Comparison of Exponential Distribution Sample Mean vs. Central Limit Theorem (CLT) Normal Distribution

Course 6 (Statistical Inference) / Assessment 1 Part 1

Paul Ringsted, 20th January 2019

Synopsis

In this report we analyze the distribution of means from independent identically-distributed (IID) exponentials, and compare it to the theoretical Normal distribution as given by the Central Limit Theorem (CLT). We analyzed the distribution of 1000 simulations of the mean over 40 draws from the exponential distribution, with rate parameter $\lambda = 0.2$. We also determined the theoretical Normal distribution for this statistic per the CLT, compared these two distributions using density plots, and also compared back to the original exponential distribution. Finally, we performed further analysis using a permutation test of the difference in means between samples of these populations. The analysis supports the conclusion that, the distribution of means of IID samples drawn from an exponential distribution, conforms to a Normal distribution as given by the CLT.

Simulation of Exponential Samples

We first set up a data frame with the results of 1000 simulations of 40 exponentials with $\lambda = 0.2$, and prep some variables for analysis. Figure 1 shows a histogram of the resulting distribution and it's mean.

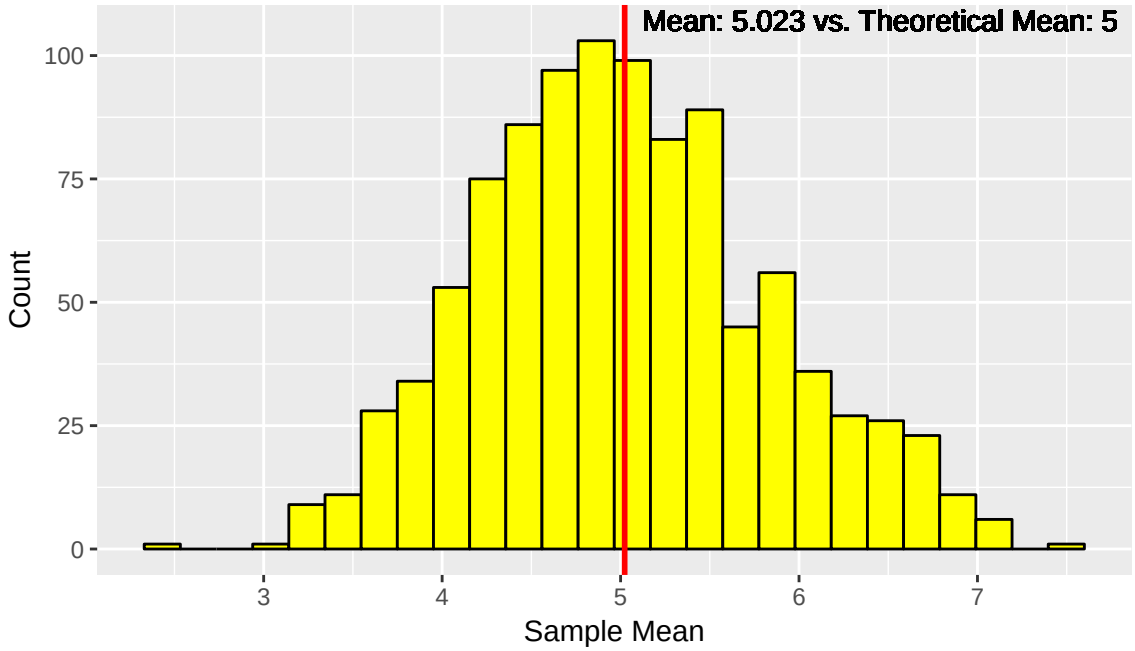

Figure 1: Distribution of Mean of IID Samples from Exponential Distribution

Central Limit Theorem - Sample vs. Theoretical Statistics

According to the **Central Limit Theorem**, the distribution of the mean of the independent, identically distributed (IID) exponentials, $\overline{X}$, will converge to a Normal distribution.

If $X \sim Exp(\lambda)$ and $\mu = E[X] = \frac{1}{\lambda}$ and $\sigma^2 = Var[X] = \frac{1}{\lambda^2}$, per CLT: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

$\implies \overline{X} \sim N(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$ for $\lambda = 0.2$, n = 40

$\implies \overline{X} \sim N(\ 5\ ,\ 0.625\ )$

Summary of Mean and Variance:

Statistic	Exponential Sample	Theoeretical (CLT)	Diff (%)
Mean	5.023	5	0.46
Variance	0.657	0.625	5.13

Distribution vs. Central Limit Theorem & Original Exponential

The histogram in Figure 1 for this simulation certainly appears to fit a Gaussian model. To get a clearer picture of this fit we can view this as density plot and overlay the theoretical Normal distribution given by the CLT. Figure 2 shows these density plots, and confirms a high degree of fit. Figure 2 also shows the density function of the original exponential distribution, which differs greatly from the distribution of its sample means.

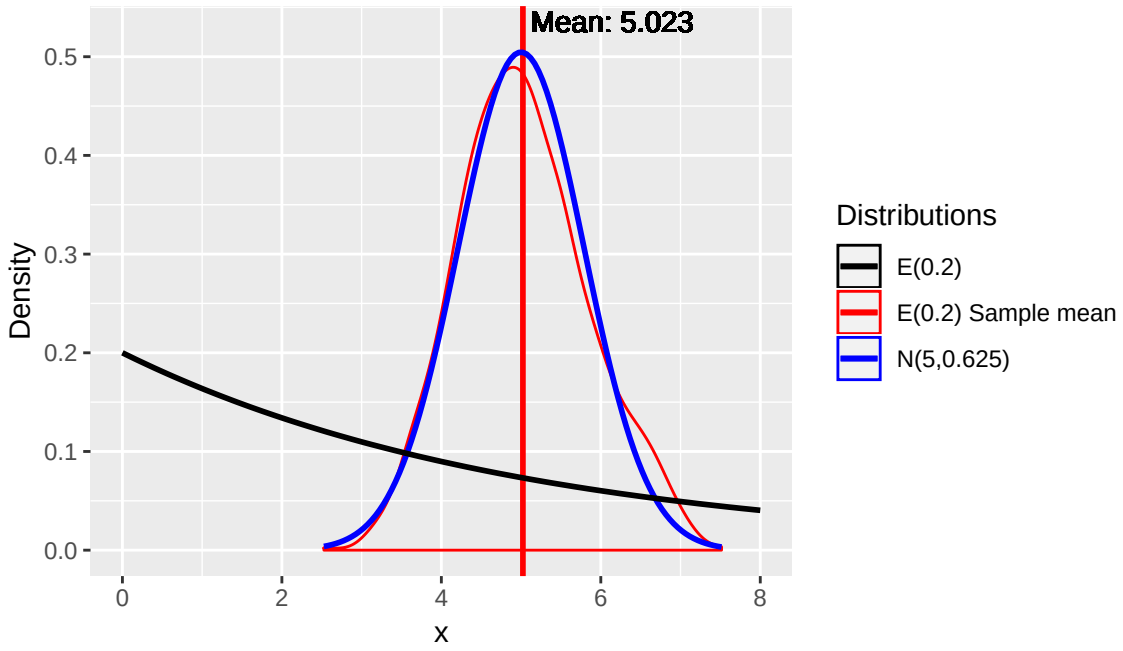

Figure 2: Density of Mean of IID Exponentials vs. CLT Normal and original Exponential Distributions

Conclusion: This plot indicates that the simulation distribution of exponential sample means certainly differs greatly from the exponential distribution from which these samples are drawn, and the simulation distribution appears to fit well to the theoretical Normal distribution as per the CLT.

Further Analysis vs. Central Limit Theorem (Permutation Test)

We can further examine this by taking a similar size random sample from this Normal distribution, and perform a permutation test of these two datasets, to see if there is any statistical difference between them. The test will take 10,000 permutations of these two datasets, with each data element randomly re-assigned to be either from the exponential sample mean or Normal populations. We take the difference in mean of the two new populations, for each permutation, and the original population, to see if there is any distinctive permutation of these populations. Our null hypothesis is

$H_0 : \overline{X} \sim N(\ 5\ ,\ 0.625\ )$

Table 2 summarizes the observed (non-permuted) means of these datasets. Figure 3 shows a plot of the resulting differences in means between the permutated populations, along with the observed (non-permuted) difference in mean, which falls within the same range.

Table 2: Mean of Permutation Test Populations

Mean of Exp Sample Means	Mean of Normals	Observed Diff
5.023	4.99	0.033

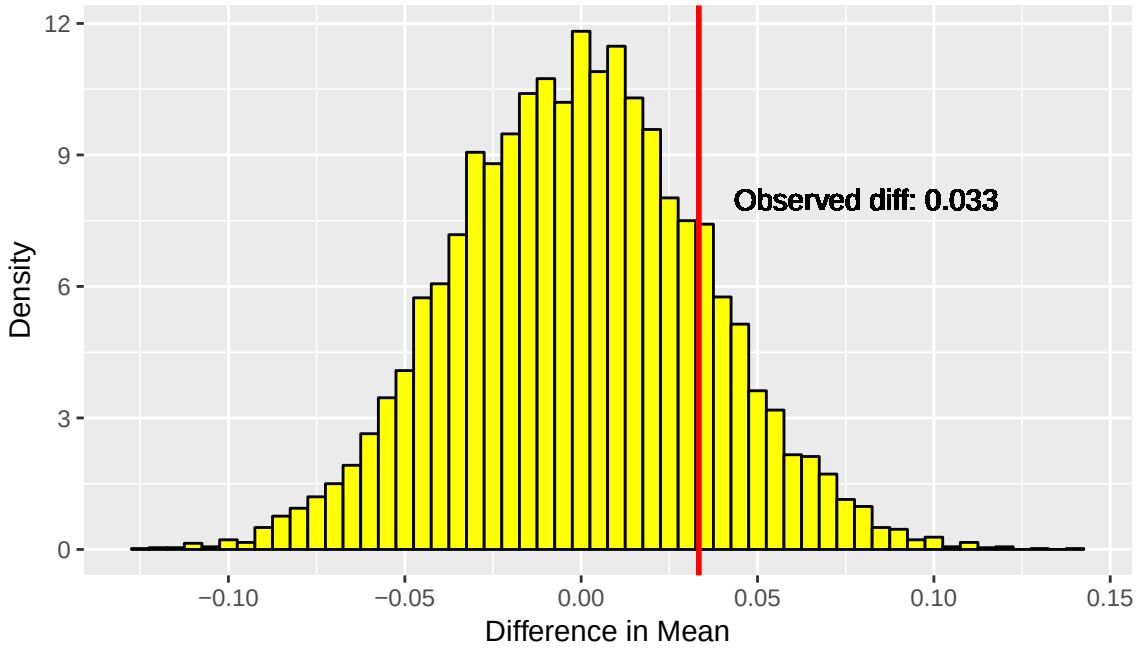

Figure 3: Permutation Test Differences

Conclusion: This test shows a very small degree of difference between these data sets, and the observed difference falls clearly within the range of differences across 10,000 permutations of these populations. There is nothing to imply the distributions of exponential sample means and the theoretical normal per the CLT, are different i.e. we accept the hypothesis that $\overline{X} \sim N($ 5 , 0.625 $)$

Appendix: R Code

```
library(knitr)
opts_chunk$set(fig.width=6, fig.height=3.5, fig.pos = "h", echo=FALSE, eval=TRUE)
library(ggplot2)
library(kableExtra)

#-------------------------------------------------------------------------------------
# Set up parameters for the simulation
set.seed(123456)        # Set RNG seed for reproducibility
lambda<-0.2             # Exponential rate parameter
num_exps<-40            # How big is the sample for each mean
num_sims<-1000          # How many samples/simulations to run

# Build a vector of length {num_sims}, for mean of different samples from exp dist
mns<-NULL
for (i in 1:num_sims) {mns <- c(mns,mean(rexp(num_exps,lambda)))}
mns_df <- data.frame(mns)             # Make a data frame for plotting
colnames(mns_df)<-"exp_mean"

# Calculate some statistics from the simulated dataset and its theoretical distribution
sim_mean<-mean(mns)
sim_sd<-sd(mns)
sim_var<-sim_sd^2
theor_mean<- as.numeric(1/lambda)                  # Theoretical mean per CLT
theor_var <- as.numeric(1/(num_exps*lambda^2))     # Theoretical variance per CLT
pct_mean <- 100*(sim_mean-theor_mean)/theor_mean
pct_var <- 100*(sim_var-theor_var)/theor_var

#-------------------------------------------------------------------------------------
# Histogram of the mean of the samples of exponentials, with vertical line for mean
g <- ggplot(mns_df,aes(x=exp_mean))
g <- g + geom_histogram(color="black",fill="yellow",binwidth=sim_sd/4)
g <- g + labs(x="Sample Mean",y="Count")
g <- g + geom_vline(xintercept=sim_mean,color="red",size=1)
g <- g + geom_text(aes(sim_mean+0.1,105,vjust=0,hjust=0,
                       label = paste0("Mean: ",round(sim_mean,3),
                                      " vs. Theoretical Mean: ",round(theor_mean,3))))
print(g)

#-------------------------------------------------------------------------------------
# Density plot of the mean of samples exponential
# vs. original exponential density & the CLT normal
norm_str<-paste0("N(",theor_mean,",",round(theor_var,3),")")
exp_str<-paste0("E(",lambda,")")
g <- ggplot(mns_df,aes(x=exp_mean))
g <- g + geom_density(aes(color=paste0(exp_str," Sample mean")))
g <- g + labs(x="x",y="Density")
g <- g + geom_vline(xintercept=sim_mean,color="red", size=1)
g <- g + geom_text(aes(sim_mean+0.1, 0.525, vjust=0, hjust=0,
                       label = paste0("Mean: ",round(sim_mean,3))))
g <- g + stat_function(fun=dnorm, geom = "line", args = list(mean = theor_mean,
                       sd = sqrt(theor_var)), size = 1, aes(colour=norm_str))
```

```r
g <- g + stat_function(fun=dexp, geom = "line", xlim=c(0,8), args = list(rate = lambda),
                       size = 1, aes(colour=exp_str))
g <- g + scale_colour_manual("Distributions",values = c("black","red","blue"))
print(g)

#-------------------------------------------------------------------------------
# Permutation Test.  # Set up the same size population drawn from the CLT Normal

# Draw {num_sims} random normals and combine with the original exponential means
norms<-rnorm(num_sims,mean=theor_mean,sd=sqrt(theor_var))
data<-c(mns,norms)

# Set up reference labels for the original data sets, "exp" then "norm"
labels<-c(rep("exp",num_sims),rep("norm",num_sims))

# Function to take the diff in means of a dataset, split by label
get_mean_diff<-function(dat,lab) {
        (mean(dat[lab == "exp"]) - mean(dat[lab == "norm"]))}

# Run a series of 10,000 simulations taking different permutations of the data labels
permStats<-data.frame(sapply(1:10000,function(i) get_mean_diff(data,sample(labels))))
colnames(permStats)<-c("mean_diff")

# Get the same statistic for the original dataset
observedStat <- get_mean_diff(data,labels)

# Print summary table of the means for the analysis
results<-data.frame(mean(mns),mean(norms),observedStat,stringsAsFactors = FALSE)
results %>% kable(col.names=c("Mean of Exp Sample Means","Mean of Normals","Observed Diff"),
                  align="r",digits=c(3,3,3),booktabs=T,
                  caption="Mean of Permutation Test Populations") %>%
                kable_styling(latex_options = "hold_position")

#-------------------------------------------------------------------------------
# Histogram of the differences in means from the permutation test
g <- ggplot(permStats,aes(x=mean_diff))
g <- g + geom_histogram(color="black",fill="yellow",aes(y=..density..),binwidth=0.005)
g <- g + labs(x="Difference in Mean",y="Density")
g <- g + geom_vline(xintercept=observedStat,color="red",size=1)
g <- g + geom_text(aes(observedStat+0.01,7.75,vjust=0,hjust=0,
                       label = paste0("Observed diff: ",round(observedStat,3))))
print(g)
```