Statistical Simulation and Inferential Data Analysis

Inferential Statistics: Course Project

John Snyder

August 26, 2018

1 Overview

The goal of this document is twofold. In the first section, simulation is used to demonstrate the Central Limit Theorem. In the second section, basic statistical inference techniques are used to analyse the `ToothGrowth` data set.

2 Central Limit Theorem

The central Limit Theorem states

2.1 Simulations

```
library(ggplot2)

lambda <- 0.2
n <- 40
m <- 1000

set.seed(2018) # for reproducibility
expSamples <- rexp(n*m, lambda)
expMatrix <- matrix(expSamples, ncol=n, nrow=m)


ggplot(data.frame(exp=expSamples), aes(x=exp)) +
    geom_histogram(color="black", bins=50, aes(y=..density..)) +
    stat_function(fun=dexp,
                  color="red",
                  size=1,
                  args=list(rate=lambda))+
    labs(y="Probability Density",
         x="Observed Value")
```

2.2 Sample vs. Theoretical Mean

Code for the graphs below is based on the code provided in a StackExchange question

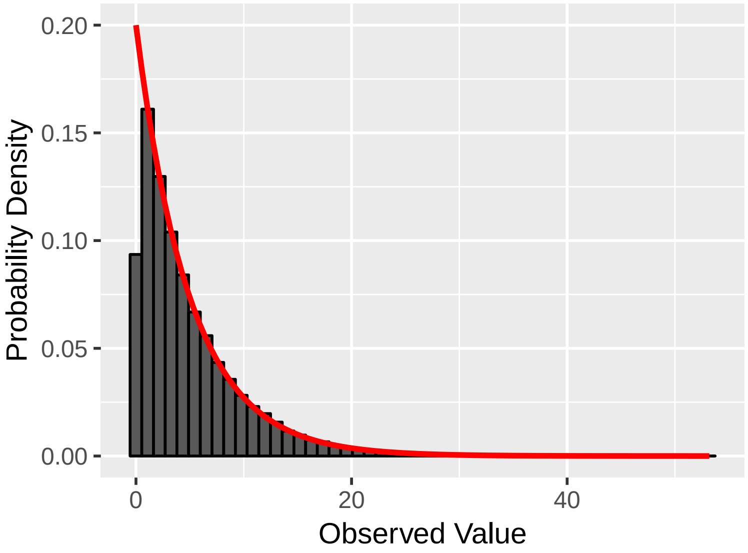

Figure 1: Exponential Distribution Simulated Sample and PDF

```
# Calculate the means of the 1000 Samples
expMeans <- apply(expMatrix, 1, mean)

# Plot the distribution of means, along with the theoretical mean for
# the exponential distribution, the measured mean, and the PDF for
# a normal distribution with a mean equal to the theoretical mean and
# a standard deviation equal to our sample standard deviation.
ggplot(data.frame(means = expMeans), aes(x=means)) +
    geom_histogram(color = "black", bins=50, aes(y=..density..)) +
    stat_function(fun=dnorm,
                  color="red",
                  size=2,
                  args = list(mean=1/lambda,
                              sd=sd(expMeans))) +
    geom_vline(xintercept = 1/lambda, color="red", size=2)+
    geom_vline(xintercept = mean(expMeans), color="blue", size=2) +
    labs(y="Probability Density",
         x="Sample Mean (n=40)")
```

2.3 Sample vs. Theoretical Variance

```
expSDs <- apply(expMatrix, 1, sd)

ggplot(data.frame(sd = expSDs), aes(x=sd)) +
    geom_histogram(color = "black", bins=50, aes(y=..density..)) +
    stat_function(fun=dnorm,
                  color="red",
                  size=2,
                  args = list(mean=1/lambda,
```

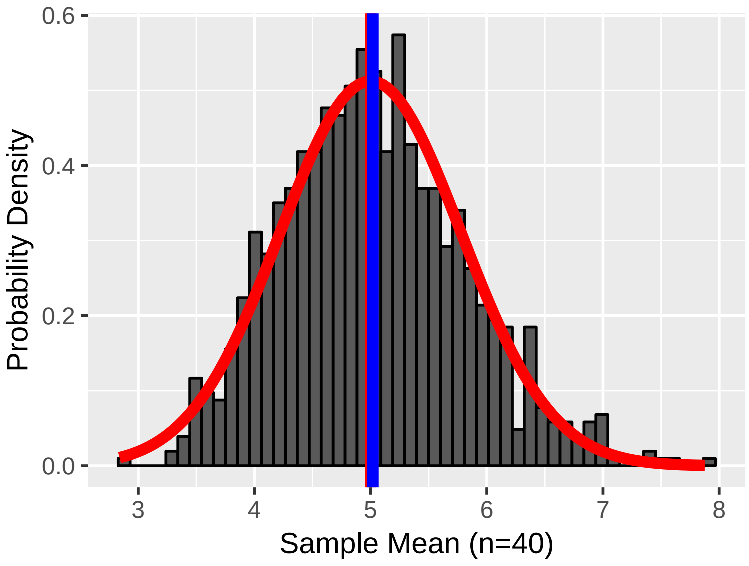

Figure 2: Sample of 1000 means from the exponential distribution, n=40 for each sample

```
                              sd=sd(expSDs))) +
    geom_vline(xintercept = 1/lambda, color="red", size=2)+
    geom_vline(xintercept = mean(expSDs), color="blue", size=2) +
    labs(y="Probability Density",
         x="Sample Standard Deviation (n=40)")
```

3 Inferential Data Analysis

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data("ToothGrowth")
summary <- ToothGrowth %>%
           group_by(dose, supp) %>%
           summarize(mean=mean(len), sd=sd(len), n=n())

knitr::kable(summary, caption = "Summary of Tooth Growth Data")
```

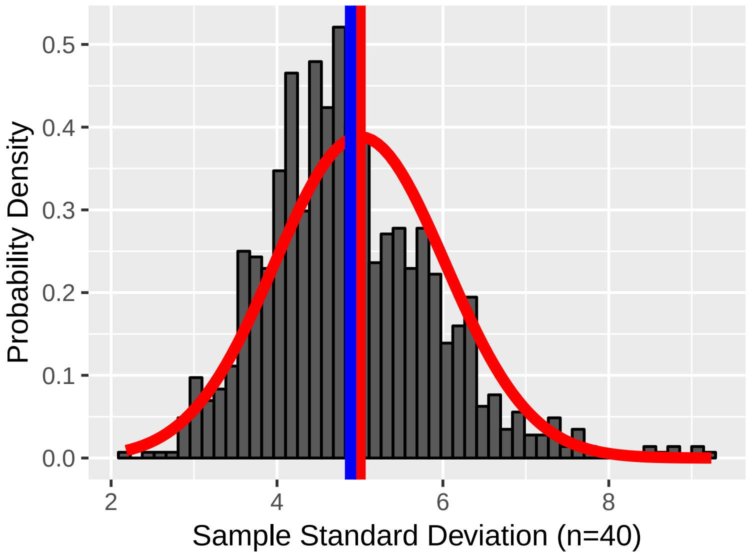

Figure 3: Sample of 1000 standard deviations from the exponential distribution, n=40 for each sample

Table 1: Summary of Tooth Growth Data

dose	supp	mean	sd	n
0.5	OJ	13.23	4.459708	10
0.5	VC	7.98	2.746634	10
1.0	OJ	22.70	3.910953	10
1.0	VC	16.77	2.515309	10
2.0	OJ	26.06	2.655058	10
2.0	VC	26.14	4.797731	10

```
ggplot(ToothGrowth, aes(x=factor(dose), y=len, fill=supp)) +
  geom_boxplot()
```

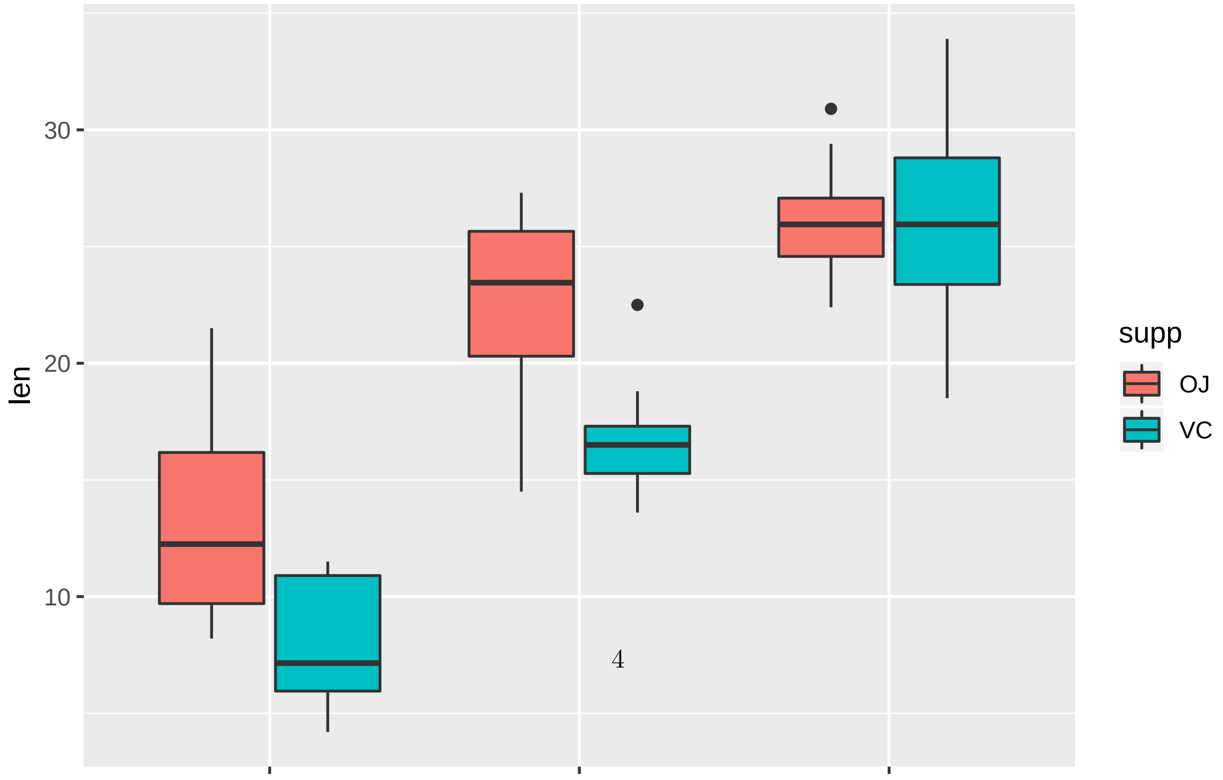

Table 2: Significance Level by Dose

dose	p
0.5	0.0063586
1.0	0.0010384
2.0	0.9638516