Coursera - Statistical Inference Project

Howard Murray

2024-07-24

Project Overview

The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

Part 1: Simulation Exercise Instructions

In this part of the project, we are to investigate the exponential distribution in R and compare it with the **Central Limit Theorem**.

The Central Limit Theorem

The *Central Limit Theorem* says that the estimate of the mean minus the mean of the estimate divided by the standard error of the estimate has a distribution like that of a normal distribution for large n. Mathematically stated:

$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{Estimate-Mean}{Std.Error of Estimate}$

This means that the sample average is approximately normally distributed with a mean given by the population mean and variance given by the standard error of the mean. $\overline{X} \sim N(\mu, \sigma^2/n)$

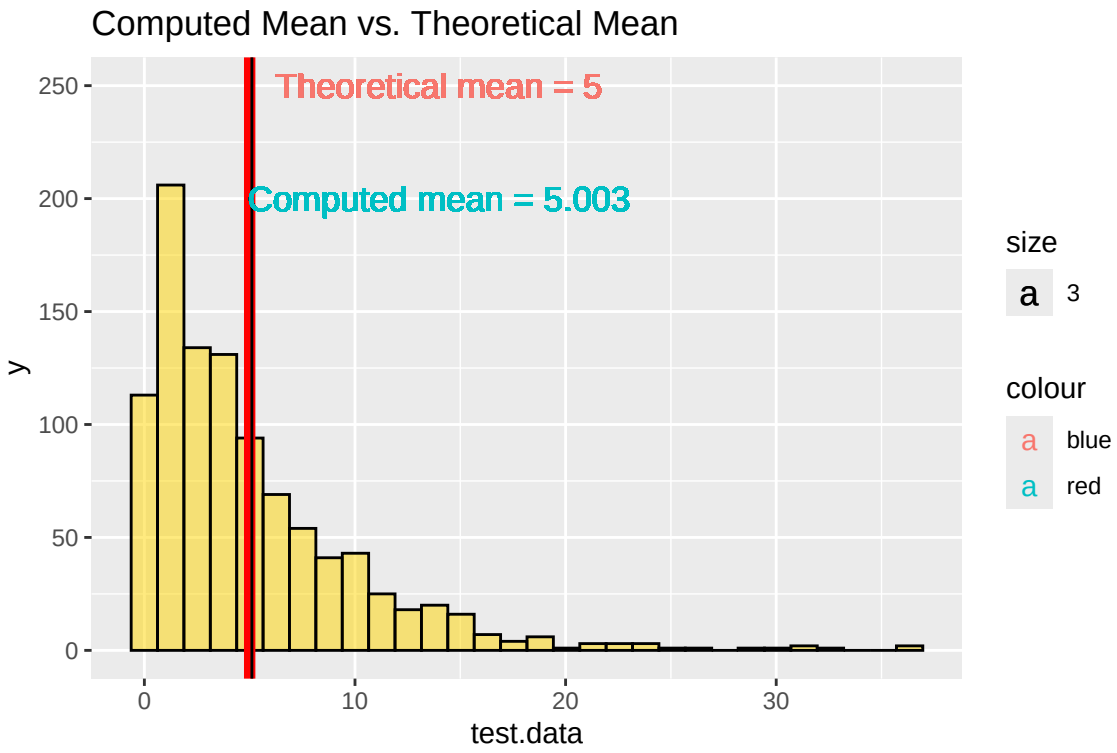
Quick look at an exponential distribution (n = 1000, $\lambda = 0.2$)

I initially plotted the distribution of a thousand random samples of an exponential distribution to analyze its shape and centrality. As expected, the histogram that an exponential distribution doesn't look have the shape of a Gaussian bell curve. Mathematically we know that given **lambda = 0.2**, we compute that the **theoretical mean is equal to 5**. We will run the R code of size n = 1000 with the same lambda to see how close we get to the expected theoretical mean. **The computed mean is 5.003.**

```
library(tidyverse)
library(ggplot2)

set.seed(1234) # It is important to set seed to a defined number so that your research is reproducible.
lambda = 0.2

# First I will run an n = 1000 exponential distribution with lambda = .2 to see what it looks like
test.data <- rexp(1000, rate = lambda)

# Compute the mean of the exponential distribution samples.
mean(test.data)
```

```
## [1] 5.003067
```

```
ggplot(data.frame(test.data), aes(test.data)) +
  geom_histogram(alpha = .5, fill = "gold", color = "black") +
  ggtitle("Computed Mean vs. Theoretical Mean") +
  geom_vline(xintercept = 5, color = "red", lwd = 2) +
  geom_vline(xintercept = mean(rexp(1000, .2), color = "blue", lwd = 2)) +
  geom_text(aes(x = 14, y = 200, label = "Computed mean = 5.003", data = test.data, size = 3,
                color = "red")) +
  geom_text(aes(x = 14, y = 250, label = "Theoretical mean = 5", size = 3, color = "blue"))
```



The exponential distribution was simulated in R with the function, **rexp(n, lambda)**, where lambda (λ) is the rate parameter. The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. Set **lambda = 0.2** for all of the simulations. We were instructed to investigate the distribution using averages of 40 exponentials. In order to approximate the Central Limit Theorem, we are do a thousand iterations.

1. Compare Sample mean and variance to theoretical values.

Mathematically we know that given **lambda = 0.2**, we compute that the **theoretical mean is equal to 5.000**. We ran the R code to prove out our assumptions related to how this simulation would perform by running 1000 iterations of 40 random samples with the same lambda to see how close we get to the expected theoretical mean. **The computed mean is 4.974.**

The computed variance is equal to **0.571**. The theoretical variance is equal to **0.625**.

```
library(tidyverse)
library(ggplot2)

set.seed(1234) # It is important to set seed to a defined number so that your research is reproducible.
lambda = 0.2
exp.samples = 40
```

```r
# Run a simulation of 40 exponential samples 1000 times and do some exploratory analysis.
sim.Means = NULL

for (i in 1:1000){
  sim.Means = c(sim.Means, mean(rexp(exp.samples, lambda)))
                }

summary(sim.Means)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.170   4.429   4.938   4.974   5.505   7.390
```

```r
# Compute mean of simulation Means
mean(sim.Means)
```

```
## [1] 4.974239
```

```r
# Compute theoretical mean
theoretical.Mean <- 1/lambda
theoretical.Mean
```

```
## [1] 5
```

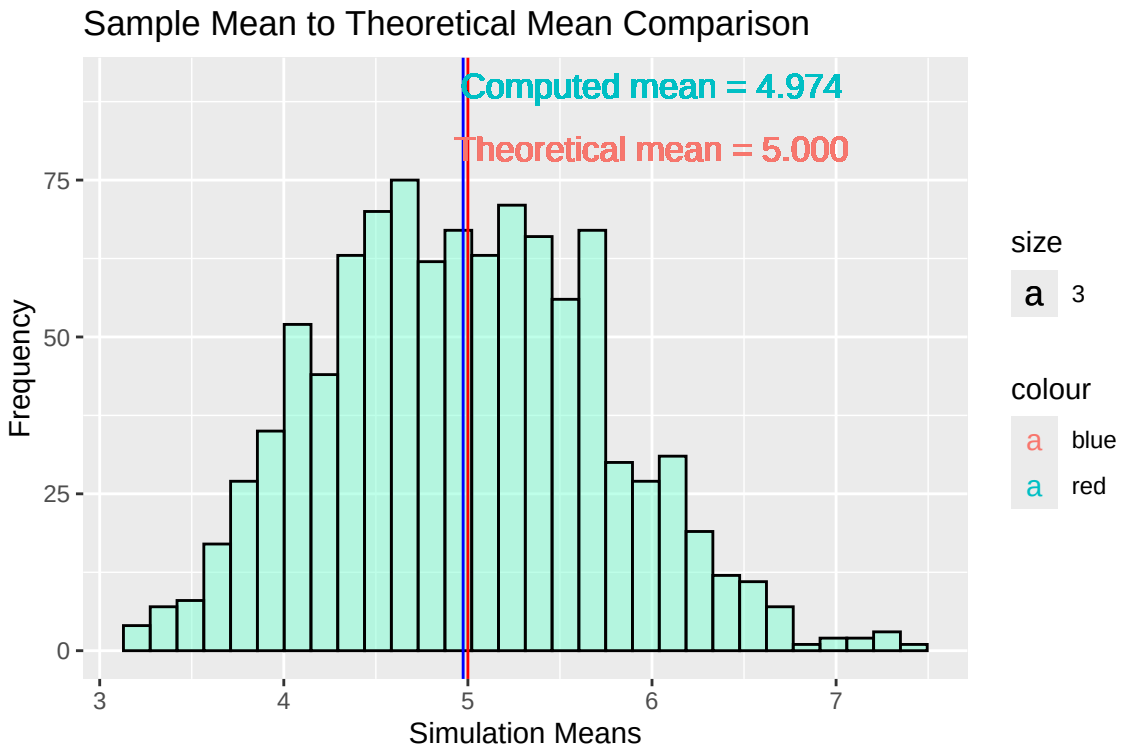
```r
# compute variance of simulation means
var.sim.Means <- var(sim.Means)
var.sim.Means
```

```
## [1] 0.5706551
```

```r
# Compute theoretical variance
theoretical.var <- (1/lambda)^2/40
theoretical.var
```

```
## [1] 0.625
```

```r
# Plot a visualization of the means to compare sample mean to the theoretical mean
ggplot(data.frame(sim.Means), aes(sim.Means)) +
  geom_histogram(alpha = 0.5, fill = "aquamarine", color = "black") +
  labs(x = "Simulation Means", y = "Frequency", title = "Sample Mean to Theoretical Mean Comparison") +
  geom_vline(xintercept = 5, color = "red") +
  geom_vline(xintercept = mean(sim.Means), color = "blue") +
  geom_text(aes(x = 6, y = 90, label = "Computed mean = 4.974", data = sim.Means, size = 3, color = "re
  geom_text(aes(x = 6, y = 80, label = "Theoretical mean = 5.000", size = 3, color = "blue"))
```



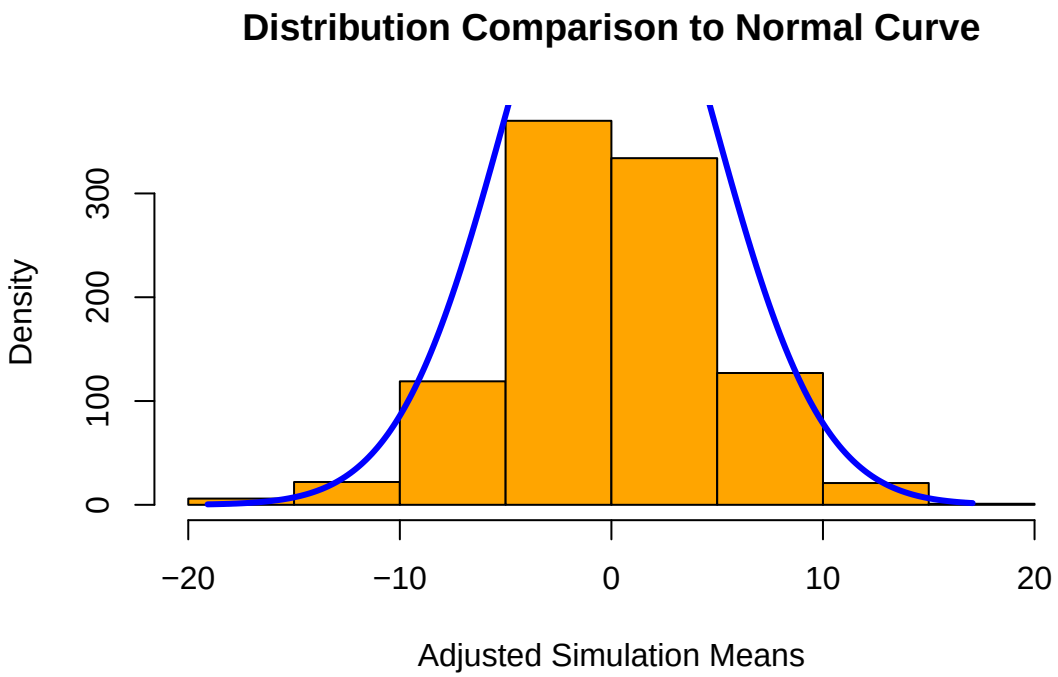
Is the distribution approximately normal?

Let's compare how closely the sample distribution aligns with that of a standard normal distribution as stipulated in *The Central Limit Theorem*.

$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{Estimate-Mean}{Std.Error of Estimate}$

We can see from graph below that when we take the average of the estimates and adjust them by the mean value and standard error that the histogram approaches a standard normal curve with the same mean and standard deviation as CLT suggests. We would expect that as the number of samples in the histogram increases, the more it aligns with a standard normal curve.

```
# Subtract mean from distribution to center at zero
adj.sim.Means <- sim.Means - mean(sim.Means)

# Create data for a standard normal distribution with sd = Computed sd from above or sd = 0.7554171
set.seed(1234)
Norm.dist <- rnorm(1000, mean = 0, sd = sim.Means)

# Plot histogram
hist_data <- hist(Norm.dist, col = "orange", xlab = "Adjusted Simulation Means", ylab = "Density", main

# Define x and y values for normal curve

x_values <- seq(min(Norm.dist), max(Norm.dist), length = 1000)
y_values <- dnorm(x_values, mean = mean(Norm.dist), sd = sd(Norm.dist))
y_values <- y_values * diff(Norm.dist[1:2]) * length(Norm.dist)

# Overlay curve on histogram
lines(x_values, y_values, lwd = 3, col = "blue")
```



Part 2: ToothGrow Data Analysis

Exploratory Analysis of ToothGrow Dataset.

This part of the project required us to examine the **ToothGrowth** dataset. This data was collected as part of a study to measure the length of *odontoblasts* (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C **(0.5 mg/day, 1 mg/day, and 2 mg/day)** by one of two delivery methods: **Orange Juice** or via **Ascorbic Acid** (a form of vitamin C coded as VC). More detailed information on the study can be found at: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html.

Boxplots of the dataset were created. The plots were faceted by the three dosage groupings (0.5 mg/day, 1.0 mg/day, and 2.0 mg/day). A look at the faceted boxplots suggests that increasing the dosage of vitamin C regardless of Supply Type has a more positive impact on the response *Tooth Length* as the overall means for each group of subjects show increased length for each dosage grouping. It also appears from the boxplots that the *OJ* Supply type tends to yield a more positive responses to tooth growth versus the *VC* method.

```
# Exploratory analysis of the ToothGrowth dataset.
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 ...
```

```
# Check for incompletes
sum(is.na(ToothGrowth))
```

```
## [1] 0
```

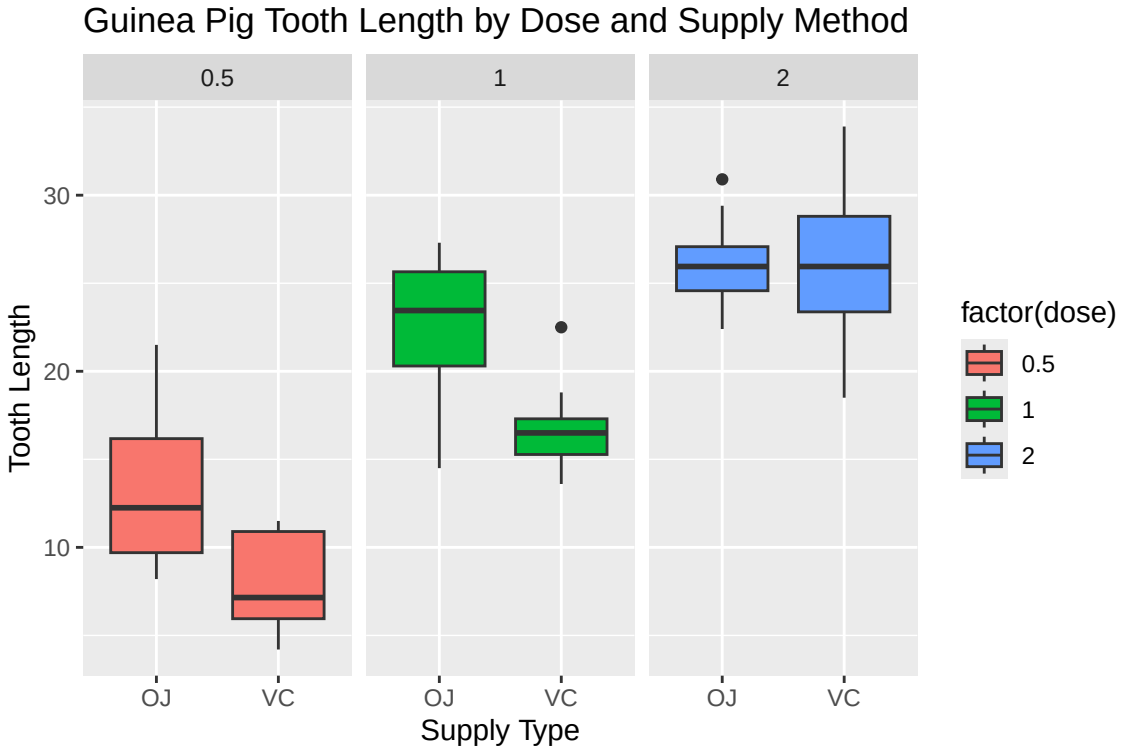
```
# Cursory review of the contents of the dataframe, summary, data types, etc.
head(ToothGrowth)
```

```
##    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
```

```
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
```

```
# Visualization: Create faceted plots to see if there are any visual indications that dosage or deliver

ggplot(ToothGrowth, aes(supp, len)) +
 geom_boxplot(aes(fill = factor(dose))) +
  facet_wrap(.~factor(dose)) +
  labs(x = "Supply Type", y = "Tooth Length", title = "Guinea Pig Tooth Length by Dose and Supply Metho
```



Inferential Statistics

While the above visualization suggests a certain relationship between dosage size and delivery method to tooth length, we cannot be certain without more statistical testing. This is an opportunity for us to apply inferential statistical methods like the 2-sample t-Test to see if we can mathematically see a relationship.

The impact of Dosage amount on Tooth Growth

First step is to establish a null (H_0) and alternative hypothesis (H_a). Our null hypothesis is: *Dosage amount has no effect on Tooth Length*. Our alternative hypothesis is that *increased dosage amount positively affects Tooth Length*. To prove or disprove our hypothesis, we will compare the response, Length for the guinea pigs that received the lowest dose (0.5 mg/day) to the response of those who received the highest dose (2 mg/day).

Using R's **t.test** function, we will compare the low dosage response to the high dosage response. The test is not paired. The alternative setting in the function was set to **"greater"**. The test yielded a **p-value = 2.199e-14**. This is much less than our threshold of 0.5, therefore **we reject the null hypothesis in favor of the alternative and say there is statistically significant evidence that a higer dosage of vitamin see leads to longer teeth**. Comparison of the means of the two populations shows $Mean_Hi = 26.1$ and $Mean_Lo = 10.61$ or a difference of **15.49 units**.

```
# Capture 0.5 mg/day data
low.dose <- subset(ToothGrowth, dose == 0.5)

# Capture 2.0 mg/day data
high.dose <- subset(ToothGrowth, dose == 2)

# Run t.test function
t.test(high.dose$len, low.dose$len, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  high.dose$len and low.dose$len
## t = 11.799, df = 36.883, p-value = 2.199e-14
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  13.27926      Inf
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

The impact of Vitamin C Supply Type on Tooth Growth

As in the previous case, establish a null (H_0) and alternative hypothesis (H_a). Our null hypothesis is: *Vitamin C Supply type has no effect on Tooth Length*. Our alternative hypothesis is that *one of the delivery method has a greater positive affect on Tooth Length than the other*. To prove or disprove our hypothesis, we will compare the response, Length for the guinea pigs that received the vitamin C via OJ to the subjects who received Ascorbic Acid (VC).

Again we use R's **t.test** function, we will compare the **"OJ supply"** to the **"VC supply"** length response. The test is not paired. The alternative setting in the function was set to **"greater"**. The test yielded a **p-value = 0.03032**. This is below our threshold of 0.5, therefore **we reject the null hypothesis in favor of the alternative and say there is statistically significant evidence that a the OJ supply method leads to longer teeth in the subjects**. Comparison of the means of the two populations shows $Mean_Oj = 20.6633$ and $Mean_Vc = 16.963$ or a difference of **3.7 units**.

```
# Capture OJ supply data
OJ.supp <- subset(ToothGrowth, supp == "OJ")

# Capture VC supply data
VC.supp <- subset(ToothGrowth, supp == "VC")

# Run t.test function
```

```r
t.test(OJ.supp$len, VC.supp$len, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ.supp$len and VC.supp$len
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```