Statistical Inference Project

by C. McBride

Part 1: Simulation Exercise

Overview

This part of the report explores distributions of a large number of random samples from the exponential distribution. The exponential distribution has a mean of 1/lambda where lambda is the rate of growth for the exponential function. The mean and variation of the samples will be compared to the population or theoretical mean and variation.

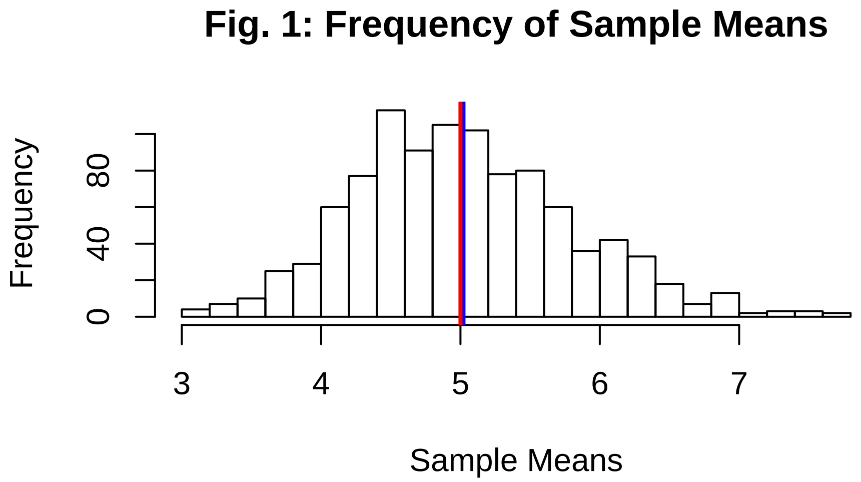
Simulations

To begin with, a value of 0.2 is assigned as the rate for the exponential function. A sampling of 40 observations is repeated 1000 times and subsequently organized into a 1000x40 matrix for ease of processing. Row-wise calculation of the mean for each 40 observation sampling is completed and then passed to the analysis in the next section of this report.

Distribution of Random Sample Distributions

Sample Mean vs Theoretical Mean

The frequency of means is plotted.

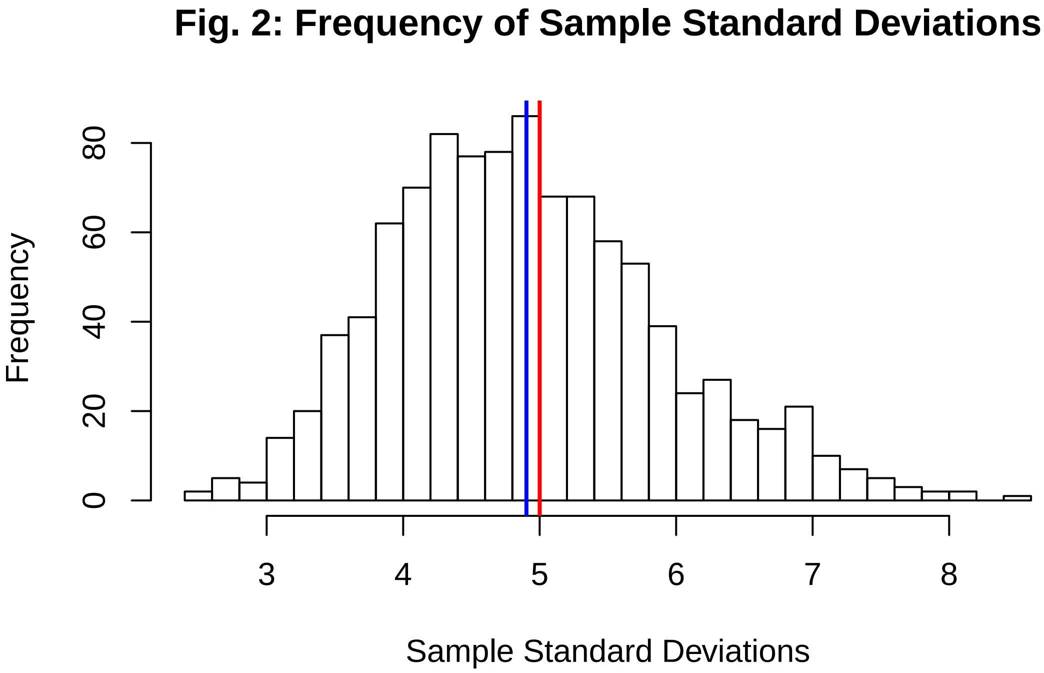


The distribution is nearly Gaussian in accord with the Central Limit Theorem. Also, the theoretical mean (mu=5) is nearly identical to the sample mean (mean=5.022) which is as expected since although the sample distribution has its own distribution it is centered at mu the population mean.

Sample Variance vs Theoretical Variance

First the standard deviation for each 40 observation sample is calculated and collectively assigned to a variable.

Just as with the sample mean, the frequencies of the sample variations as measured by standard deviation is plotted. The mean of the sample standard deviations (blue) and the population standard deviation (red) are also plotted as vertical lines.



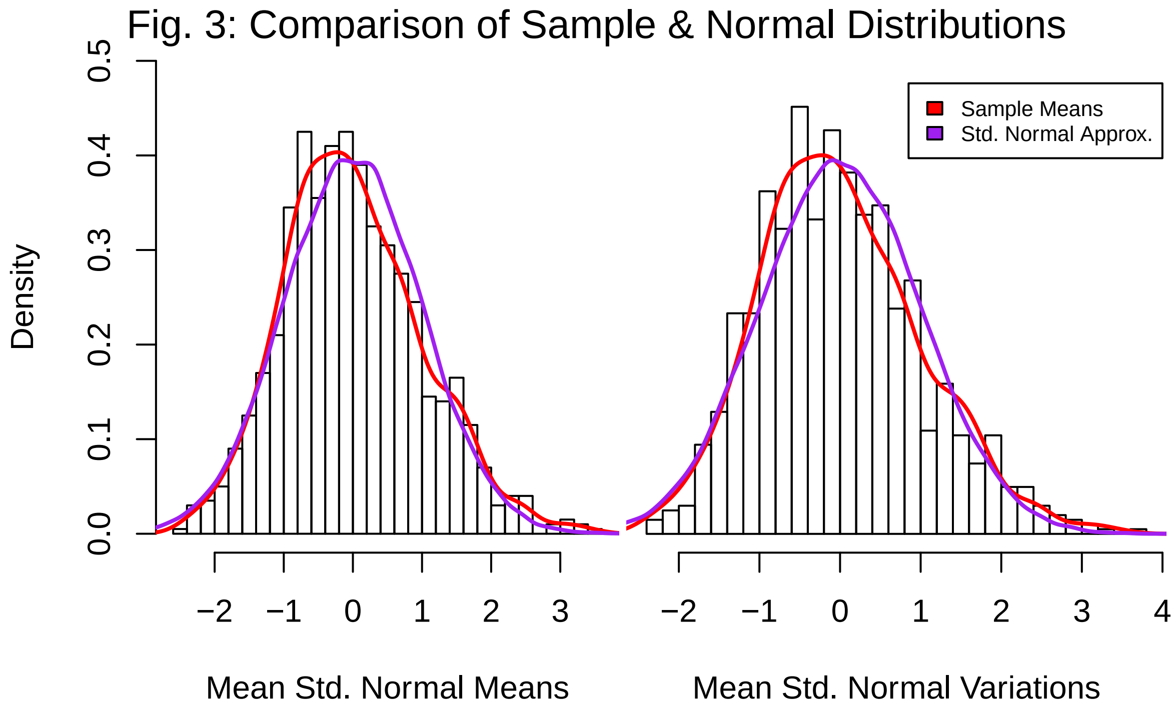
In the case of sample variation, the distribution is near Gaussian despite a slight right skew. The mean of the sample standard deviations (s=5) is very close as an estimate to the population or theoretical standard deviation (sigma=5). This makes sense since we know that the distribution of the sample variance (in this case standard deviation) is centered on the parameter that is is estimating (i.e. the population variation).

Distribution

Next, we'll show that the two distributions that we found, one for the sample means and the other for the sample variations, are both near normal in their distributions.

First we'll scale each of the vectors to be standard normal.

Then we'll plot them along with a density plot for the standard normal distribution.



The distribution of sample means is a bit right-skewed but is nearly normal. As with the distribution of sample means, this distribution also proves very close to normal despite some outliers that show some leverage over the density curve.

Part 2: Basic Inferential Data Analysis

Overview

In this section of the report, hypothesis testing is conducted using a toy dataset (ToothGrowth) from the datasets package.

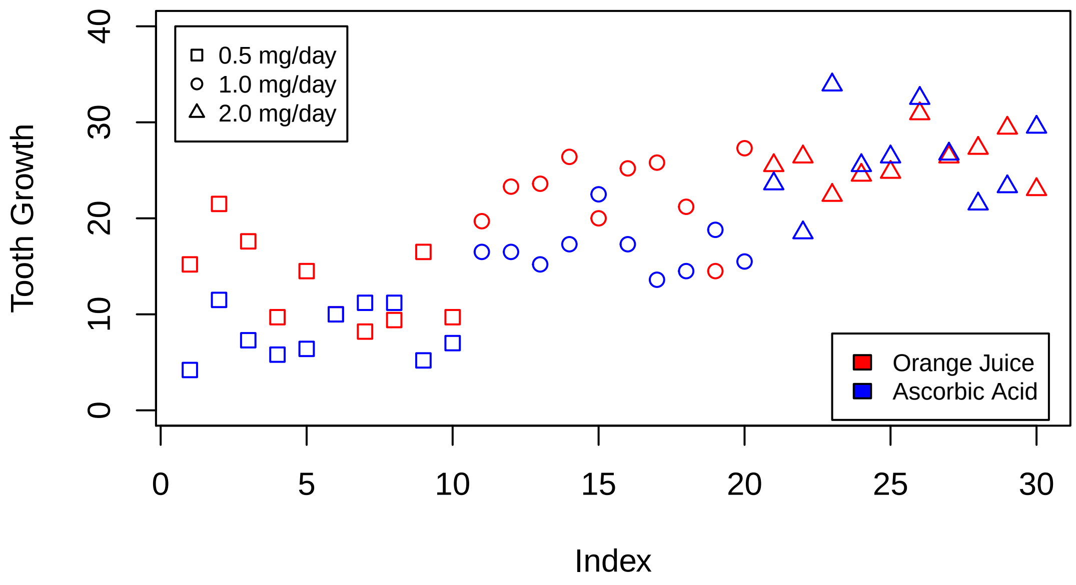
Basic Data Summary

The dataset contains the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dosages of vitamin C (0.5, 1, 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid. The animals were evenly distributed across the six groups created by the intersection of the variables `supplement` and `dose`.

The data read into a tibble object with dimension 60 by 3 for analysis.

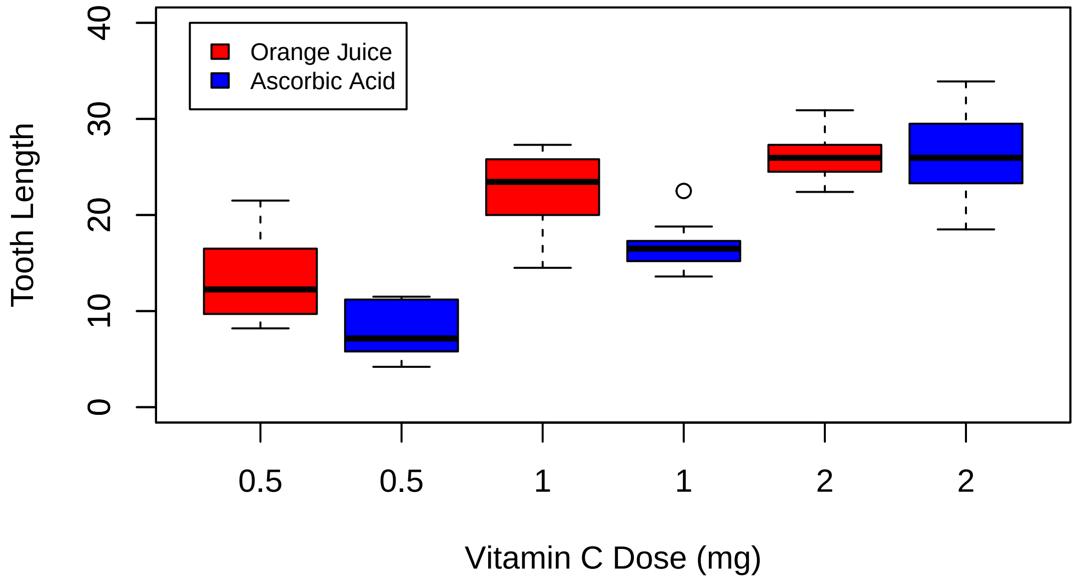
A scatterplot of tooth length by observation index is plotted. The variables of `supplement` and `dose` are coded using color and shape respectively.

Fig. 4: Tooth Growth by Supplement and Dose



The plot shows a clear relationship between `dose` and tooth length. Any relationship between `supplement` and tooth length is not as clear. To explore that relationship further a boxplot is created.

Fig. 5: Guinea Pigs' Tooth Growth



Here again, the positive relationship between `dose` and tooth length is evident. The boxplot also helps to clarify the relationship between the categorical variable `supplement` and tooth length. For all dose levels except 2.0 mg/day, there appears to be significant difference between tooth length related to orange juice compared to that related to ascorbic acid. In the case of the exception, the 2.0 mg/day dose, the means appear to be nearly identical.

We'll next test the differences in means using t-tests.

Inferential Testing

To start, a test is performed to look for a significant difference in mean tooth length by supplement without controlling for dose. Since the subjects for each subgroup are unique, that is the study is not paired, we'll set the `paired` parameter to `FALSE`. Also, because the variations of the two subsets are unequal ($s_{oj}^2 =$ 43.6334368, $s_{aa}^2 =$ 68.3272299), the parameter `var.equal` is set to `FALSE`

```
# t-test for difference in means
overall_test <- t.test(vd_all$len, oj_all$len, paired = FALSE)
```

The p-value for the difference in overall means (p = 0.0606345) is very close to the commonly accepted alpha value for significance 0.05. At the same time, the corresponding confidence interval of 95% includes zero (lower bound = -7.571, upper bound = 0.171), but only barely. This is at once reason to not reject H_0 and cause for further investigation since the borderline statistics and the exploratory plots suggest that controlling for dosage might yield significant differences.

So next we'll perform t-tests on the difference in means of the two supplements this time controlling for dose. Since none of the tests are paired and because the variations for all dose groups compared by supplement are different, we'll keep the parameters `paired` and `var.equal` as both `FALSE`.

```
# t-test for difference in means by dose
dose_0.5 <- t.test(oj_all[oj_all$dose==0.5,]$len, vd_all[vd_all$dose==0.5,]$len,
                   paired = FALSE)

dose_1.0 <- t.test(oj_all[oj_all$dose==1,]$len, vd_all[vd_all$dose==1,]$len,
                   paired = FALSE)

dose_2.0 <- t.test(oj_all[oj_all$dose==2,]$len, vd_all[vd_all$dose==2,]$len,
                   paired = FALSE)
```

T-tests comparing the two vitamin C supplement types by similar dose show a significant difference between the mean tooth length between the doses of 2.0 mg was practically non-existent (t = -0.046, p = 0.9638516, $\bar{x}_{oj} =$ 26.06, $\bar{x}_{aa} =$ 26.14).

Overall the difference in apparent effect of the two supplements while not controlling for dose is near significant and warranted further inquiry. At lower doses, orange juice showed a significantly stronger correlation to increased tooth growth, but at the highest dose of 2.0mg/day the two associated mean growth values were nearly identical. Since the t-test used compensates for unequal variation, the results assume that the samples are from normally distributed populations and as representative of those populations are themselves normally distributed.

Citations

Formatting for the boxplot was largely inspired by code by Roger Bivand as shown in the documentation for the boxplot() function.

Information about the dataset was copied from its documentation at help(ToothGrowth).

Appendices

Code Chunks

Code for simulations:

```r
# set seed for reproducible results
set.seed(7463)

# simulation variables
n <- 40  #number of random observations taken per sample
B <- 1000  #number of samples

# create simulation matrix
M <- matrix(rexp(n*B, 0.2), B, n)

# get mean of each sample
M_means <- apply(M, 1, mean)
```

Code for Figure 1:

```r
# histogram of means of sampled distributions
hist(M_means, breaks=30,
     main="Fig. 1: Frequency of Sample Means", xlab = "Sample Means")

# add vertical lines for sample and theoretical means
abline(v=mean(M_means), col="blue", lwd=2) # mean of sampled means
abline(v=1/.2, col="red", lwd=2) # theoretical mean
```

Code for calculation of variances:

```r
# calculate sd of samples
M_variations <- apply(M, 1, sd)
```

Code for Figure 2:

```r
# histogram of means of sampled distributions
hist(M_variations, breaks=30,
     main="Fig. 2: Frequency of Sample Standard Deviations",
     xlab = "Sample Standard Deviations")

# add vertical lines for sample and theoretical means
abline(v=mean(M_variations), col="blue", lwd=2) # mean of sampled means
abline(v=1/.2, col="red", lwd=2) # theoretical mean
```

Code for standard normalization of data:

```r
# scale sample means and variations for standard normal
means_sn <- scale(M_means)
variations_sn <- scale(M_variations)
```

Code for Figure 3:

```r
# set plot parameters for two panel plot
par(mfrow=c(1, 2), mar=c(5.1, 4.1, 4.1, 0.1))

# histogram of standardized normalized sample means
hist(means_sn, breaks=30, prob=TRUE,
     main=NA,
     xlab="Mean Std. Normal Means", ylab="Density", ylim=c(0,0.5))
lines(density(means_sn), col="red", lwd=2)
lines(density(rnorm(100000)), col="purple", lwd=2)


# reset margins for second plot
par(mar=c(5.1, 0.1, 4.2, 2.1))

# histogram of standardized normalized sample means
hist(variations_sn, breaks=30, prob=TRUE,
     main=NA,
     xlab="Mean Std. Normal Variations",ylab=NA, ylim=c(0,0.5), yaxt="n")
lines(density(means_sn), col="red", lwd=2)
lines(density(rnorm(100000)), col="purple", lwd=2)

# add common title
mtext("Fig. 3: Comparison of Sample & Normal Distributions", cex=1.25, adj=1.25)

# add common legend
legend(.85, 0.48, c("Sample Means", "Std. Normal Approx."),
       fill=c("red", "purple"), cex=0.65)
```

Code for Figure 4:

```
# subset by supplement
oj_all <- ToothGrowth[ToothGrowth$supp=="OJ",]
vd_all <- ToothGrowth[ToothGrowth$supp=="VC",]

# scatterplot of growth by dose by supplement
plot(oj_all$len, pch=oj_all$dose, col="red",
     main="Fig. 4: Tooth Growth by Supplement and Dose",
     ylab="Tooth Growth",
     ylim=c(0,40))
points(vd_all$len, pch=vd_all$dose, col="blue")
legend(0.5, 40, c("0.5 mg/day", "1.0 mg/day", "2.0 mg/day"),
       pch=c(0,1,2), cex=0.75)
legend(23, 8, c("Orange Juice", "Ascorbic Acid"),
       fill=c("red", "blue"), cex=0.75)
```

Code for Figure 5:

```
# boxplots of distributions of tooth growth by supplement and dose

boxplot(len~dose, data=oj_all, col="red",
        boxwex=0.4, at=1:3-.25,
        main="Fig. 5: Guinea Pigs' Tooth Growth",
        xlab = "Vitamin C Dose (mg)", ylab = "Tooth Length",
        xlim=c(0.5, 3.5), ylim=c(0,40))
boxplot(len~dose, data=vd_all, boxwex=0.4, at=1:3+.25, col="blue", add=TRUE)
legend(0.5, 40, c("Orange Juice", "Ascorbic Acid"),
       fill=c("red", "blue"), cex=0.75)
```