Practical Analysis of the CLM and Hypothesis Testing

Kate Prendergast

Part 1

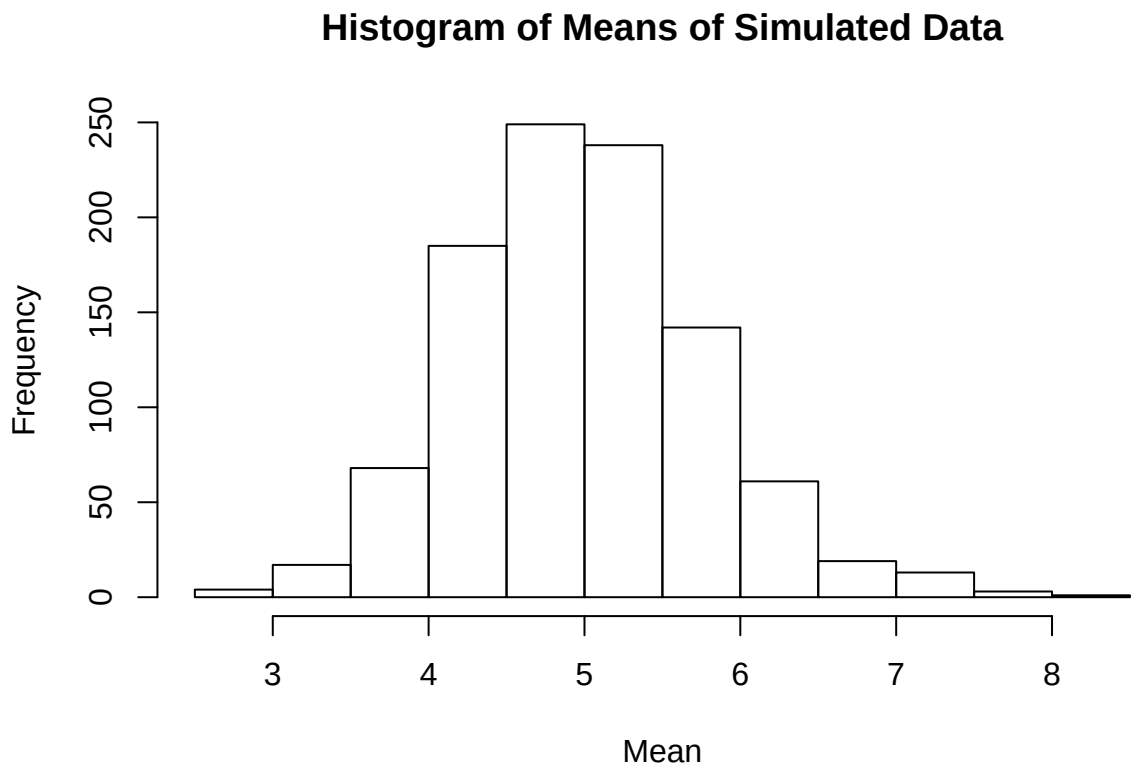
Overview

In Part 1, we will examine the distribution of means of 40 random exponentials, using the exponential distribution. This will be based on running 1000 simulations, where each one takes the mean of 40 random exponentials. We will analyze the mean and variance of our simulated data and look at the distribution.

Simulations

We first create our 1000 simulations, taking the mean of 40 exponentials in each. For this analysis, we will use a lambda of 0.2.

```
## set the seed for the sake of reproducibility of this report
set.seed(22)
mns = replicate(1000, mean(rexp(40, 0.2)))
hist(mns, xlab = "Mean", main = "Histogram of Means of Simulated Data")
```



We can see that the sample means are centered around 5.

Sample Mean vs. Theoretical Mean

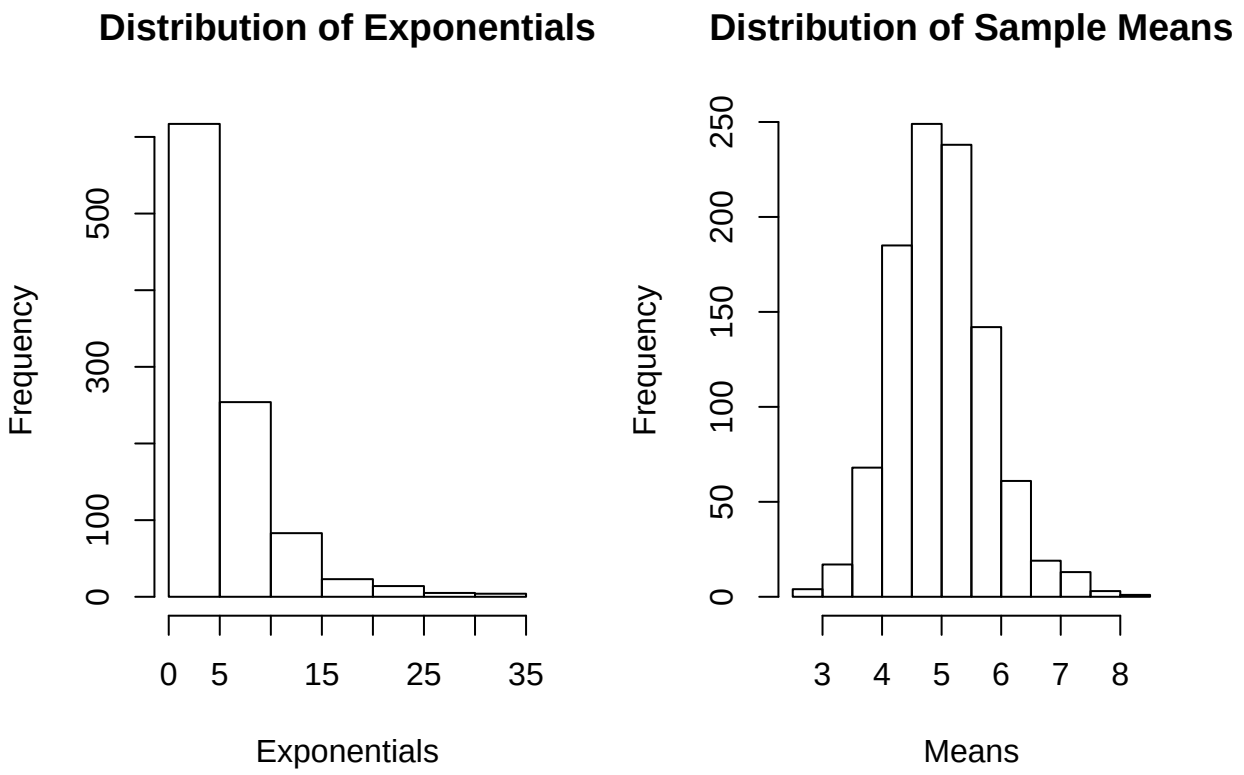
We now calcualte the exact mean of our data, which gives us **4.989191**. The theoretical mean of the exponential distribution (1/lambda) gives an expected mean of **5**. Hence the mean of our simulated data is very close to the expected value.

Sample Variance vs. Theoretical Variance

Next, we calculate the variance of our simulated data, which is **0.6337023**. The theoretical variance of the exponential distribution (with n = 40 draws from the population) is (1/n) * (1/lambdaˆ2), which gives us **0.625**. Once again, the variance of our simulated data is very close to the expected value.

Distribution

Finally, we see if the distribution of our sample means is approximately normal. We show a comparison of a plot of 1000 random exponentials (using our same lambda of 0.2) and the same histogram of our means of our simulations of 40 samples.



As expected, the first plot shows the frequency of values decreasing exponentially, with the majority of values near 5. Our second plot also shows the means centered around 5, however in this plot we see that our distribution of means takes on a Gaussian shape and is normally distributed.

Summary

In summary, our analysis shows that the simulated data does behave as predicted by the CLT. Our mean and variance approach the theoretical values, and the distribution of our sample means is approximately normal.

Part 2

Overview

In Part 2, we will examine a fairly small dataset that shows tooth growth results from Vitamin C, using one of two delivery methods and given at three different doses. We will examine whether the data appears to be normally distributed, and if so, determine if the differences in resutls appear to be statistically significant.

Data Summary

We first load the ToothGrowth dataset and use the stat.desc() function to perform some initial analysis of the data. We are most interested in the Shapiro-Wilk normality test results, shown below, but see the appendix for the full analysis.

```
##                   len supp         dose
## normtest.p 0.1091005   NA 1.990132e-08
```

Based on the Shapiro-Wilk normality test statistic p-value of 0.11 calculated for this dataset and an assumed significance level of 0.05, we will not reject the hypothesis that the data is normally distributed and can proceed with confidence estimates using a t-distribution.

Testing

We start by plotting the results by dose and delivery method. Dose values are 0.5, 1, and 2. Delivery methods are OJ and VC, which stand for Orange Juice and Ascorbic Acid, respectively.

See Figure 1 in the appendix for the accompanying boxplot.

There doesn't appear to be significant differences by delivery method, but the dose does seem to affect the length of tooth growth. Let's further examine the differences in results by dose usingt-test confidence intervals.

After subsetting the data by dose (see Appendix for code) so that we can compare two at a time, we then compare the 0.5 to 1.0 doses, 1.0 to 2.0, and 0.5 to 2.0 We look at the confidence interval for each comparison, which will show with 95% confidence what the difference in the mean is between the two data sets involved.

```
rbind(as.numeric(t.test(dose05$len, dose1$len)$conf.int),
as.numeric(t.test(dose1$len, dose2$len)$conf.int),
as.numeric(t.test(dose05$len, dose2$len)$conf.int))
```

```
##            [,1]       [,2]
## [1,] -11.983781  -6.276219
## [2,]  -8.996481  -3.733519
## [3,] -18.156167 -12.833833
```

In all 3 cases, since 0 is not in the interval, we can say with 95% confidence that the difference in the means is significant (non-zero) and that dose does affect the tooth growth results.

Appendix

Part 1

Code for calculating the sampling mean and variance.

```
simMean <- mean(mns)
simVar <- var(mns)
```

Code for creating the side-by-side comparison graphs.

```
par(mfrow=c(1,2))
hist(rexp(1000, 0.2), xlab = "Exponentials", main = "Distribution of Exponentials")
hist(mns, xlab = "Means", main = "Distribution of Sample Means")
```

Part 2

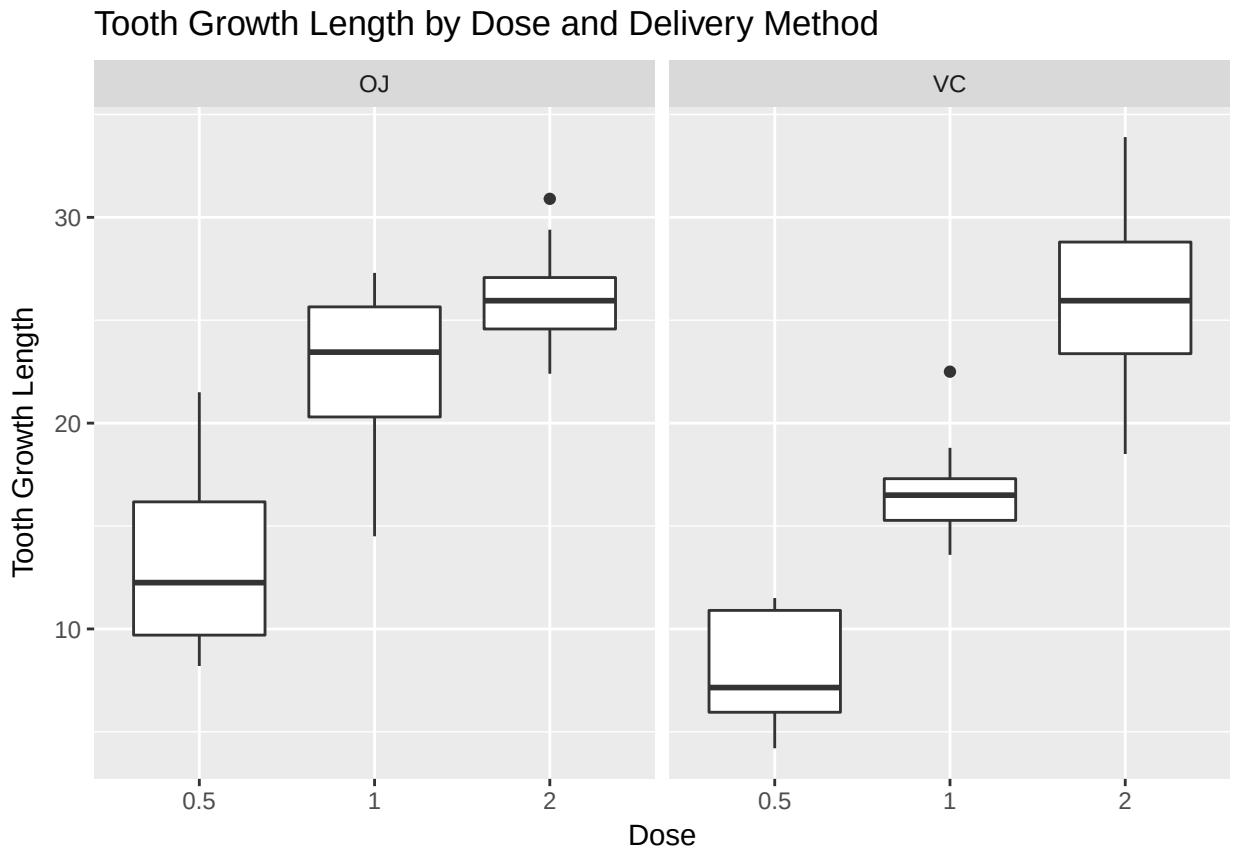
Code for loading the ToothGrowth data and performing a basic statistical analysis.

```
data("ToothGrowth")
stat.desc(ToothGrowth, norm = TRUE)
```

```
##                      len supp          dose
## nbr.val        60.0000000   NA  6.000000e+01
## nbr.null        0.0000000   NA  0.000000e+00
## nbr.na          0.0000000   NA  0.000000e+00
## min             4.2000000   NA  5.000000e-01
## max            33.9000000   NA  2.000000e+00
## range          29.7000000   NA  1.500000e+00
## sum          1128.8000000   NA  7.000000e+01
## median         19.2500000   NA  1.000000e+00
## mean           18.8133333   NA  1.166667e+00
## SE.mean         0.9875223   NA  8.118705e-02
## CI.mean.0.95    1.9760276   NA  1.624549e-01
## var            58.5120226   NA  3.954802e-01
## std.dev         7.6493152   NA  6.288722e-01
## coef.var        0.4065901   NA  5.390333e-01
## skewness       -0.1425376   NA  3.722966e-01
## skew.2SE       -0.2308721   NA  6.030190e-01
## kurtosis       -1.0425144   NA -1.549583e+00
## kurt.2SE       -0.8566377   NA -1.273298e+00
## normtest.W      0.9674286   NA  7.649050e-01
## normtest.p      0.1091005   NA  1.990132e-08
```

Figure 1 - code and boxplot

```
library(ggplot2)
qplot(factor(dose), len, data = ToothGrowth, geom = "boxplot", facets = .~supp,
      xlab = "Dose", ylab = "Tooth Growth Length",
      main = "Tooth Growth Length by Dose and Delivery Method")
```



Code for subsetting tooth decay data.

```
dose05 <- subset(ToothGrowth, dose == 0.5)
dose1 <- subset(ToothGrowth, dose == 1.0)
dose2 <- subset(ToothGrowth, dose == 2.0)
```