements. Perhaps guinea pig teeth are valuable in Chinese medicine, in which case benefit will trump cost in the decision-making process. This analysis is out of the scope of this

project, since we have been given neither the cost of supplements nor the means to evaluate the potential sale price of guinea pig teeth.

Staying within our scope, we will look at the significance of length gains length for each dosage level and the next higher increment in dosage. We will do this by supplement, treating it as a potentially confounding variable.

For each supplement, we will compare 0.5 versus 1.0 mg/day and 1.0 versus 2.0 mg/day.

Our null hypothesis throughout is that the difference between the two dosage levels under consideration is insignificant with 95% confidence.

Orange Juice (OJ)

0.5 versus 1.0 mg/day

```
## [1] 8.784919e-05
```

The p-value is less than 0.05 and positive, hence the higher dosage is significantly more effective at stimulating tooth growth.

1.0 versus 2.0 mg/day

We recall from the box plot chart that there was considerable overlap between the boxes for orange juice at these two dosages. This test will provide insight into whether the higher dosage is actually significantly more effective.

```
## [1] 0.03919514
```

Once again, the p-value is less than 0.05 and positive, hence the higher dosage is significantly more effective at stimulating tooth growth.

Pure Ascorbic Acid (VC)

0.5 versus 1.0 mg/day

```
## [1] 6.811018e-07
```

In this case as well, the p-value is less than 0.05 and positive, hence the higher dosage is significantly more effective at stimulating tooth growth.

1.0 versus 2.0 mg/day

```
## [1] 9.155603e-05
```

And also in our final test case, the p-value is less than 0.05 and positive, hence the higher dosage is significantly more effective at stimulating tooth growth.

Conclusions Of Tooth Growth Analysis

- The higher the dosage the better in terms of stimulating tooth growth, regardless of supplement. For both supplements, the difference between each incrementally adjacent dosage pair (0.5-1.0 and 1.0-2.0 mg/day) is significant.
- At lower dosages (0.5 and 1.0 mg/day), the performance of orange juice is significantly better than pure ascorbic acid. The box plot showed that the difference is most pronouced at the 1.0 dosage level.
- At the highest dosage level tested, 2.0 mg/day, the performance of orange juice and pure ascorbic acid are roughly equivalent.