StatInf_Proj1

Brett McLaughlin

Tuesday, January 20, 2015

Section 1: Exponential Distribution Analysis

Overview: This project reviews the exponential distribution. We simulate 1000 averages of 40 random, exponentially-distributed random variables, and then examine its properties.

Part 1a: Populate the sample sets.

```
library(ggplot2);
X_mns40 <- NULL;
X_lambda <- 0.2;
X_count <- 1000;
X_sample <- 40;
for(i in 1:X_count) X_mns40 = c(X_mns40, mean(rexp(X_sample, X_lambda)));
```

Part 1b: Sample Means & Variances

Here, we inspect our Sample Means, and compare them to the theoretical mean of the distribution.

Given that we are using a lambda of 0.2, that suggests that our mean should be theoretically centered at 5 (since on an exponential distribution, the mean is 1/lambda). As our sample sizes grow, the variance of the averages of the groups approaches zero.

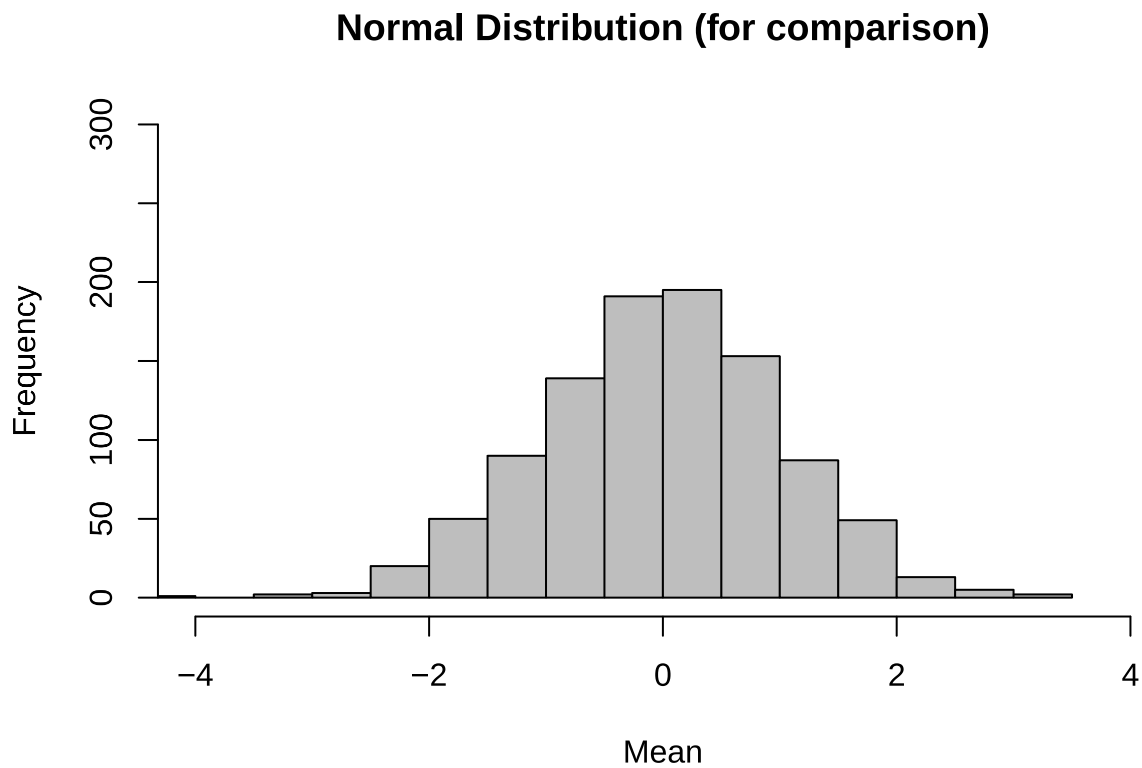
We can see from the results below that as the sample set gets bigger, our mean stays the same (always centered on 5), but the tails move in, and the standard deviation moves towards zero – just like we expected!

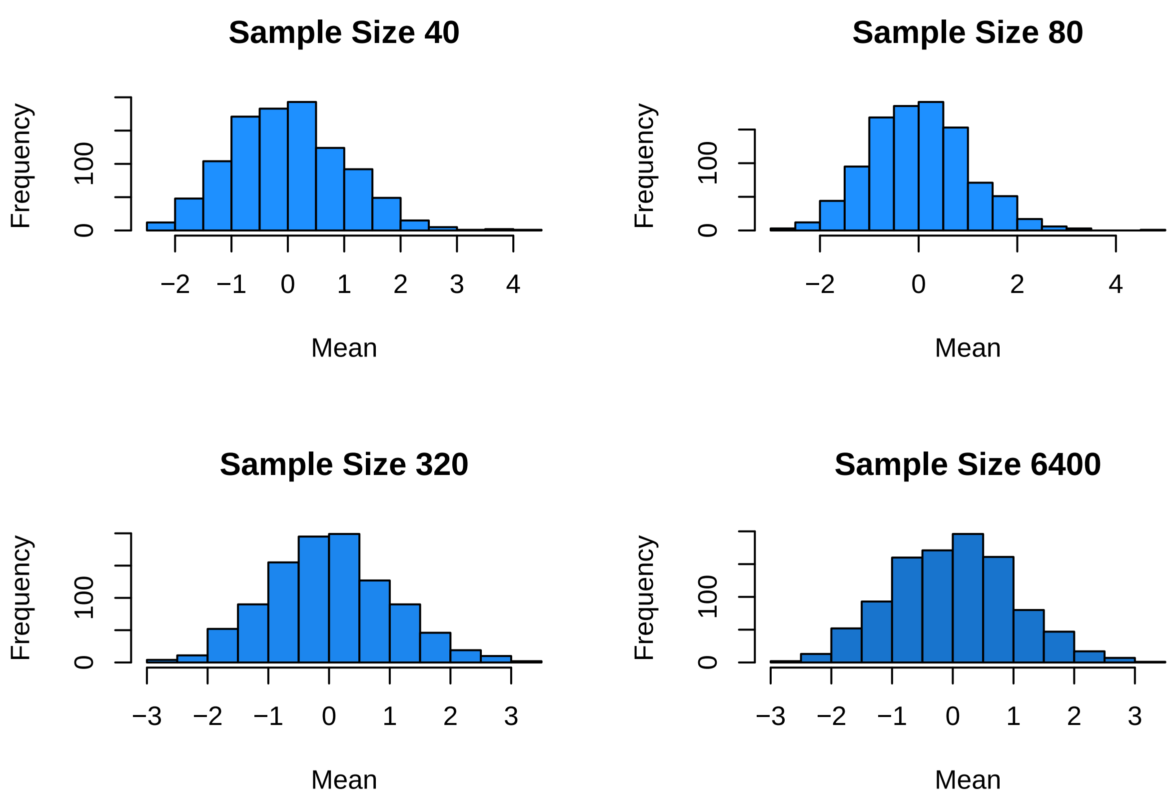
```
result_table
```

```
##      Sample Size Average Average Min Average Max Average     Variance
## [1,]          40        5.025345    3.053688    8.194210 0.6236261558
## [2,]          80        5.007388    3.440557    7.497646 0.3025938259
## [3,]         320        5.002306    4.238039    5.957874 0.0768844541
## [4,]       64000        5.000130    4.943072    5.061767 0.0003867844
```

Part 1c: Normal Form

Below, I convert our groups of 1000 samples (based upon averages of 40, 80, 320 and 64000) to "normal form" - subtracting the mean and dividing by the standard deviation – so that our values are centered around 0 and have a standard deviation of 1. Let's see if the results are distributed normally, and how it changes as our average sample size grows.


Normal Distribution (for comparison)
Frequency
0
50
100
200
300
−4
−2
0
2
4
Mean


Sample Size 40
Frequency
0
100
−2
−1
0
1
2
3
4
Mean
Sample Size 80
Frequency
0
100
−2
0
2
4
Mean
Sample Size 320
Frequency
0
100
−3
−2
−1
0
1
2
3
Mean
Sample Size 6400
Frequency
0
100
−3
−2
−1
0
1
2
3
Mean

All of our distributions are in normal form – you can see that each is symmetrical and drops off evenly to both sides.

Section 2: ToothGrowth

Part 2a: Populate the data.

```
# This didn't really warrant its own section...
data(ToothGrowth);
```

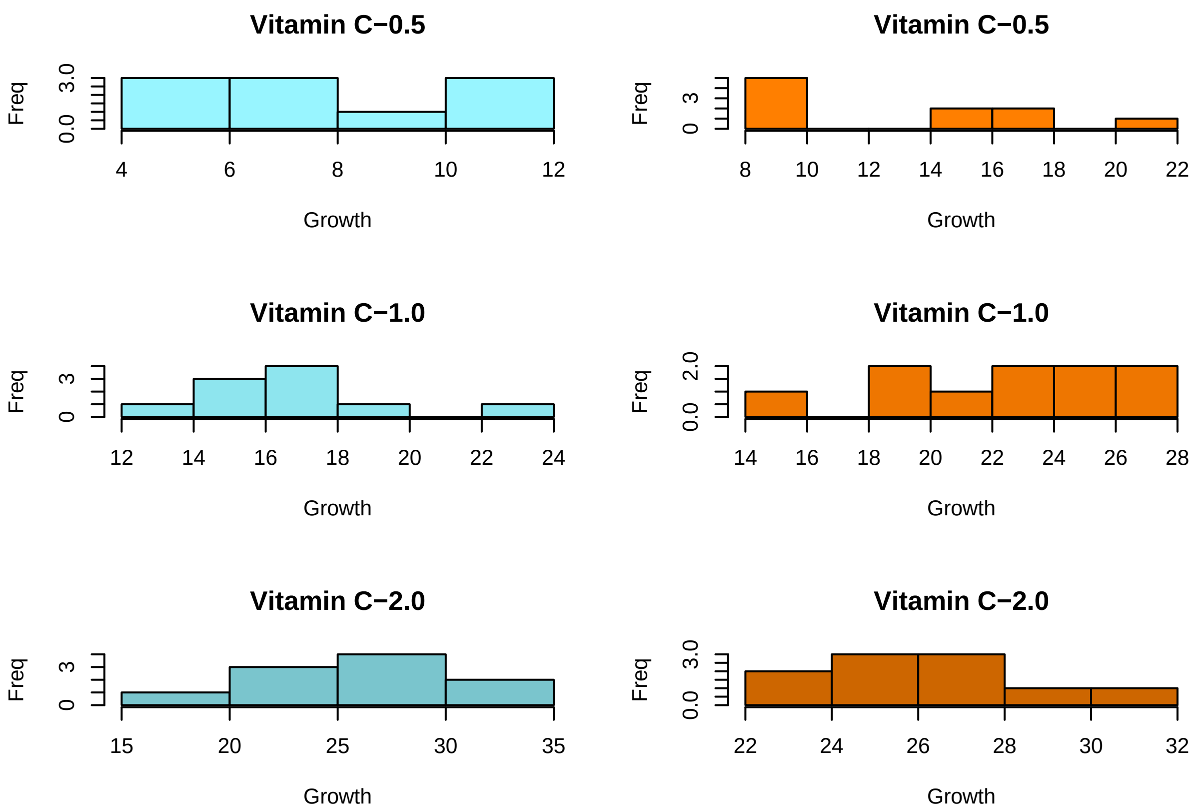
Part 2b: Initial data inspection.

First, let's load the data and generate some general totals: min, max, mean, count, and our 95% confidence intervals (so 2.5% low and 97.5% high).

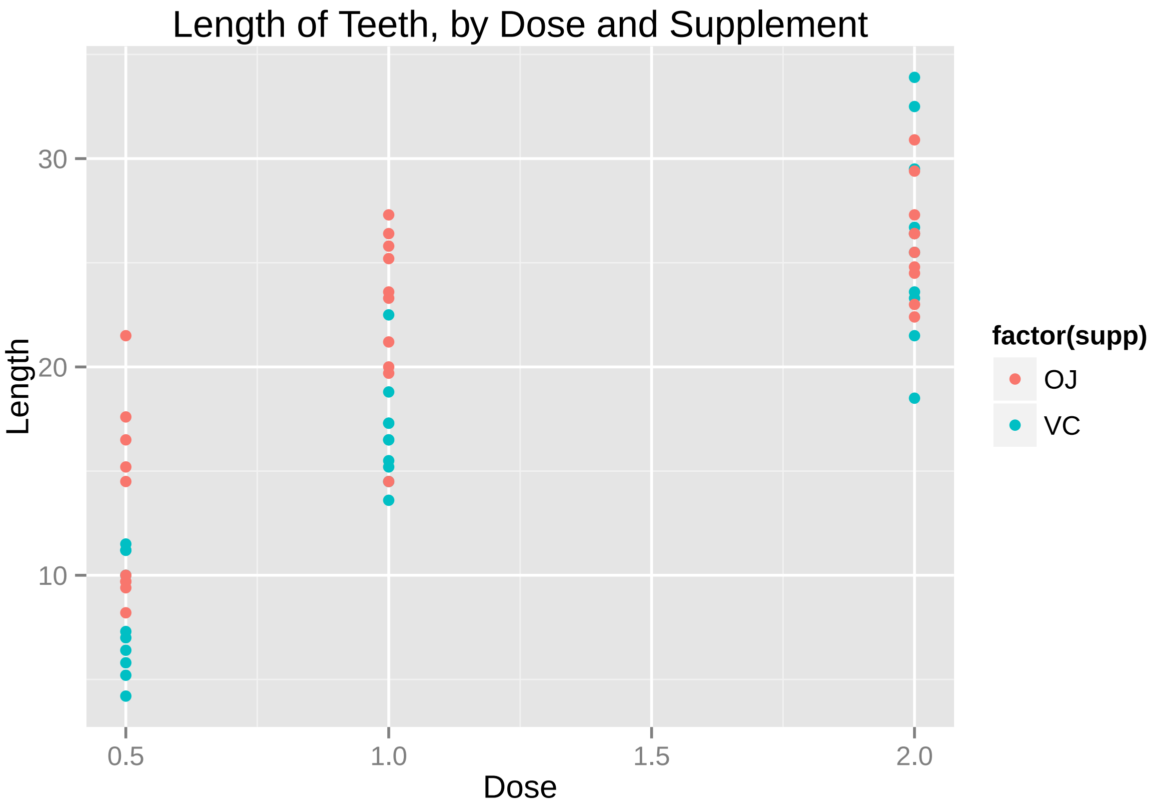
Because our test sets are so small, we should use t.test to calculate the confidence intervals.

```
##      Supplement Dose Min  Max  Mean  Count 2.5%  97.5%
## [1,] VC         0.5  4.2  11.5 7.98  10    6.02  9.94
## [2,] VC         1    13.6 22.5 16.77 10    14.97 18.57
## [3,] VC         2    18.5 33.9 26.14 10    22.71 29.57
## [4,] OJ         0.5  8.2  21.5 13.23 10    10.04 16.42
## [5,] OJ         1    14.5 27.3 22.7  10    19.9  25.5
## [6,] OJ         2    22.4 30.9 26.06 10    24.16 27.96
```

Now let's look at the distributions of each set.



. . . For what it's worth, the distributions are not normal. Other than that, it doesn't look very interesting. Let's scatter plot the results.



Part 2c: Results/Conclusions

- Both Vitamin C and Orange Juice, even in small doses, are correlated with tooth growth. And higher doses of either leads to more growth.
- No matter what the Supplement/Dose, the results are not normally distributed.
- For doses of .5 or 1.0, Orange Juice is correlated with higher tooth growth than Vitamin C. For these doses, OJ's lower confidence interval is higher than Vitamin C's confidence interval – so it's extremely likely to yield more tooth growth.
- For doses of 2.0, Orange Juice and Vitamin C return similar growth rates.

Assumptions:

- It would be nice to have a Control Group, with no Vitamin C. We're implicitly assuming that no supplements would yield NO tooth growth, which is probably not true.
- I'm not sure if we can assume a shared variance across unique Supp/Doses, so I'm treating each Supp/Dose as an independent set. If we did assume a shared variance, we'd likely get narrower confidence intervals.