Statistical Inference Course Project

Dylan Hazlett

Thank you for reviewing my final report for the Statistical Inference Course. In Part 1, we will simulate a collection exponential variables, then compare the sample's summary statistics to their theoretical values and distribution. In Part 2, we will analyze the differences in tooth-growth among varying doses and delivery methods of Vitamin C using the ToothGrowth dataset in R.

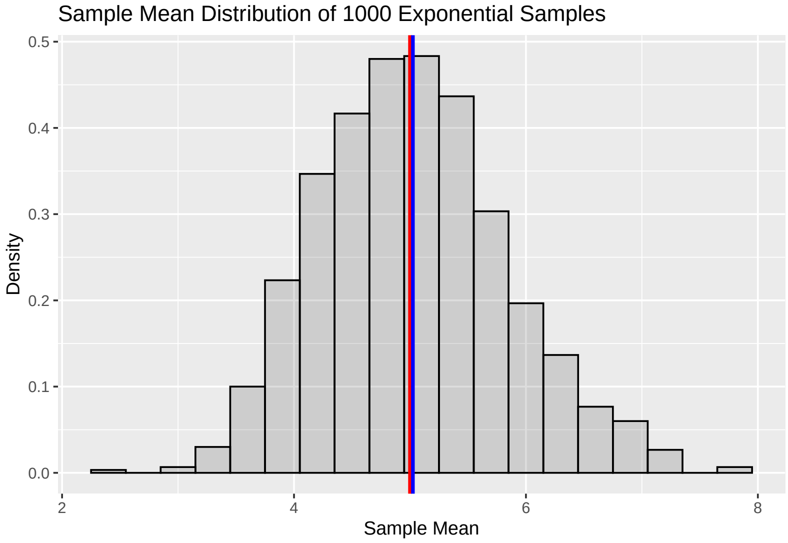
Part 1: Simulation of Expontentially Distributed Variables

```r
mns = NULL
for (i in 1 : 1000) {
  mns = c(mns, mean(rexp(40, .2)))
}
```

The above block of code simulates the distribution of averages of 40 exponentials with lambda value .2, 1000 times. After initializing 'mns' to record the means, a for loop runs the exponential distribution expressions and gathers each mean 1000 times. The output is an array of 1000 independently sampled means.

```
## [1] "Using the mean() function on mns, we calculate the average Sample Mean: 5.02528893363976"
```

Theoretically, and with enough simulations, this value will converage to the mean of the exponential distribution, 1/lambda or 5 since lambda = .2.



This figure shows the distribution of the 1000 sample means with two vertical lines, the average sample mean in blue and the theoretical mean in red. The two may appear to be touching on the chart due to how close they are in value.

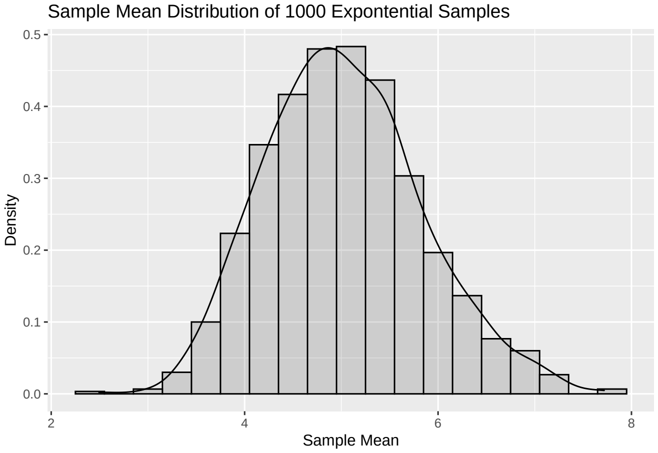
```
## [1] "Using the var() function, we calculate that the Sample Variance is 0.651939499915466"
```

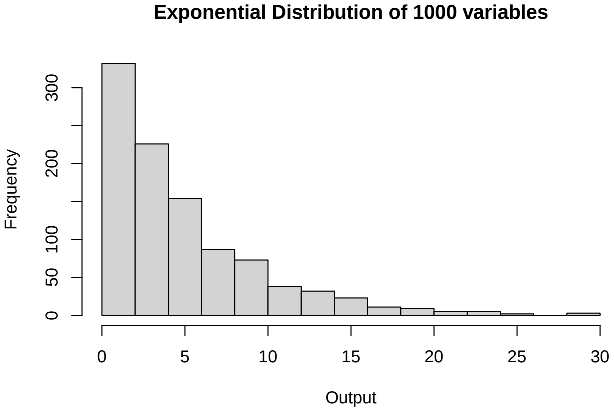
To compare the sample variance to its theoretical value for the exponential distribution, we use the following formula given that n equals 40 and sigma equals the standard deviation, given that the standard deviation of the exponential distribution is 1/lambda where lambda equals 0.2. $\bar{V}ar = \sigma^2/n$

```
## [1] "The Theoretical Variance for 40 exponential variables where lambda equal 0.2 is 0.625"
```

Given infinite simulations, the sample variance will converge to this value of 0.625.

This simulation is an example of the Central Limit Theorem, where the mean samples resulting from the simulations are normally distributed around the theoretical mean.



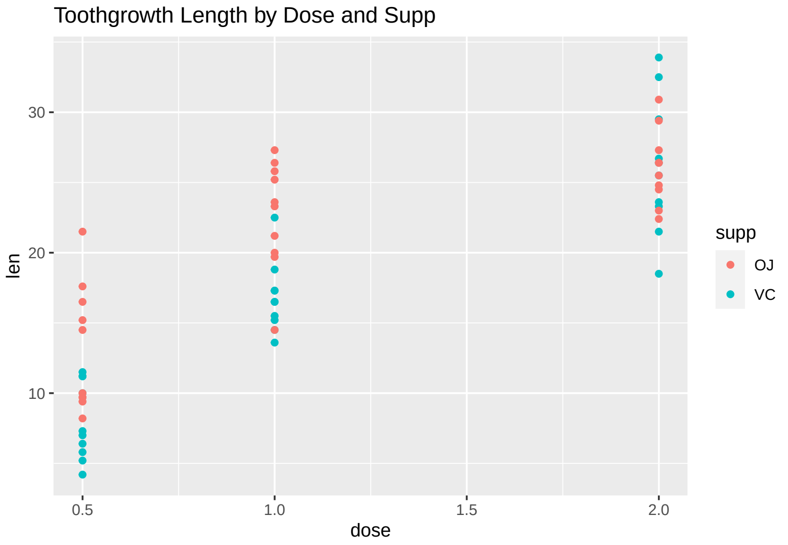


The exponential distribution is of course not normally distributed; but, after taking 1000 random and independent samples of the mean of this distribution, we get the below distribution that appears to be a normal Guassian distribution of samples. Although the right tail of this distribution stretches further from the mean than that of the right, that is expected due to the nature of the exponential distribution stretching to infinity on the x-axis. We accept this drawback and conclude that the distribution is approximately normal. Using the central limit theorem, we assume iid, meaning our exponential variables were sampled randomly and independently of one another. We also assume that the sample size of 40 independent variable was sufficiently large, which is a safe assumption due to the resulting distribution chart.

Part 2: ToothGrowth

Now, for Part 2, lets load and explore the ToothGrowth dataset in R.

As stated in the R help file, the ToothGrowth dataset has 60 data entries of tooth length (len), dosage mg/day (dos), and supplement type (supp). With two supplements (VC or OJ) and 3 dosage levels (0.5, 1, and 2 mg/day), there are 10 data entries to each of the 6 group combinations of dosage and supplement. Since these entries are not paired, any comparison between these groups will use the mean.



Here is a visual comparison of all of the length data points split by dosage on the x-axis and supplement for color. We can see that, at a dosage of 2 mg/day, there is little difference in the distribution of length among supplements. We will keep this and our earlier hypothesis in mind as we now construct confidence intervals for all 6 groups.

```
ToothGrowth%>% group_by(dose, supp)%>%summarise(mean = mean(len),
    lower_bound = mean(len)-(qt(.975,9)*sd(len)/ sqrt(n())),
    upper_bound = mean(len)+(qt(.975,9)*sd(len)/ sqrt(n())))
```

```
## `summarise()` has grouped output by 'dose'. You can override using the `.groups` argument.
```

```
## # A tibble: 6 x 5
## # Groups:   dose [3]
##    dose supp   mean lower_bound upper_bound
##   <dbl> <fct> <dbl>       <dbl>       <dbl>
## 1   0.5 OJ    13.2        10.0        16.4
## 2   0.5 VC     7.98        6.02        9.94
## 3   1   OJ    22.7        19.9        25.5
## 4   1   VC    16.8        15.0        18.6
## 5   2   OJ    26.1        24.2        28.0
## 6   2   VC    26.1        22.7        29.6
```

The figure above shows the 95% t confidence interval for each of the 6 testing groups. We can conclude that for any of the intervals that have overlapping ranges, a t test will not have a significant result at 95% confidence. We can make several significant conclusions from the individual group comparisons that do not intersect eachother's t confidence interval, but it may be more useful to go over the higher level implication with this dataset. It would appear that, at lower doses (0.5 and 1.0), the OJ supplement has a significant greater effect on tooth growth than that of the VC supplement; however, at the highest dosage of 2.0, the tooth growth differences between supplements are negligable. Though tooth growth for the OJ groups does increase as the dosage increases (we are more than 95% confident that there is a length difference between OJ groups of 0.5 mg/day and 2.0mg/day), lengths at consecutive doses are not different at 95% confidence; whereas VC supplement groups show steep, significant differences between all dosage levels. Because the data groups here are small, 10 subjects in each, we must be cautious with definitive conclusions about the population.

Lastly, let's run a t test between supplements on all subjects with equal variances set to False: Given OJ, Var(len) = 43.63 Given VC, Var(len) = 68.33 Note: This test cannot be preformed with dosage as there are more than two groups.

```
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, data = ToothGrowth)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC 
##         20.66333         16.96333
```

Using the t test result where 0 difference is inside the 95% confidence interval, we cannot conclude with 95% confidence that tooth growth is different between supplement groups.

To use the t confidence intervals and the t test we assumes that the data are iid and have a roughly symmetrical distribution.