Comparing Distributions of Random Exponentials with Samples of Random Exponentials and an Analysis of the Toothgrowth Data Set

Max R

10/15/2017

Overview

These two analyses are part of the final project for Coursera's Statistical Inference Course as part of the Data Science Specialization offered by Johns Hopkins University. The first part of the project will explore the difference in distributions of a random exponential variable versus the distribution of random samples of exponential variables. For the first simulation, n = 1000 and $\lambda = .2$. For the second simulation, there will be 1000 samples of size 40 with $\lambda = .2$.

Simulation 1

The first simulation simulates 1000 random exponentials with a lambda = .2. The mean of an exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$, which means the variance is $1/\lambda^2$. For $\lambda = .2$, the expected (theoretical) mean and variance would equal 5 and 25, respectively. I'm setting the simulated and theoretical values to variables to be displayed in a matrix.

Table 1: Theoretical vs Simulated Mean & Variance for Simulation 1

	Mean	Variance
Simulated	4.86	25.6
Theoretical	5.00	25.0

As we can see in the above matrix, the sample values are quite close to the theoretical values. However, given the nature of the exponential distribution, if we plot a histogram of this simulation, it will have a strong right skew. The vertical black line indicates the position of the simulated mean, while the theoretical is 5 (see page 2).

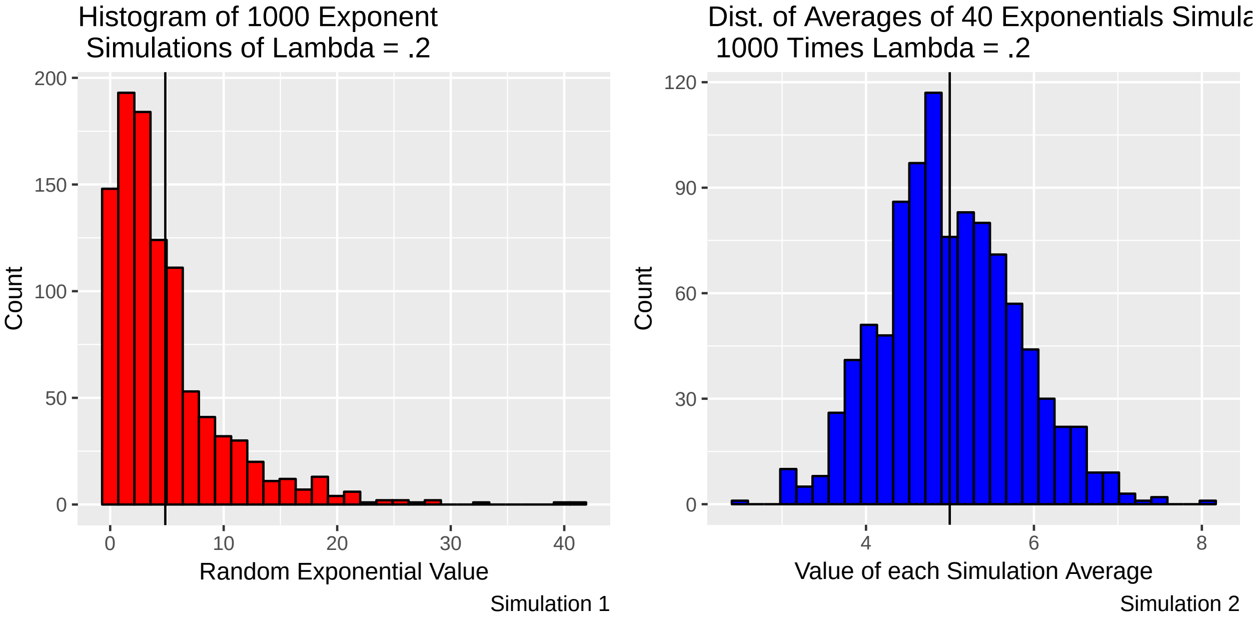
Simulation 2

For the second simulation, we are simulating samples of 40 random exponentials with lambda = .2 one thousand times. Since we are finding the distribution of samples, we would expect the distribution to look much more Gaussian than the first distribution. We would expect the mean to be equal to $1/\lambda$ (just as before) but the expected variance will be different. The standard deviation of a sample mean of exponentials (called the standard error) is equal to $(1/\lambda)/\sqrt{(}n)$. With our given values, this would be $(1/.2)/\sqrt{(}40)$ which approximately equals .79. Since we want the variance, we square this to get $((1/.2)/\sqrt{(}40))^2$.

Table 2: Theoretical vs Simulated Mean & Variance for Simulation 2

	Mean	Variance
Simulated	5	0.648
Theoretical	5	0.625

As we can see by the matrix, the simulated values are quite close to the theoretical values. Now lets see if the histogram becomes a normal distribution. The vertical black line indicates the mean of the simulation.



This distribution is much more Gaussian than the first distribution, as it resembles a bell curve.

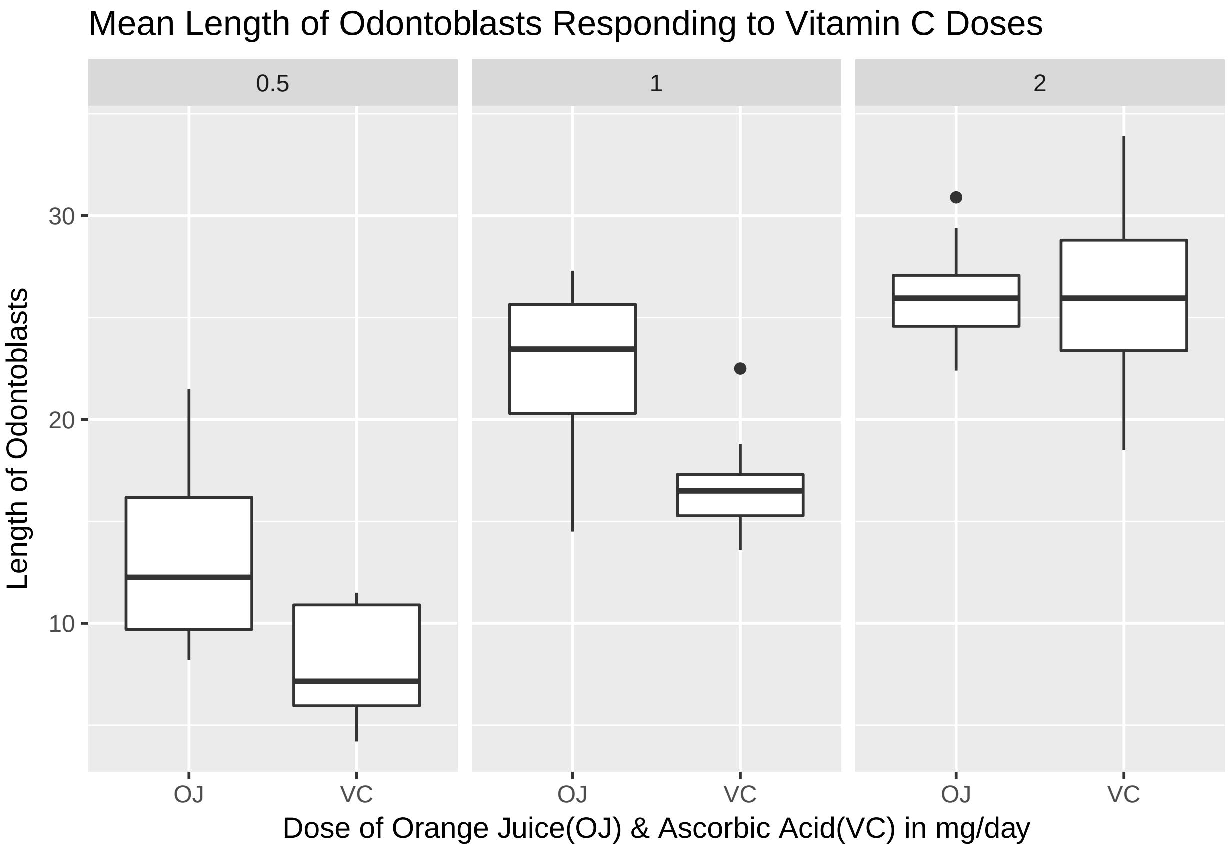
Part 2

For the second part of the assignment, we are asked to provide a basic summary of the ToothGrowth data. This dataset compares different dose levels (0.5, 1 and 2 mg/day) of either orange juice or ascorbic acid to 60 different guinea pigs to determine the response of odontobalsts (cells responsible for tooth growth). To do this, I will show the average length of odontoblasts as well as the standard deviation for the 6 groups (i.e. OJ at the 3 dose levels and ascorbic acid(VC) at the 3 dose levels in a matrix.

Table 3: Mean & Standard Deviations for Different Dosages and Supplement Type

	OJ_0.5	OJ_1.0	OJ_2.0	VC_0.5	VC_1.0	VC_2.0
Mean	13.23	22.70	26.06	7.98	16.77	26.1
Standard Deviation	4.46	3.91	2.65	2.75	2.52	4.8

There are clear differences in means so I will focus the analysis on differences in means. Below is a plot showing the averages for the two types of treatments for the three dose sizes.



There seem to be some clear differences but the 2.0 mg/day dosages don't seem to differ from each other. We will examine this by creating multiple t.tests comparing each of the groups with one another. First, I'm going to create 6 variables containing the relevant information for each group.

Now, t.tests are testing whether each group differs significantly from one another (Ho: they are the same; Ha: they are different). We will use the standard significance level of alpha = 0.05 and only look at the *p*-values for each test. I set the paired argument equal to false (which is the default) since we are comparing means with different guinea pigs. I set the alternative argument to "two.sided" since this is standard for comparison of means. The matrix that returns shows all of the p-values.

Table 4: P-Value Table for Different Supplements and Dosage Levels

	VC_0.5	VC_1.0	VC_2.0
OJ_0.5	0.006	0.046	0.000
OJ_1.0	0.000	0.001	0.097
OJ_2.0	0.000	0.000	0.964

Conclusions

With alpha = 0.05, we reject the Null hypothesis that the means are the same between all groups except between orange juice 2.0mg/ascorbic acid 2.0mg and orange juice 1.0mg/ascorbic acid 2.0mg. It appears that at smaller doses, orange juice has a greater impact on the length of odontoblasts but as doses reach 2 mg/day, the differences between orange juice and ascorbic acid appear to be null. This is assuming that the data are normally distributed and that variances are not equal. ##Appendix(R code)

```
options(width = 100, scipen = 999, digits = 3)
library(ggplot2)
library(knitr)
library(datasets)
library(dplyr)
library(gridExtra)
set.seed(100)
expon <- rexp(1000, .2)
mean0 <- mean(expon)
var0 <- var(expon)
vartheory <- 1/.2^2
meantheory <- 1/.2
matrix1 <- matrix(c(mean0, var0, meantheory, vartheory), byrow = T, 2, 2)
colnames(matrix1) <- c("Mean", "Variance")
rownames(matrix1) <- c("Simulated", "Theoretical")
kable(matrix1, caption = "Theoretical vs Simulated Mean & Variance for Simulation 1")

mns <- NULL
for (i in 1:1000) mns = c(mns, mean(rexp(40, .2)))
mean1 <- mean(mns)
var1 <- var(mns)
meantheory1 <- 1/.2
vartheory1 <- ((1/.2)/sqrt(40))^2
matrix2 <- matrix(c(mean1, var1, meantheory1, vartheory1), byrow = T, 2, 2)
colnames(matrix2) <- c("Mean", "Variance")
rownames(matrix2) <- c("Simulated", "Theoretical")
kable(matrix2, caption = "Theoretical vs Simulated Mean & Variance for Simulation 2")

p1 <- qplot(expon) +
        geom_histogram(fill = "red", col = "black") +
        geom_vline(xintercept = mean0) +
        xlab("Random Exponential Value") +
        ylab("Count") +
        labs(caption = "Simulation 1") +
        ggtitle("Histogram of 1000 Exponent\n Simulations of Lambda = .2 ")

p2 <- qplot(mns) +
        geom_histogram(fill = "blue", col = "black") +
        geom_vline(xintercept = mean1) +
        labs(caption = "Simulation 2") +
        ggtitle("Dist. of Averages of 40 Exponentials Simulated\n 1000 Times Lambda = .2") +
        xlab("Value of each Simulation Average") +
        ylab("Count")

grid.arrange(p1, p2, ncol=2)
tooth_groups_means <- ToothGrowth %>%
        group_by(supp, dose) %>%
        summarise(length = mean(len))
tooth_groups_sd <- ToothGrowth %>%
        group_by(supp, dose) %>%
        summarise(length = sd(len))
tooth_matrix <- matrix(c(tooth_groups_means$length, tooth_groups_sd$length), byrow = T, 2, 6)
rownames(tooth_matrix) <- c("Mean", "Standard Deviation")
```

```r
colnames(tooth_matrix) <- c("OJ_0.5", "OJ_1.0", "OJ_2.0", "VC_0.5", "VC_1.0", "VC_2.0")
kable(tooth_matrix, caption = "Mean & Standard Deviations for Different Dosages and Supplement Type")
ggplot(ToothGrowth, aes(supp, len)) +
        geom_boxplot() +
        facet_wrap(~dose) +
        xlab("Dose of Orange Juice(OJ) & Ascorbic Acid(VC) in mg/day") +
        ylab("Length of Odontoblasts") +
        ggtitle("Mean Length of Odontoblasts Responding to Vitamin C Doses")

VC_0.5 <- ToothGrowth %>%
        filter(supp == "VC" & dose == 0.5)
VC_1.0 <- ToothGrowth %>%
        filter(supp == "VC" & dose == 1.0)
VC_2.0 <- ToothGrowth %>%
        filter(supp == "VC" & dose == 2.0)
OJ_0.5 <- ToothGrowth %>%
        filter(supp == "OJ" & dose == 0.5)
OJ_1.0 <- ToothGrowth %>%
        filter(supp == "OJ" & dose == 1.0)
OJ_2.0 <- ToothGrowth %>%
        filter(supp == "OJ" & dose == 2.0)
VC_0.5vsOJ_0.5 <- t.test(VC_0.5$len, OJ_0.5$len, paired = F, alt = "two.sided")$p.value
VC_0.5vsOJ_1.0 <- t.test(VC_0.5$len, OJ_1.0$len, paired = F, alt = "two.sided")$p.value
VC_0.5vsOJ_2.0 <- t.test(VC_0.5$len, OJ_2.0$len, paired = F, alt = "two.sided")$p.value
VC_1.0vsOJ_0.5 <- t.test(VC_1.0$len, OJ_0.5$len, paired = F, alt = "two.sided")$p.value
VC_1.0vsOJ_1.0 <- t.test(VC_1.0$len, OJ_1.0$len, paired = F, alt = "two.sided")$p.value
VC_1.0vsOJ_2.0 <- t.test(VC_1.0$len, OJ_2.0$len, paired = F, alt = "two.sided")$p.value
VC_2.0vsOJ_0.5 <- t.test(VC_2.0$len, OJ_0.5$len, paired = F, alt = "two.sided")$p.value
VC_2.0vsOJ_1.0 <- t.test(VC_2.0$len, OJ_1.0$len, paired = F, alt = "two.sided")$p.value
VC_2.0vsOJ_2.0 <- t.test(VC_2.0$len, OJ_2.0$len, paired = F, alt = "two.sided")$p.value

matrix3 <- matrix(c(VC_0.5vsOJ_0.5, VC_0.5vsOJ_1.0, VC_0.5vsOJ_2.0,
                    VC_1.0vsOJ_0.5, VC_1.0vsOJ_1.0, VC_1.0vsOJ_2.0,
                    VC_2.0vsOJ_0.5, VC_2.0vsOJ_1.0, VC_2.0vsOJ_2.0), byrow = F, 3, 3)

row.names(matrix3) <- c("OJ_0.5", "OJ_1.0", "OJ_2.0")
colnames(matrix3) <- c("VC_0.5", "VC_1.0", "VC_2.0")
kable(matrix3, caption = "P-Value Table for Different Supplements and Dosage Levels")
```