Inference Course Project

Joe Nguyen

31 October 2015

In this project, we use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

- A simulation exercise
- Basic inferential data analysis

A. Simulation for Distribution of Means of Exponentials

We investigate the exponential distribution and the Central Limit Theorem (CLT). The mean of the exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. We set $\lambda = 0.2$ for all simulations. We investigate the distribution of averages of 40 exponentials using 1000 simulations.

1. Sample Mean

We show the sample mean (from simulations) and compare it with the theoretical mean of the distribution.

```
## Sample mean: 4.971972 vs. Theoretical mean: 5
```

In Fig. 1 (in Appendix), the sample mean (red dashed line) is close to the theoretical mean (blue solid line).

2. Sample Variance

We compare the sample variance to the theoretical variance ($1/\lambda^2$) by reporting the mean sample variance and theoretical variance, and also plotting the sample variance distribution from the simulations in Fig. 2.

```
## Sample variance: 25.39236 vs. Theoretical variance: 25
```

3. Distribution of Mean of Multiple Draws from Exponential is Approximately Normal

Here, we compare the distribution of 1000 draws of the exponential distribution to the distribution of 1000 "averages of 40 draws from the exponential distribution".

Figure 3 highlights the CLT which states that in the limit as more samples are drawn, the average of these draws follows a Gaussian distribution. The consequence of the CLT is evident in Fig. 3 (right) where the histogram of averages of draws fits in the Gaussian density plot (red line). A beneficial result of the CLT is that in the limit, we can use the sample mean to approximate the mean of the underlying distribution; for the exponential, this is $\frac{1}{\lambda} = \frac{1}{0.2} = 5$ as seen in Fig. 3 (right).

B. Inference on *ToothGrowth* Data

Information about the dataset:

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

1. Load the *ToothGrowth* data

Explore the dataset.

```
data("ToothGrowth")
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 ...
```

```
rbind(head(ToothGrowth,2), tail(ToothGrowth,2))
```

```
##     len supp dose
## 1   4.2   VC  0.5
## 2  11.5   VC  0.5
## 59 29.4   OJ  2.0
## 60 23.0   OJ  2.0
```

2. Provide a Basic Summary of the Data

```
##       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
```

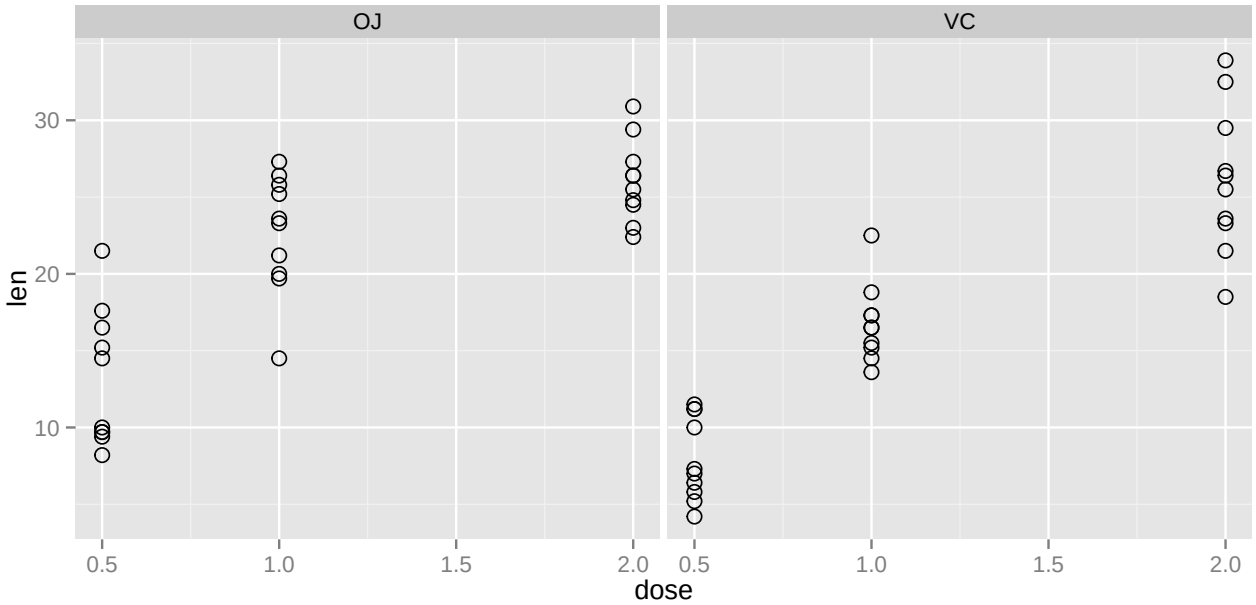

Figure 3. Response (len) based on dose (0.5, 1, 2)mg and supplement type (OJ, VC) from 10 guinea pigs.

3. Compare Tooth Growth

Tooth growth (len) is compared by dose and supplement type (supp). T-tests are used to determine the difference between pairs of dose and the supp pair. We use t-tests because the response (len) appears to by roughly symmetric and mound shaped in Fig. 3.

i. Dose First, differences in len are evaluate for pairs of dose levels = {0.5, 1, 2} mg. The data are paired by dose and so three paired t-test are performed:

t-test Index	Dose A	Dose B
1	0.5	1
2	0.5	2
3	1	2

Also, group variances are assumed to be unequal as they are unknown. Assuming unequal variance results in wider confidence intervals than equal variance, which is more conservative. The 95% confidence intervals (CIs) are:

```
##            [,1]       [,2]
## [1,] -11.872879  -6.387121
## [2,] -18.367198 -12.622802
## [3,]  -9.258186  -3.471814
```

As the 95% CIs are entired below zero, these t-tests suggest the smaller doses result in less tooth growth; or, increasing the dose increases tooth growth. Additionally, the p-values are small ($<< 0.05$), strongly suggesting more tooth growth with higher doses.

```
##              [,1]
## [1,] 1.225437e-06
## [2,] 7.190255e-10
## [3,] 1.934186e-04
```

ii. Supplement Second, a paired t-test is performed to evaluate a difference in tooth growth (len) between the two supplments (supp = {OJ, VC}). Again, unequal variance is assumed to be conservative.

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

```
##
##  Paired t-test
##
## data:  len by supp
## t = 3.3026, df = 29, p-value = 0.00255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.408659 5.991341
## sample estimates:
## mean of the differences
##                     3.7
```

The 95% CI = (1.409, 5.991) suggests the supplment OJ results in more tooth growth than supplment VC. This is corroborated by a small p-value = 0.00255.

Appendix

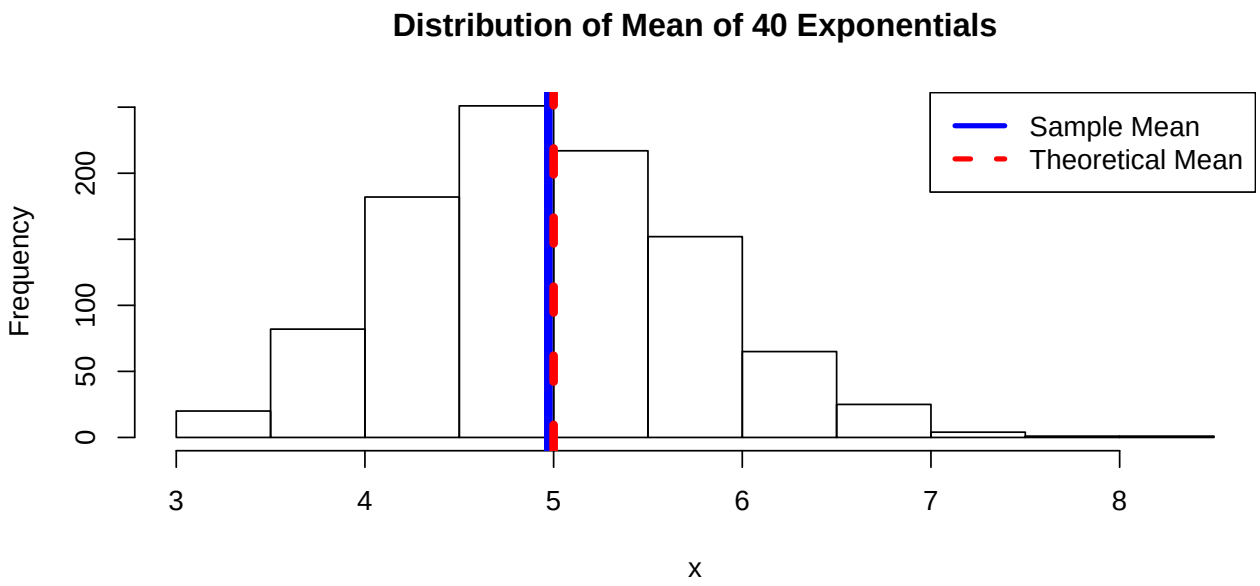

Figure 1. Distribution of mean of 40 exponentials using 1000 simulations.

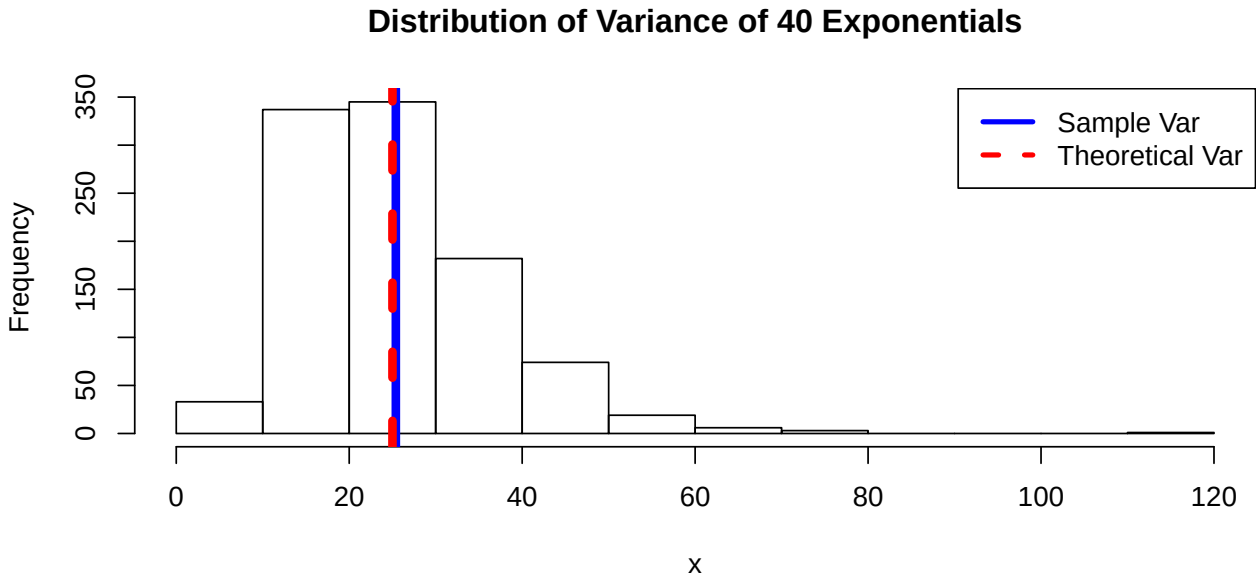

Figure 2. Distribution of variance of 40 exponentials using 1000 simulations.

```
datSample <- data.frame(x = c(rexp(nSim), datExpMn),
                        group = factor(rep(c(1,2), each = nSim),
                        labels = c("Single Draw", "Average of 40 Draws"))
                        )
library(ggplot2)
g <- ggplot(datSample, aes(x)) +
    geom_histogram() +
    facet_wrap(~ group) +
    ggtitle("Exponential Draws Distribution")

# Overlay Normal distribution
```

```r
x <- mnSample + sd(datExpMn) * seq(-3, 3, length = 1000)
y <- dnorm(x, mnSample, sd(datExpMn))

# Scale density to count of datExpMn
y <- y * max(ggplot_build(g)$data[[1]]$count) / max(y)
datNorm <- data.frame(x = c(x, x),
                      y = c(rep(NaN, length(y)), y),
                      group = factor(rep(c(1,2), each = nSim),
                      labels = c("Single Draw", "Average of 40 Draws"))
                      )
# Plot
g + geom_line(aes(x,y), data = datNorm, col = "red") + facet_wrap(~ group)
```

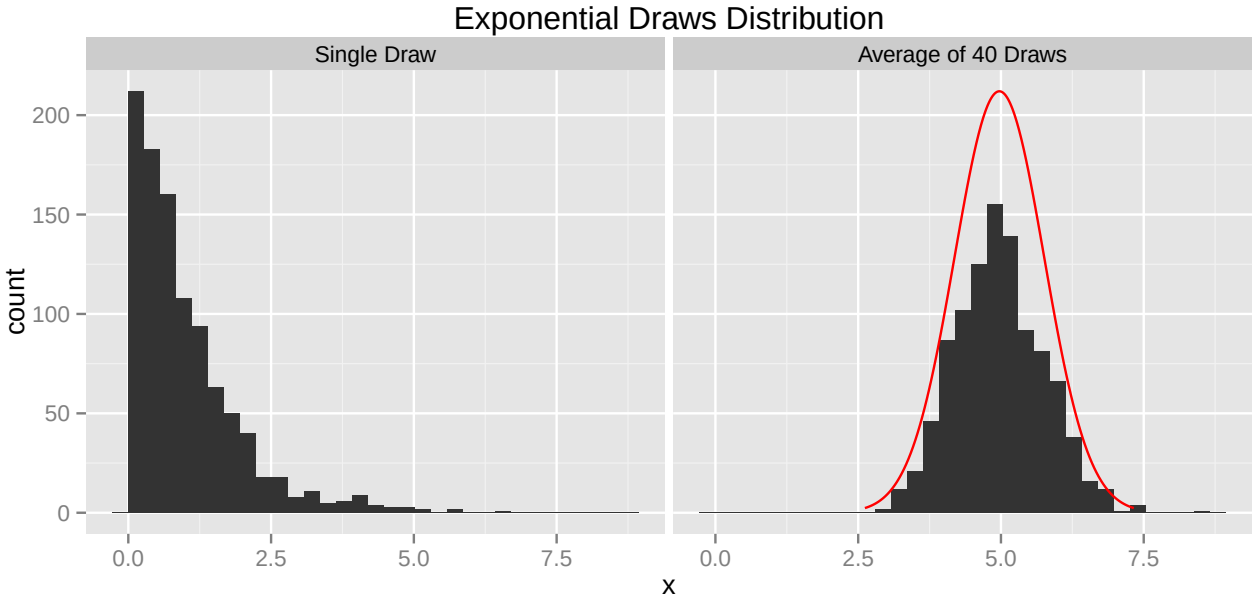

Figure 3. Comparison between sampling distributions with single draws and average of 40 draws from an exponential distribution.

Code for summary of *ToothGrowth* data

```r
summary(ToothGrowth)

tg <- ggplot(ToothGrowth, aes(x = dose, y = len), col = supp) +
    geom_point(shape = 1, size = 3) +
    facet_wrap(~ supp)
tg
```

Code for Dose CIs and P-values

```r
lenDose1 <- subset(ToothGrowth, dose %in% c(0.5,1))
lenDose2 <- subset(ToothGrowth, dose %in% c(0.5,2))
lenDose3 <- subset(ToothGrowth, dose %in% c(1,2))
```

```r
rbind(
    t.test(len ~ dose, paired = TRUE, var.equal = FALSE, data = lenDose1)$conf,
    t.test(len ~ dose, paired = TRUE, var.equal = FALSE, data = lenDose2)$conf,
    t.test(len ~ dose, paired = TRUE, var.equal = FALSE, data = lenDose3)$conf
)

rbind(
    t.test(len ~ dose, paired = TRUE, var.equal = FALSE, data = lenDose1)$p.value,
    t.test(len ~ dose, paired = TRUE, var.equal = FALSE, data = lenDose2)$p.value,
    t.test(len ~ dose, paired = TRUE, var.equal = FALSE, data = lenDose3)$p.value
)
```