Statistical Inference Project (Coursera)

Jason Murray

2017-01-17

Overview

In this project we will be exploring the exponential distribution using R. Looking at how the theoretical mean and variance compare to a sample mean and variance. We will also be looking at how the values of the exponential function are distributed compared to how the values of the means of samples of exponential function are distributed as a demonstration of the central limit theorem in action.

Simulations

Required Libraries

```
library(ggplot2)
library(dplyr)
```

First let's set a random seed and generate a 1000 means of sample size 40 from the exponential distribution. We will be using $\lambda = .2$ throughout.

```
set.seed(42)

# Generate a 1000 means of random values form the exponential distribution
re1000m = NULL
for (i in 1 : 1000) re1000m = c(re1000m, mean(rexp(n= 40, rate = .2)))
```

Sample Mean vs Theoretical Mean

The theoretical mean of the exponetial distribution is equal to $1/\lambda$. So with $\lambda = .2$ the theoretical mean would be equal to

$1/\lambda = 1/.2 = 5$

If we now calculate the mean for our 1000 means we get.

```
mre1000m <- mean(re1000m)
mre1000m
```

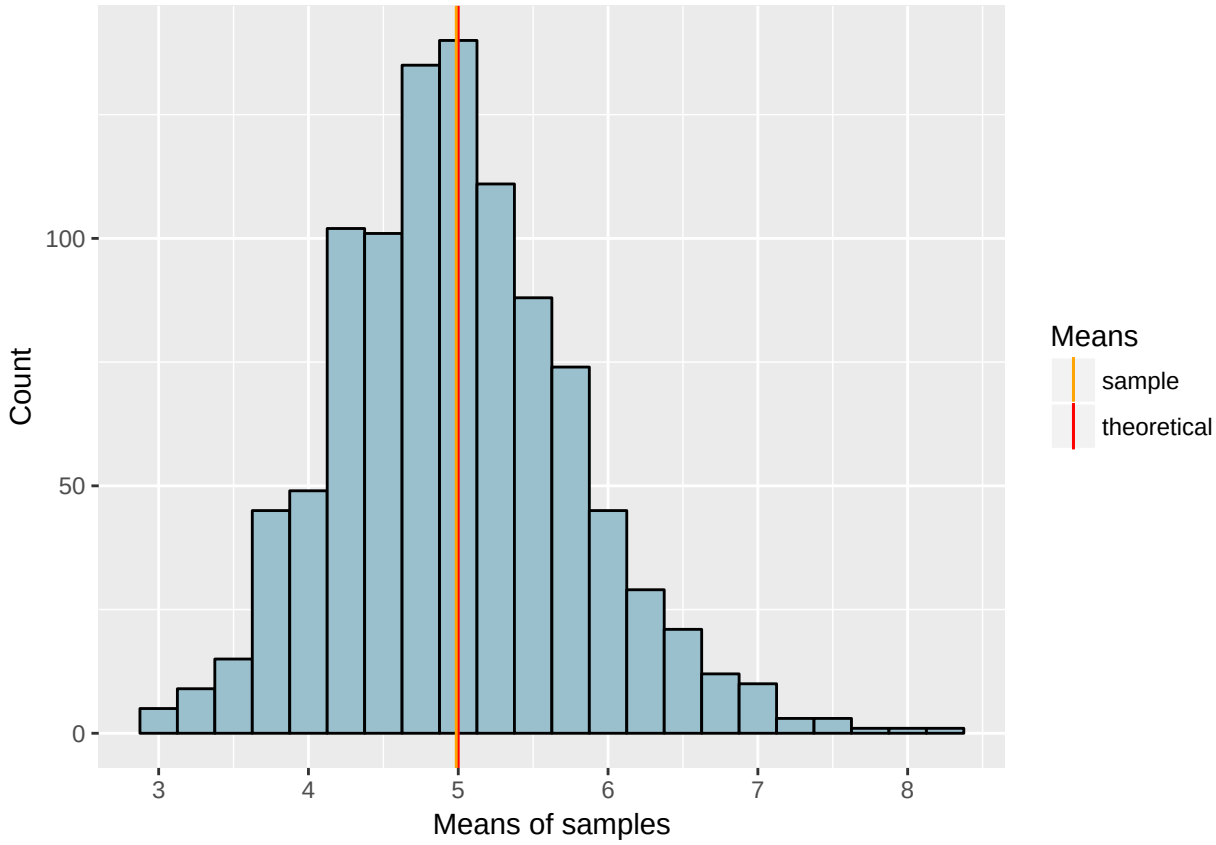
```
## [1] 4.986508
```

Which is very close to the theoretical value of 5.

Let's graph the distribution of our sample means and so we can see how close.

```
ggplot() +
    aes(re1000m) +
    geom_histogram(binwidth = .25, fill = "light blue 3", color = "black") +
    geom_vline(aes(xintercept = 5, color = "theoretical")) +
```

```
    geom_vline(aes(xintercept = mre1000m, color = "sample")) +
    scale_color_manual(name = "Means", values = c(theoretical = "red", sample = "orange")) +
    labs(x = "Means of samples", y = "Count")
```



Sample Variance vs Theoretical Variance

The theoretical variance for the distribution of samples means is σ^2/N where sigma is equal to $1/\lambda$. We already know $1/\lambda = 5$ so plugging in we get: $5^2/N = 25/40 = .625$

```
var(re1000m)
```

```
## [1] 0.6344405
```

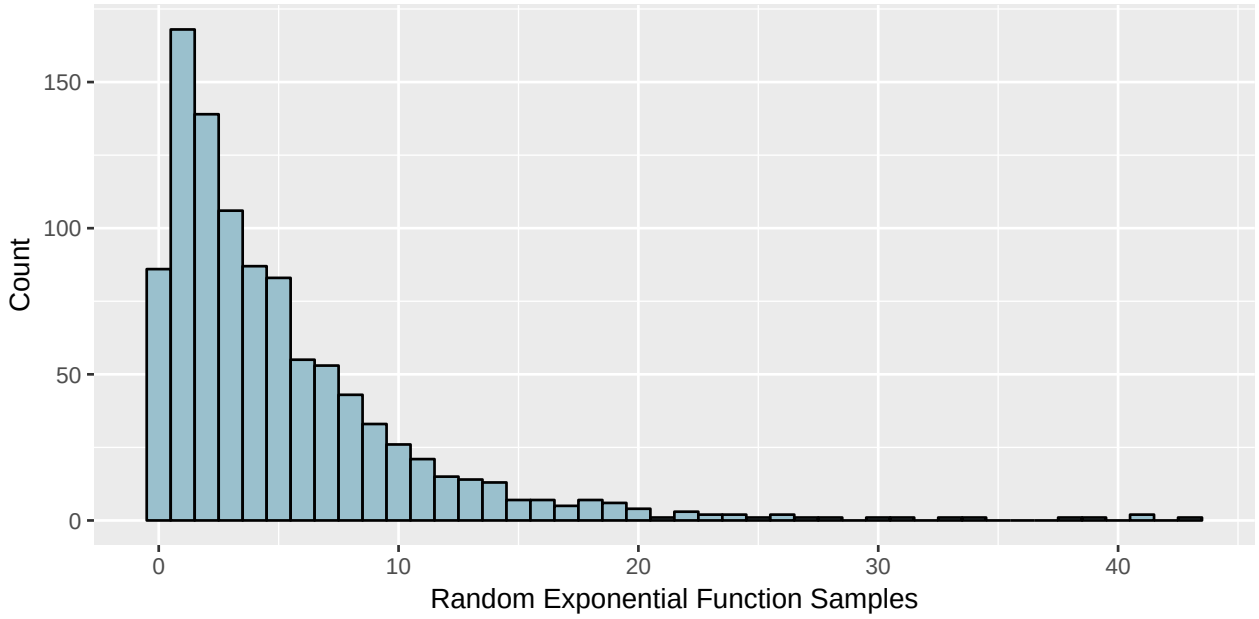
Which is again very close to the theoretical variance of .625.

Distribution

We have so far been looking at the distribution of sample means but what does the distribution of the sample exponential function variables actually look like?

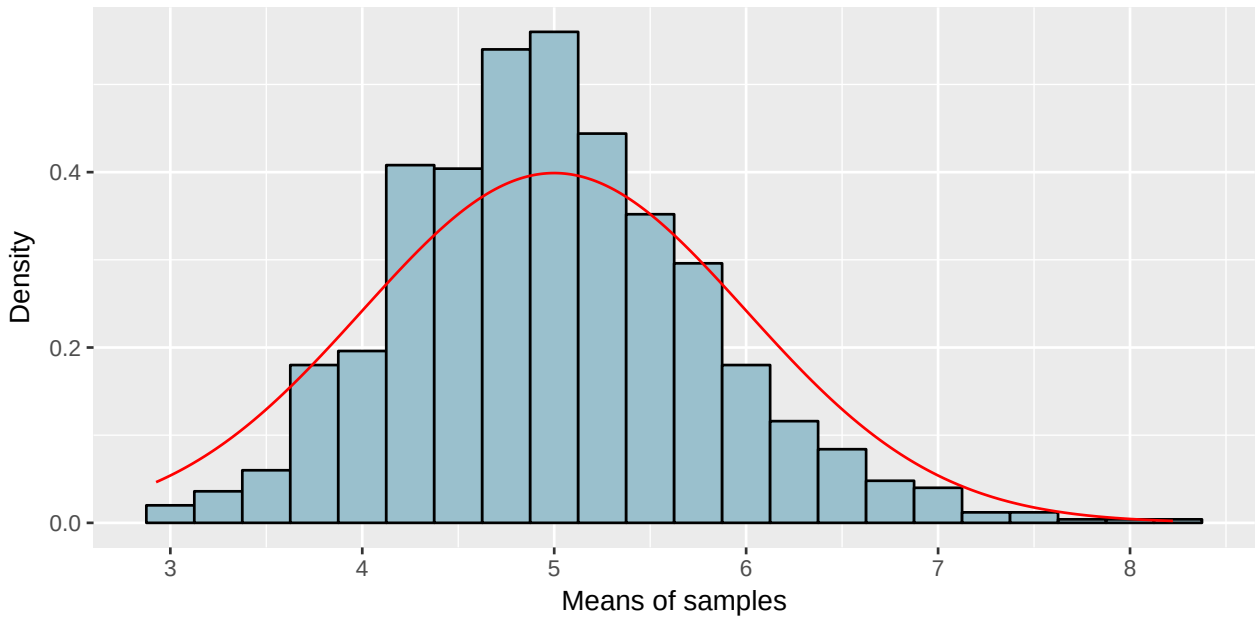
Let's sample a 1000 random variables from the exponential function and see how they are distributed

```
re1000 <- rexp(1000, rate = .2)
ggplot() +
    aes(re1000) +
    geom_histogram(binwidth = 1, fill = "light blue 3", color = "black") +
    labs(x = "Random Exponential Function Samples", y = "Count")
```



So why does the distribution of 1000 means of 40 samples look different than 1000 sample values?

```
ggplot() +
    aes(re1000m) +
    geom_histogram(aes(y=..density..),binwidth = .25, fill = "light blue 3", color = "black") +
    stat_function(fun = dnorm, color = "red", n = 1000, args = list(mean = 5)) +
    labs(x = "Means of samples", y = "Density")
```



This is because of the central limit theorem. Regardless of how the original data is distributed the distribution of the means of samples approaches a normal distribution centered at the mean of the population as the sample size get's larger. I changed the y axis to density above and overlayed a normal curve to show that the shape of the distribution of means is approaching normal.

Part 2 - Tooth Growth Data Analysis

Initial Exploration

Let's load up the Tooth Growth dataset and do some exploration

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
summary(ToothGrowth)
```

```
##       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
```

So we have 60 observations across 3 variables. Looks like it's broken into 2 groups OJ and VC.

Looking up the help page it looks like this was from a study done with 60 guinea pigs so they were all unique subjects.

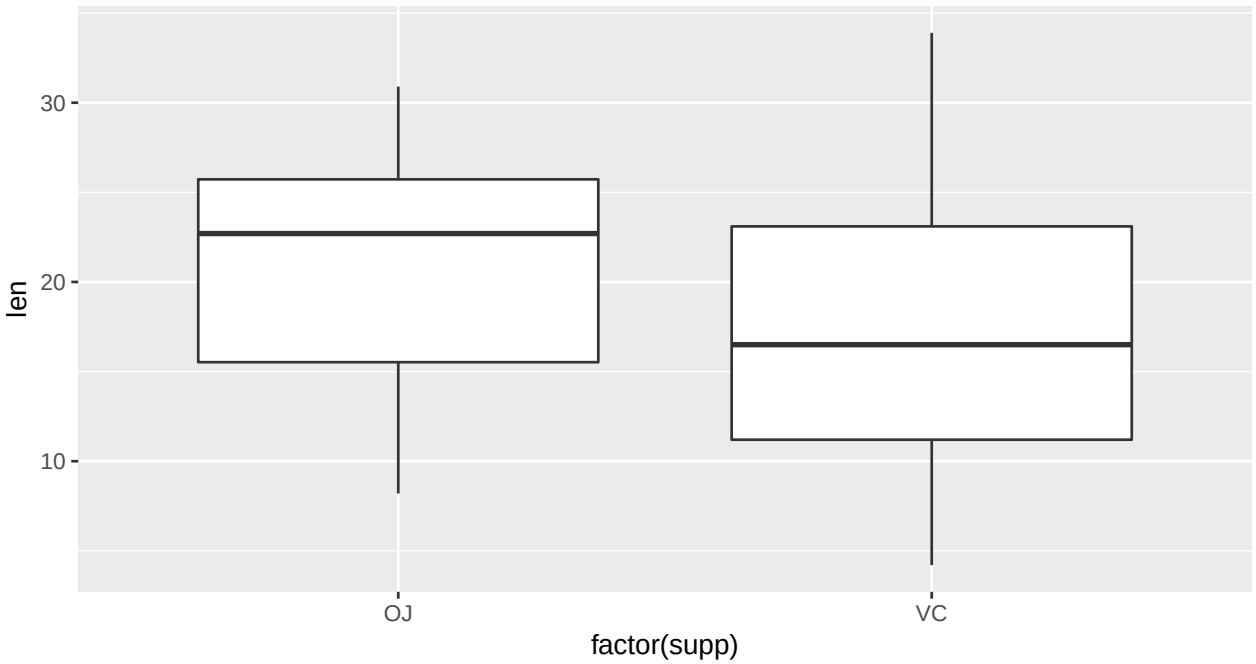
The first column (len) is defined as numeric Tooth length. I'm assuming this is a differential from the original measured length at the conclusion of the study. It's not really clear from the description of the data but for our analysis I don't think it relevant. It also doesn't include the units of the measurement.

The second column(supp) is the type of supplment used, so they gave half of the subjects vitamin C as absorbic acid (VC) and half in the form of orange juice (OJ).

The final column(dose) is the dosage in milligrams/day. This is the amount of vitamin C regarldess of of the supplement used. They were 3 groups per supplement consisting of .5, 1, 2 mg/day dosages.

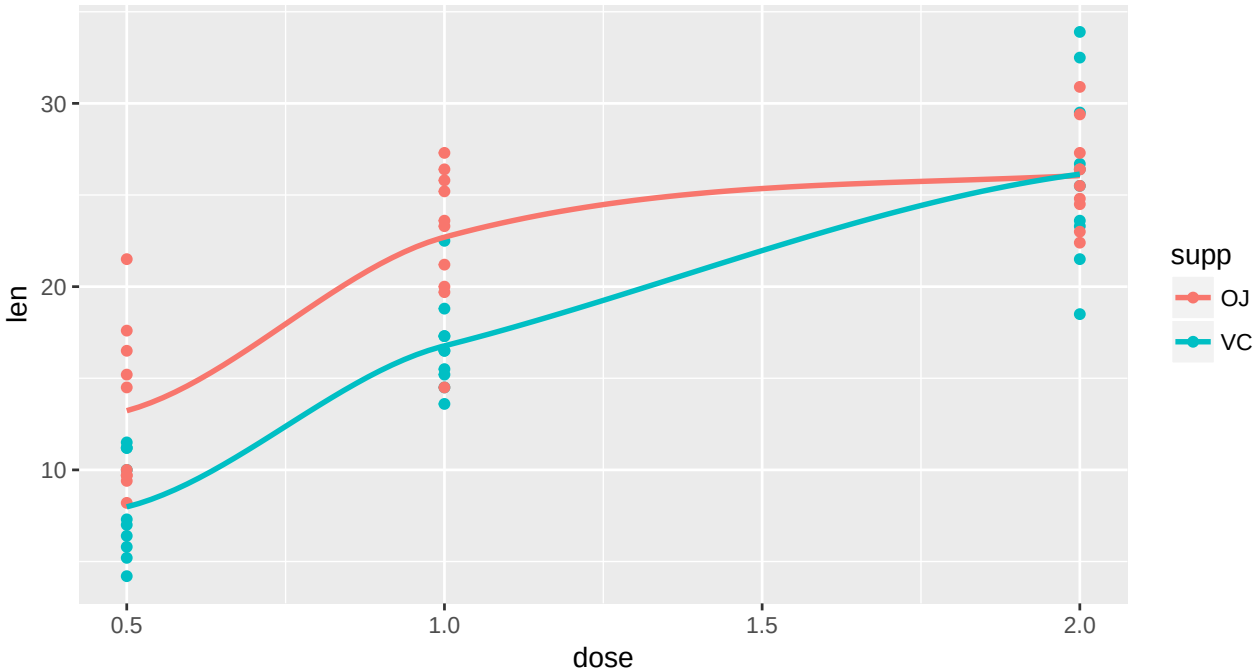
Let's do a boxplot to compare the to delivery methods to see if it look there is an obvious difference in the application method.

```
ggplot(data = ToothGrowth, aes(factor(supp),len)) +
    geom_boxplot()
```



So intially it looks like OJ might have a larger effect on toothgrowth. Let's look at the data by dosage and supplement type and do some smoothing

```
ggplot(data = ToothGrowth, aes(dose, len, color = supp)) +
    geom_point() +
    geom_smooth(se = FALSE)
```



Looking at this it's less clear that one is more effective than the other. At lower doses it does appear that OJ results in more tooth growth. However at the full dosage of 2mg/day they appear to be nearly equal.

Hypothesis testing

Now that we've done some inital exploration let's get down to the hypothesis testing. I will examine 2 questions here. Is the type of supplement a factor in tooth growth? Does increasing dosage increase growth?

Supplement type effect

First let's split the data in to 2 sets based on the supplement type.

```
TG_OJ <- filter(ToothGrowth, supp == 'OJ')
TG_VC <- filter(ToothGrowth, supp == 'VC')
```

Let's use a 2 sided t-test to see if there is a difference between the supplements across all dose levels. Our null hypothese will be

$H_0 : \mu_v = \mu_o$

that the mean of length will be the same for both supplements.

The alternative hypothesis in this case

$H_A : \mu_v != \mu_o$

that the mean of the length is different for the supplements.

We will use $\alpha = .05$ for our testing.

```
 t.test(TG_OJ$len, TG_VC$len)
```

```
##
##  Welch Two Sample t-test
##
## data:  TG_OJ$len and TG_VC$len
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

So the p-value here is .061, based on this we can't reject our null hypothesis. Meaning that the there is not enough a difference between the 2 supplements to say one is more effective at producing tooth growth.

Dosage Effect

Let's start by breaking our original data into 3 pieces based on dosage so we can compare between them.

```
TG_.5 <- filter(ToothGrowth, dose == 0.5)
TG_1 <- filter(ToothGrowth, dose == 1.0)
TG_2 <- filter(ToothGrowth, dose == 2.0)
```

In this case we are looking to see if by increasing the dose we also increase tooth growth. We will use a one sided t-test between each dosage level to see if there is a significant increase.

Our null hypothesis in this case is that the mean of the length will not change with dose.

First let's compare .5 to 1.0

```
t.test(TG_1$len, TG_.5$len, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  TG_1$len and TG_.5$len
## t = 6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.753323      Inf
## sample estimates:
## mean of x mean of y
##    19.735    10.605
```

So based on a pvalue of 6.342e-08 we can definately reject the null hypothesis here.

Let's take a look at 1.0 vs 2.0 now

```
t.test(TG_2$len, TG_1$len, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  TG_2$len and TG_1$len
## t = 4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.17387     Inf
## sample estimates:
## mean of x mean of y
##    26.100    19.735
```

This difference in dose has a p-value of 9.532e-06. So we can reject our null hypothese based on this change in dosage as well.

Conclusions

From our hypothesis testing we can conclude that although there does appear to be some difference between supplement type, it is not enough of a difference to be considered significant. When it comes to dosage there is a much clearer correlation and we can safely say that increasing dosage does increase tooth growth.

Appendix

Full ToothGrowth dataset

```
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
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
## 11 16.5   VC  1.0
## 12 16.5   VC  1.0
## 13 15.2   VC  1.0
## 14 17.3   VC  1.0
## 15 22.5   VC  1.0
## 16 17.3   VC  1.0
## 17 13.6   VC  1.0
## 18 14.5   VC  1.0
## 19 18.8   VC  1.0
## 20 15.5   VC  1.0
## 21 23.6   VC  2.0
## 22 18.5   VC  2.0
## 23 33.9   VC  2.0
## 24 25.5   VC  2.0
## 25 26.4   VC  2.0
## 26 32.5   VC  2.0
## 27 26.7   VC  2.0
## 28 21.5   VC  2.0
## 29 23.3   VC  2.0
## 30 29.5   VC  2.0
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
## 34  9.7   OJ  0.5
## 35 14.5   OJ  0.5
## 36 10.0   OJ  0.5
## 37  8.2   OJ  0.5
## 38  9.4   OJ  0.5
## 39 16.5   OJ  0.5
## 40  9.7   OJ  0.5
## 41 19.7   OJ  1.0
## 42 23.3   OJ  1.0
## 43 23.6   OJ  1.0
## 44 26.4   OJ  1.0
## 45 20.0   OJ  1.0
## 46 25.2   OJ  1.0
## 47 25.8   OJ  1.0
```

```
## 48 21.2   OJ  1.0
## 49 14.5   OJ  1.0
## 50 27.3   OJ  1.0
## 51 25.5   OJ  2.0
## 52 26.4   OJ  2.0
## 53 22.4   OJ  2.0
## 54 24.5   OJ  2.0
## 55 24.8   OJ  2.0
## 56 30.9   OJ  2.0
## 57 26.4   OJ  2.0
## 58 27.3   OJ  2.0
## 59 29.4   OJ  2.0
## 60 23.0   OJ  2.0
```