Statistical Inference Peer Graded Assignment

Ryan Gavin

3/16/2017

This document explores some of the topcis covered in the *Statistical Inference* course. It is the final assignment.

In **Part 1** we explore characteristics of the *exponential distribution* and how sample means and sample variance distributions of the *exponential distribution* conform to the *Central Limit Theorem.* In **Part 2**, an examination of the data collected during the *The Statistics of Bioassay** study takes place, including exploratory data analysis and statistical inference.

R Requirements

The following `R` libraries are required for our analysis.

```r
require(ggplot2)
require(dplyr)
require(reshape2)
```

Part 1

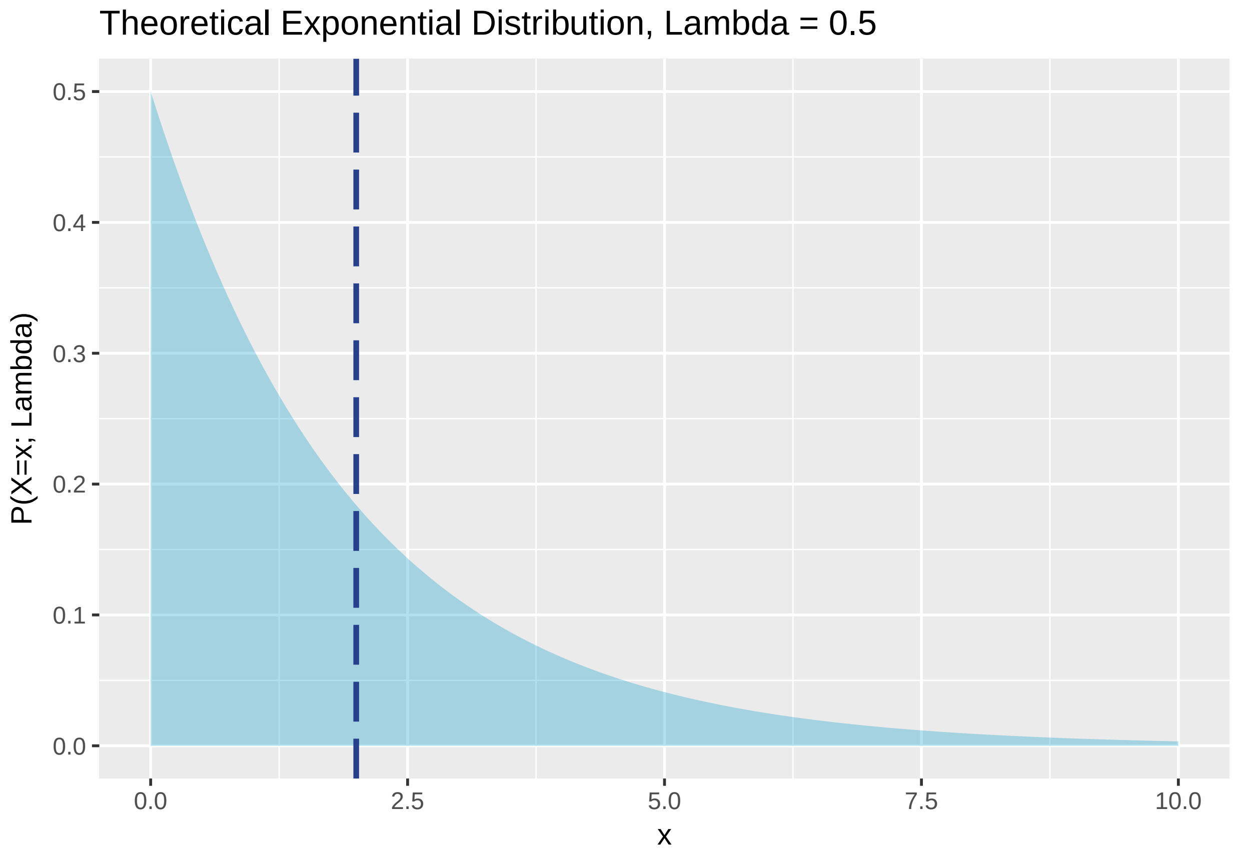
In **Part 1** we investigate the *exponential distribution* and the *Central Limit Theorem (CLT).* We will simulate events taken from the exponential distribution, repeat the observation, and draw some conclusions from the results.

Exponential Distribution

The exponential distribution (a.k.a. negative exponential distribution) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate.** It is described by the single parameter λ and the probability density function takes the following form:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0\,, \\ 0 & x < 0\,. \end{cases}$$

The mean of the exponential distribution is $1/\lambda$ and the standard deviation is $1/\lambda$. The mean is equal to the standard deviation. Below is an example of the distribution with $\lambda = 0.5$. The *mean* has been included - a vertical, blue dotted-line located at $x = 2$ ($\bar{X} = 1/\lambda = 1/0.5 = 2$).



Simulations

To investigate this distribution we'll run some simulations, randomly pulling events from an exponential. A single sample will consist of `n = 40` observations. We will repeat this sampling `B = 1000` times.

We will also set $\lambda = 0.2$ so that $\bar{X} = \sigma = 5$.

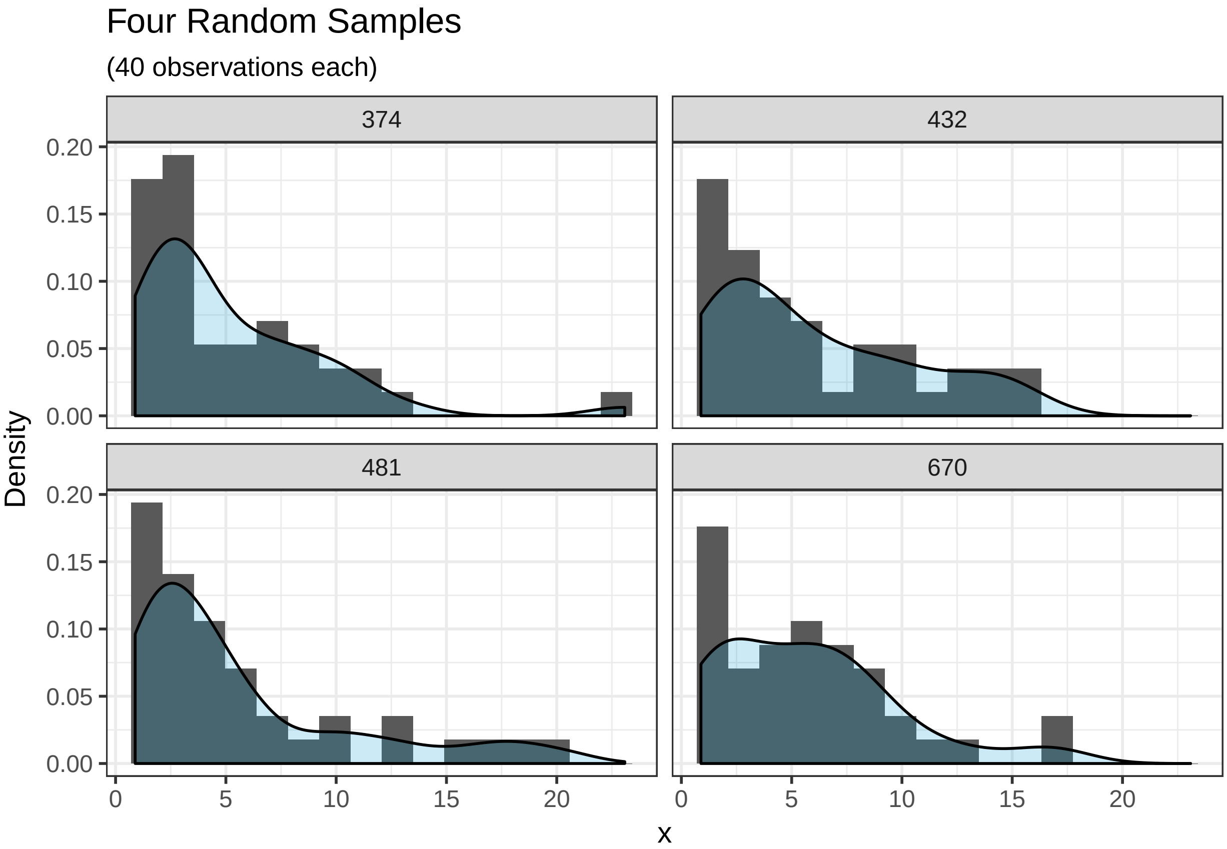
```
set.seed(1002)

lambda <- 0.2
n <- 40
B <- 1000

exp_sim <- matrix(rexp(n*B,lambda),nrow = B,ncol = n)
row.names(exp_sim) <- 1:B
```

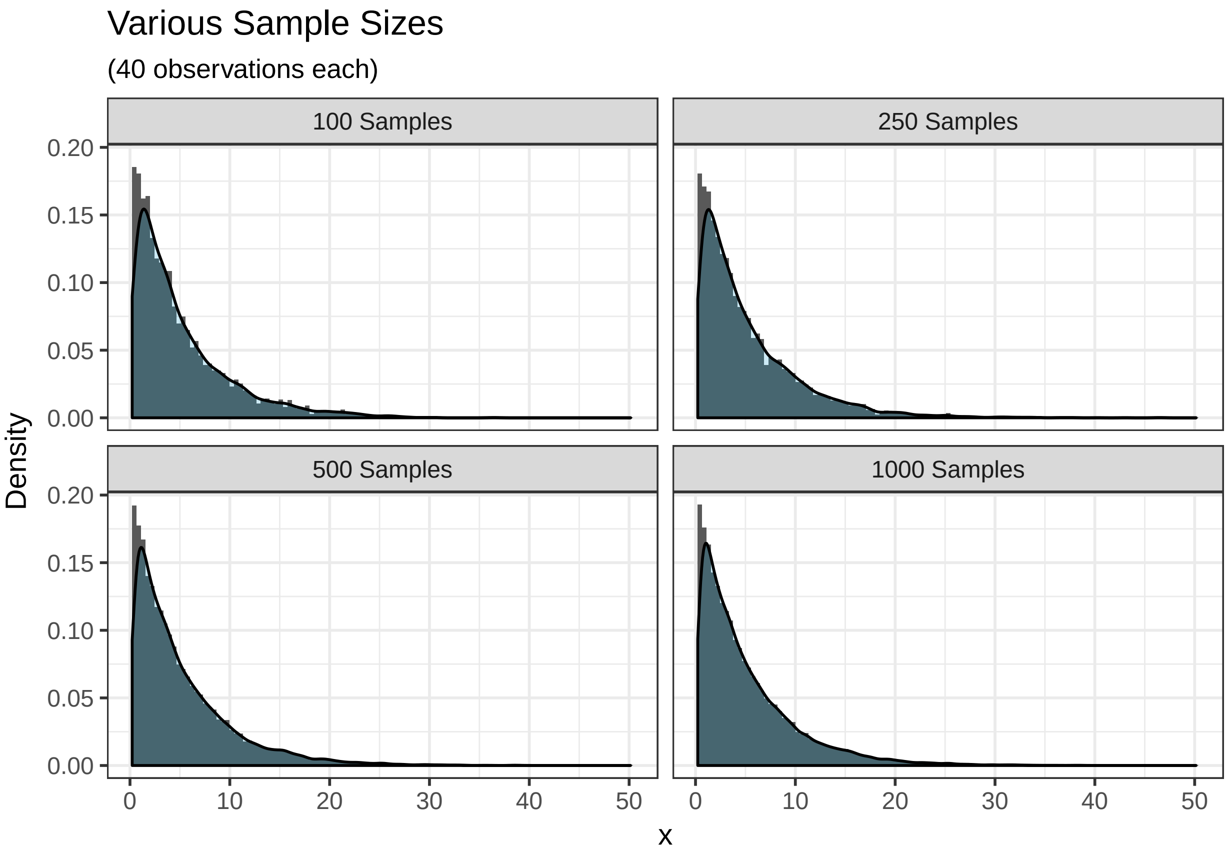
Examples

Let's examine these simulations a bit closer. We randomly picked four samples out of the one thousand generated and show their distributions below.



Although 40 observations for each sample isn't much, we can see that they start to resemble the exponential distribution. Density curves have been added to help illustrate the shape.

As a comparison, we show four more distributions of cumulative random samples: 100 samples, 250 samples, 500 samples, and all 1000 samples.



It becomes clear quite quickly that our samples added together resemble the exponential population they were taken from. The distribution becomes smoother and more like the exponential distribution with the more samples that are included, until we include all 1000 samples, for a total of 40000 observations. As before, density curves have been added.

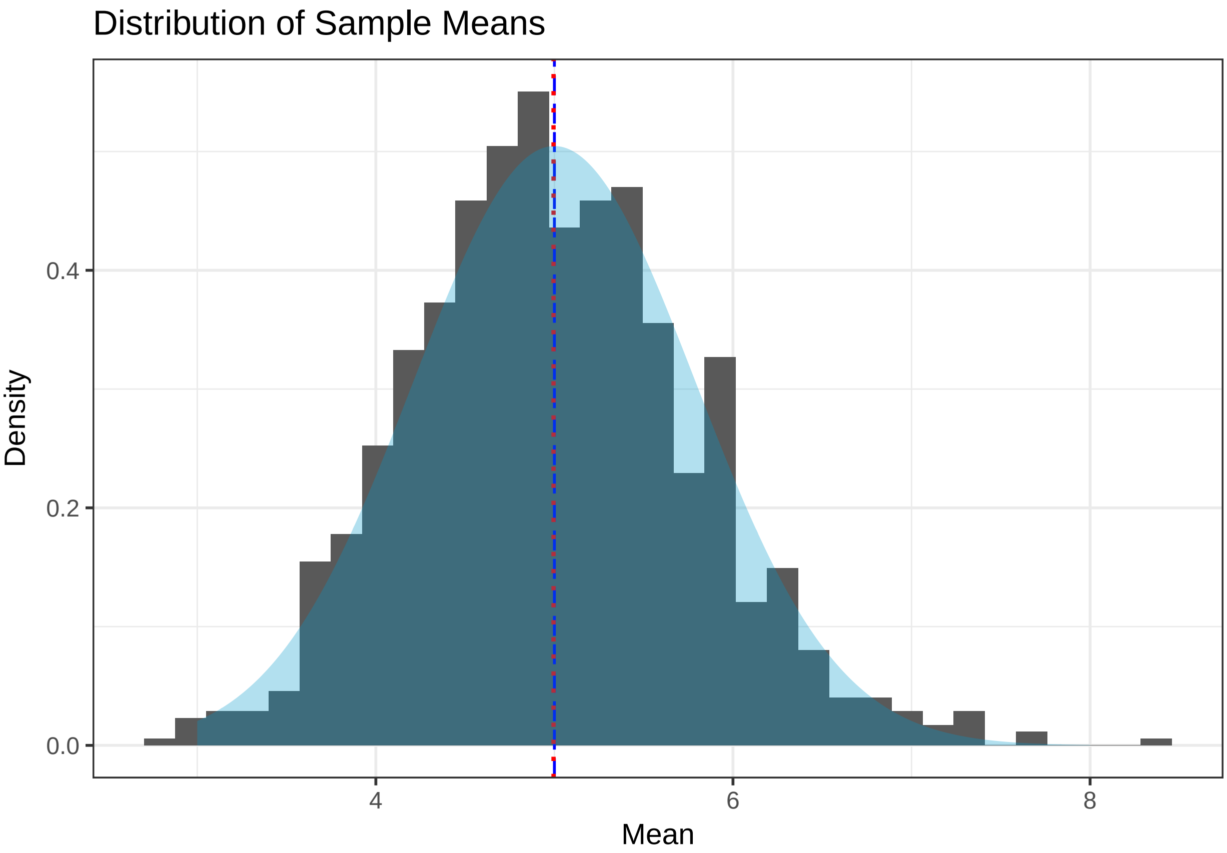
Analysis of Simulations

Calculating the mean and variance of each sample, we'll examine the distributions of the sample means and sample variance.

```
sim_mean <- apply(exp_sim,1,mean)
sim_var <- apply(exp_sim,1,var)
```

Sample Means

If we create a histogram of the 1000 sample means, we find the following distribution:



There are a few interesting characterisics to point out. The **red, short dotted-line** represents the *mean* of the sample means.

```
mean(sim_mean)
```

```
## [1] 4.995158
```

According to the *Central Limit Theorem*, the *mean* of the sample means should approach the theoretical mean of the sample population. In this scenario, the theoretical mean, $\bar{X}$, is equal to $1/\lambda = 5$. The **blue, long dotted-line** represents the theoretical mean. It is easy to see that the two means overlap and are nearly the same.

Another characteristic worth investigating is the *variabililty* of the sample means distribution. Again, according to the *CLT* the *standard deviation* should follow the the variance of the population, σ^2. The *standard deviation* of the sample means is, in fact, $\sigma/\sqrt{n}$. However, a property of the exponential distribution is that the variance, σ^2, is related to λ by $\sigma^2 = 1/\lambda^2 = 25$. Therefore, the *standard error of the means* should approach $\sigma/\sqrt{n} = 1/\lambda * 1/\sqrt{n} = 5/sqrt(40)$:

```
5/sqrt(40)
```

```
## [1] 0.7905694
```

We can calculate the *standard deviation* of our sample means:

```
sd(sim_mean)
```

```
## [1] 0.7930618
```

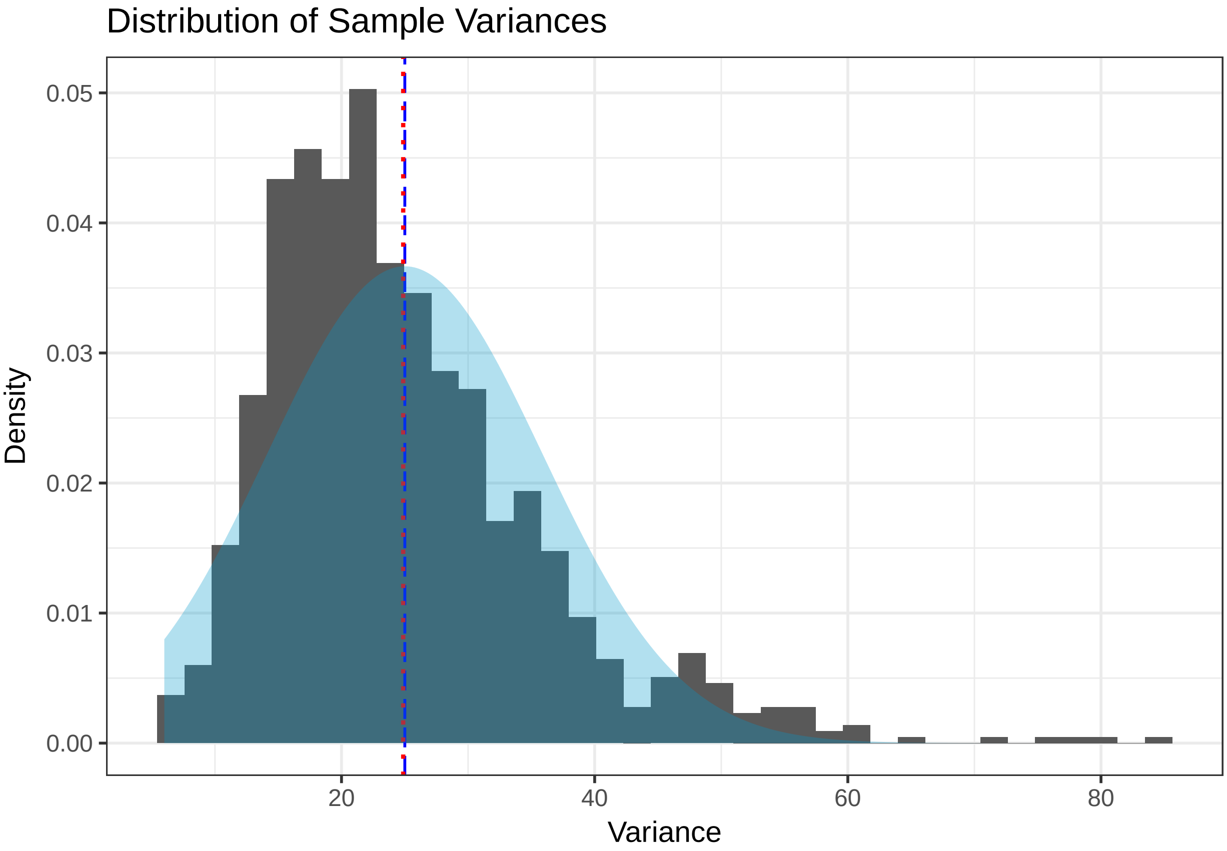
We see that they are quite close.

Finally, the *CLT* states that the distribution of sample means should form a *normal (Gaussian) distribution* centered around the mean of the population ($\mu = 5$) with a standard deviation of $\sigma/\sqrt{n} = 5/\sqrt{40}$: $N(\mu, s) =$

$N(5, 5/\sqrt{40})$. This normal distribution has been ploted with the histogram, with its area shaded a light blue. We can see that this *bell curve* does a very nice job of characterizing the sample means distribution.

Sample Variance

Lastly, we'll take a look at the sample variance. As before, the *CLT* tells us that the distribution of the sample variance should be normal, centered at the theoretical variance of the population - in this case $\sigma^2 = 1/\lambda^2 = 5^2 = 25$.



The average of the sample variance is represented by the **red, short dotted-line**. It has the following value:

```
mean(sim_var)
```

```
## [1] 24.87467
```

The theoretical variance of our exponentially distributed population is $\sigma^2 = 1/\lambda^2 = 25$, and is represented in the distribution by a **blue, long dotted-line**. We can see that they overlap and are nearly the same. The distribution is starting to take normal shape, but is shifted slightly to the left. The distribution of sample variance would become more normal with sufficiently more samples. To roughly fit the histogram with a normal distribution around $\bar{\sigma}^2 = 25$, we manually calculate the standard deviation of the sample variance,

```
sd(sim_var)
```

```
## [1] 10.88064
```

and use that for the normal distribution. It has been added with an area that is shaded light blue.

Part 2

In **Part 2** we will investigate the dataset *ToothGrowth*. This dataset contains data taken while tracking the tooth growth in 60 different guinea pigs while receiving doses of Vitamin C. The vitamin was delivered via two different supplements: Orange Juice (`supp = OJ`) and Ascorbic Acid (`supp = VC`). For each supplement, three different dosages were administered: 0.5, 1, or 2 mg/day. See the following table for a complete layout of the data taken in the study.

Supplement (`supp`)	Dosage, mg/day (`dose`)	Length of Tooth Growth (`len`)
`OJ`	`0.5`	`tooth growth length (#) for 10 guinea pigs`
`OJ`	`1.0`	`tooth growth length (#) for 10 guinea pigs`
`OJ`	`2.0`	`tooth growth length (#) for 10 guinea pigs`
`VC`	`0.5`	`tooth growth length (#) for 10 guinea pigs`
`VC`	`1.0`	`tooth growth length (#) for 10 guinea pigs`
`VC`	`2.0`	`tooth growth length (#) for 10 guinea pigs`

With 10 guinea pigs in each of the six studies, we get our 60 *total* guinea pigs.

We'll load the data set and confirm these unique variables.

```
data("ToothGrowth")
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
unique(ToothGrowth$supp)
```

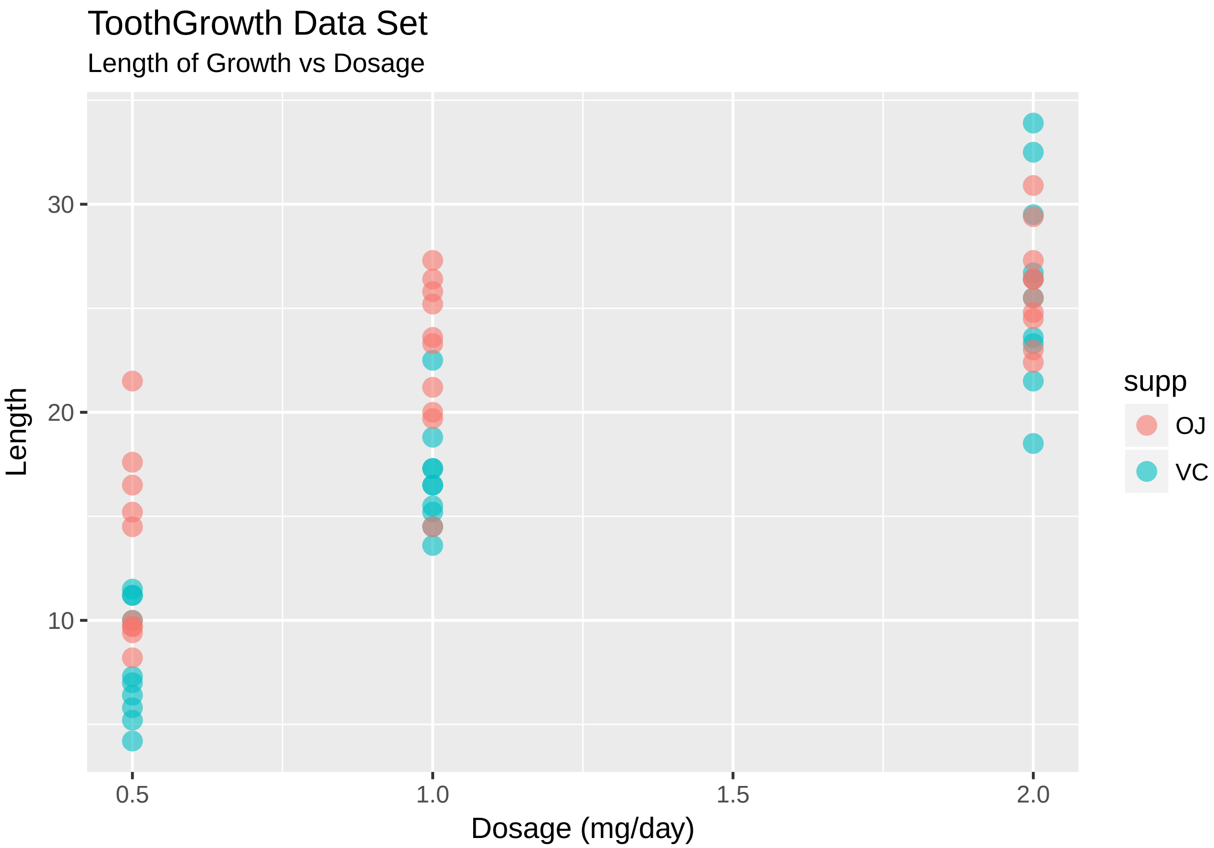
```
## [1] VC OJ
## Levels: OJ VC
```

```
unique(ToothGrowth$dose)
```

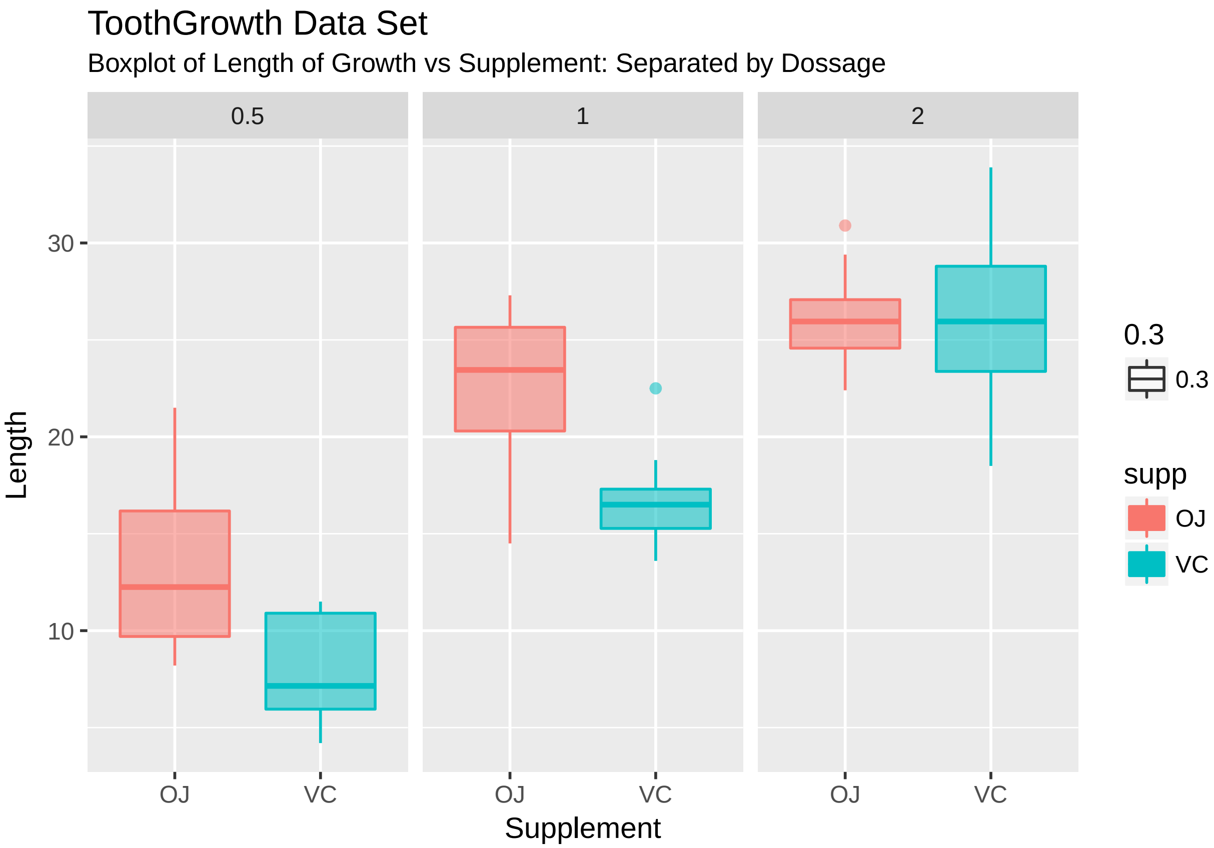
```
## [1] 0.5 1.0 2.0
```

Exploratory Analysis

A quick plot showing *length* versus *dosage*, with the distinction of *supplement*.



A *boxplot* might be more informative, giving us, perhaps, a better idea of the median and variability of each scenario.



To summarize the results, we can find the *mean* and *standard deviation* for each of the six scenarios:

```
summarise(group_by(ToothGrowth,supp,dose),mean(len))
```

```
## Source: local data frame [6 x 3]
## Groups: supp [?]
##
##     supp  dose `mean(len)`
##   <fctr> <dbl>       <dbl>
## 1     OJ   0.5       13.23
## 2     OJ   1.0       22.70
## 3     OJ   2.0       26.06
## 4     VC   0.5        7.98
## 5     VC   1.0       16.77
## 6     VC   2.0       26.14
```

```
summarise(group_by(ToothGrowth,supp,dose),sd(len))
```

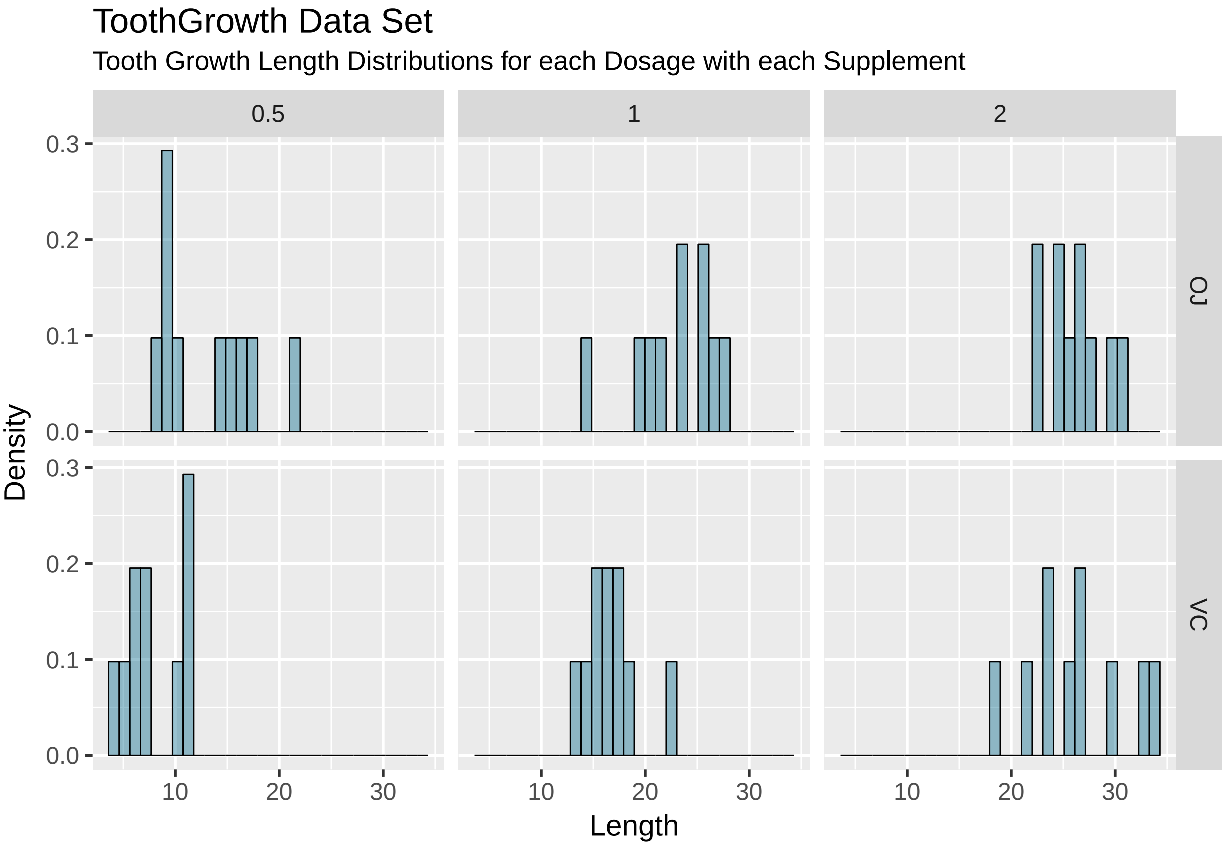
```
## Source: local data frame [6 x 3]
## Groups: supp [?]
##
##     supp  dose `sd(len)`
##   <fctr> <dbl>     <dbl>
## 1     OJ   0.5  4.459709
## 2     OJ   1.0  3.910953
## 3     OJ   2.0  2.655058
## 4     VC   0.5  2.746634
## 5     VC   1.0  2.515309
## 6     VC   2.0  4.797731
```

It looks as if the `0.5` and `1.0` doses of `OJ` provide more growth than `VC`. However, the variability of `VC` is less than that of `OJ` for these two doses. When comparing a dose of `2.0`, `OJ` and `VC` perform comparably but now `VC` has the larger spread.

Let's investigate these claims further.

Testing

Let's use *Hypothesis Testing* and *Confidence Intervals* to reach conclusions about our summary (from above) of the data. This testing will be done using **Student's (Gosset's) t-test**. As a good rule of thumb, when using t-tests with small sample sizes, like we are here, the sample should follow, roughly, a normal distribution. Below is a histogram of tooth growth versus dosage versus supplement:



Although some of the plots look skewed or spread out, each looks "normal" enough for us to proceed with our analysis using **t-tests**.

`Dosage = 0.5 mg/day`

We begin with the (null) hypothesis that the Orange Juice (`OJ`) promotes the same tooth growth in guinea pigs as Ascorbic Acid (`VC`). As an alternative hypothesis, we suggest that `OJ` promotes *more* tooth growth than `VC`. See the following formal statement:

$$H_0: \ \mu_{OJ} = \mu_{VC}$$
$$H_a: \ \mu_{OJ} > \mu_{VC}$$

To build our **independent, two sample t-test statistic**, let's summarize our findings from the `0.5 mg/day` dosage:

Supplement (`supp`) at `0.5 mg/day`	Mean Length ($\bar{X}$)	Sample Standard Deviation (s)
`OJ`	13.23	4.4597085
`VC`	7.98	2.7466343

Given the sample sizes, we choose to treat the variances as *unequal*. This should result in a slightly more conservative result. In doing so, we need to use the following equations to build our t-test statistic:

$$t = \frac{\bar{X}_{OJ} - \bar{X}_{VC}}{s_{\bar{\Delta}}}$$

$$s_{\bar{\Delta}} = \sqrt{\frac{s_{OJ}^2}{n_{OJ}} + \frac{s_{VC}^2}{n_{VC}}}$$

Special attention must also be paid to the *degrees of freedom* (*d.o.f.*) when treating the sample variances as unequal:

$$d.o.f. = \frac{(s_{OJ}^2/n_{OJ} + s_{VC}^2/n_{VC})^2}{(s_{OJ}^2/n_{OJ})^2/(n_{OJ}-1) + (s_{VC}^2/n_{VC})^2/(n_{VC}-1)}$$

We've created two R functions, `s_delta(s1,s2,n1,n2)` and `dof(s1,s2,n1,n2)`, to help with the calculation of the t-test statistic:

```r
s_delta <- function(s1,s2,n1,n2) sqrt(s1^2/n1 + s2^2/n2)
dof <- function(s1,s2,n1,n2) {
      (s1^2/n1+s2^2/n2)^2 / ( (s1^2/n1)^2 / (n1-1) + (s2^2/n2)^2 / (n2-1) )
}
```

So, to complete our testing, we need to complete the following:

```r
diff_in_means <- 13.23 - 7.98
t <- diff_in_means/s_delta(4.4597085,2.7466343,10,10)
the_dof <- dof(4.4597085,2.7466343,10,10)
conf_95 <- 13.23 - 7.98 + c(-1,1) * qt(.95,the_dof) * s_delta(4.4597085,2.7466343,10,10)
```

We can finally conclude our analysis.

> After an inspection of the `OJ` and `VC` sample distributions, we determined that they were sufficently normal to proceed with the **Student's t-test**. Given the sample size and our lack of knowledge of the sample standard deviations, we decided to continue with *unequal* variances. An independent t-test was run on data with a 95% confidence interval (CI) for the mean difference ($H_0: \bar{X}_{OJ} = \bar{X}_{VC}$ or $H_0: \bar{X}_{OJ} - \bar{X}_{VC} = 0$). It was found that Orange Juice promoted more tooth growth ($\bar{X}_{OJ} = 13.23 \pm 4.46$) than Ascorbic Acid ($\bar{X}_{VC} = 7.98 \pm 2.75$) with a difference in means of 5.25 (`diff_in_means`) and 95% CI of (2.3460404, 8.1539596) (`conf_95`). The t-test statistic is `t(the_dof) = t` which is `t(`14.9687537`) =` 3.1697328. The corresponding p-value is `pt(t,the_dof,lower.tail=FALSE) =` 0.0031793 and is statistically significant. Therefore, we can reject our null hypothesis and favor the alternative hypothesis, $H_a: \bar{X}_{OJ} > \bar{X}_{VC}$.

We can check this 'by-hand' calculation against the built-in `R` function `t.test`.

```
t.test(
      filter(ToothGrowth,supp=="OJ",dose==0.5)$len,
      filter(ToothGrowth,supp=="VC",dose==0.5)$len,
      alternative="greater",paired=FALSE,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(ToothGrowth, supp == "OJ", dose == 0.5)$len and filter(ToothGrowth, supp == "VC", dose
## t = 3.1697, df = 14.969, p-value = 0.003179
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.34604     Inf
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

We see that our analysis matches the analysis by the `R` function `t.test`. As a novice statistician, this should give us confidence in our analysis.

Dosage = 1.0 mg/day

We begin with the (null) hypothesis that the Orange Juice (`OJ`) promotes the same tooth growth in guinea pigs as Ascorbic Acid (`VC`). As an alternative hypothesis, we suggest that `OJ` promotes *more* tooth growth than `VC`. See the following formal statement:

$$H_0: \ \mu_{OJ} = \mu_{VC}$$
$$H_a: \ \mu_{OJ} > \mu_{VC}$$

As before, let's summarize our findings from the `1.0 mg/day` dosage:

Supplement (`supp`) at `1.0 mg/day`	Mean Length ($\bar{X}$)	Sample Standard Deviation (s)
`OJ`	22.7	3.9109533
`VC`	16.77	2.5153087

With understanding of the underlying procedure and trust in the `R` function `t.test`, we can confidently perform our analysis with `t.test`:

```
t.test(
      filter(ToothGrowth,supp=="OJ",dose==1.0)$len,
      filter(ToothGrowth,supp=="VC",dose==1.0)$len,
      alternative="greater",paired=FALSE,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(ToothGrowth, supp == "OJ", dose == 1)$len and filter(ToothGrowth, supp == "VC", dose =
## t = 4.0328, df = 15.358, p-value = 0.0005192
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.356158      Inf
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

From this, we can conclude our analysis at 1.0 mg/day.

> After an inspection of the OJ and VC sample distributions, we determined that they were sufficently normal to proceed with the **Student's t-test**. Given the sample size and our lack of knowledge of the sample standard deviations, we decided to continue with *unequal* variances. An independent t-test was run on data with a 95% confidence interval (CI) for the mean difference. It was found that Orange Juice promoted more tooth growth ($\bar{X}_{OJ} = 22.7 \pm 3.9$) than Ascorbic Acid ($\bar{X}_{VC} = 16.77 \pm 2.52$) with a difference in means of 5.93 and 95% CI of (3.356161, 8.503839). The t-test statistic is 4.0328. The corresponding p-value is 0.0005192 and is statistically significant. Therefore, we can reject our null hypothesis and favor the alternative hypothesis, $H_a:\ \bar{X}_{OJ} > \bar{X}_{VC}$.

Dosage = 2.0 mg/day

We begin with the (null) hypothesis that the Orange Juice (OJ) promotes the same tooth growth in guinea pigs as Ascorbic Acid (VC). As an alternative hypothesis, we suggest that OJ promotes *more* tooth growth than VC. See the following formal statement:

$$H_0:\ \mu_{OJ} = \mu_{VC}$$
$$H_a:\ \mu_{OJ} > \mu_{VC}$$

As before, let's summarize our findings from the 2.0 mg/day dosage:

Supplement (supp) at 2.0 mg/day	Mean Length ($\bar{X}$)	Sample Standard Deviation (s)
OJ	26.06	2.6550581
VC	26.14	4.7977309

With understanding of the underlying procedure and trust in the R function t.test, we can confidently perform our analysis with t.test:

```
t.test(
      filter(ToothGrowth,supp=="OJ",dose==2.0)$len,
      filter(ToothGrowth,supp=="VC",dose==2.0)$len,
      alternative="greater",paired=FALSE,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(ToothGrowth, supp == "OJ", dose == 2)$len and filter(ToothGrowth, supp == "VC", dose =
## t = -0.046136, df = 14.04, p-value = 0.5181
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.1335     Inf
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

From this, we can conclude our analysis at 2.0 mg/day.

> After an inspection of the OJ and VC sample distributions, we determined that they were sufficently normal to proceed with the **Student's t-test**. Given the sample size and our lack of knowledge of the sample standard deviations, we decided to continue with *unequal* variances. An independent t-test was run on data with a 95% confidence interval (CI) for the mean difference. It was found

that Orange Juice promoted the same tooth growth ($\bar{X}_{OJ} = 26.06 \pm 2.66$) as Ascorbic Acid ($\bar{X}_{VC} = 26.14 \pm 4.80$) with a difference in means of -0.08 and 95% CI of (`-3.133497, 2.973497`). The t-test statistic is `-0.046136`. The corresponding p-value is `0.5181`. Therefore, we can accept our null hypothesis, $H_0 : \bar{X}_{OJ} = \bar{X}_{VC}$.

Conclusion

In this assignment we examined the *exponential distribution* and some of its characteristics. By simulating events, we constructing distributions of sample means and sample variance. We observed that these distributions are normal and subscribe nicely to the *Central Limit Theorem*.

We also reevaluated data that was taken during a previous experiment to investigate the effectiveness of Vitamin C to promote tooth growth in guinea pig. Vitamin C was administered via Orange Juice or Ascorbic Acid in three different doses. We found that Orange Juice given in doses of 0.5 and 1.0 mg/day promoted more tooth growth than Ascorbic Acid that was statistically significant. A dosage of 2.0 mg/day resulted in no discernable difference in growth promotion between Orange Juice and Ascorbic Acid.

* C. I. Bliss (1952) *The Statistics of Bioassay.* Academic Press. ** https://en.wikipedia.org/wiki/Exponential_distribution