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

Course 6 (Statistical Inference) / Assessment 1 Part 1

Paul Ringsted, 18th 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 (in Appendix) 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 and plot a histogram of the resulting distribution. We set the random number generator seed for reproducibility.

```
library(ggplot2)
```

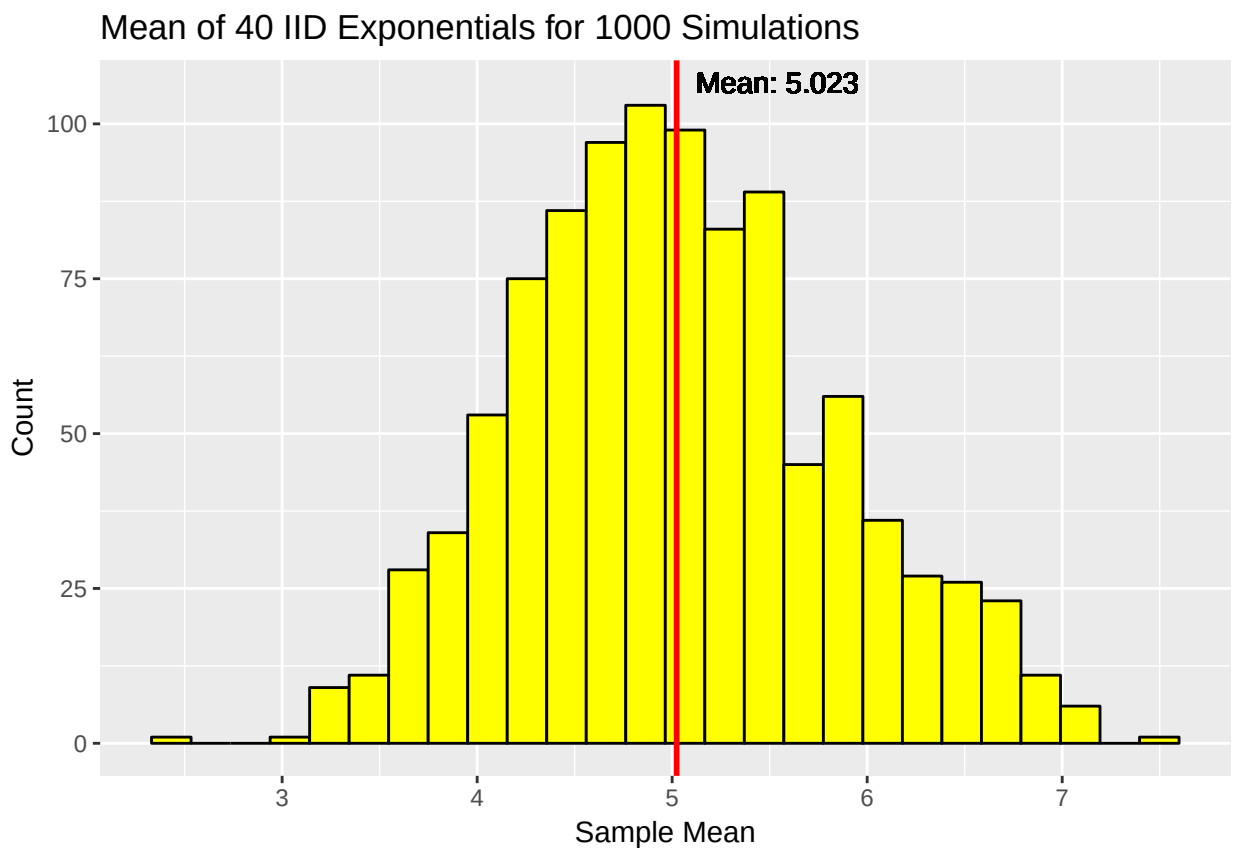
```
set.seed(123456)
lambda<-0.2
num_exps<-40            # How big is the sample for each mean
num_sims<-1000          # How many samples/simulations to run

# 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
c(sim_mean,theor_mean,sim_var,theor_var)
```

```
## [1] 5.0229151 5.0000000 0.6570391 0.6250000
```

```
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 + labs(title=paste0("Mean of ",num_exps," IID Exponentials for ",
                           num_sims," Simulations"))
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))))
print(g)
```



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

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

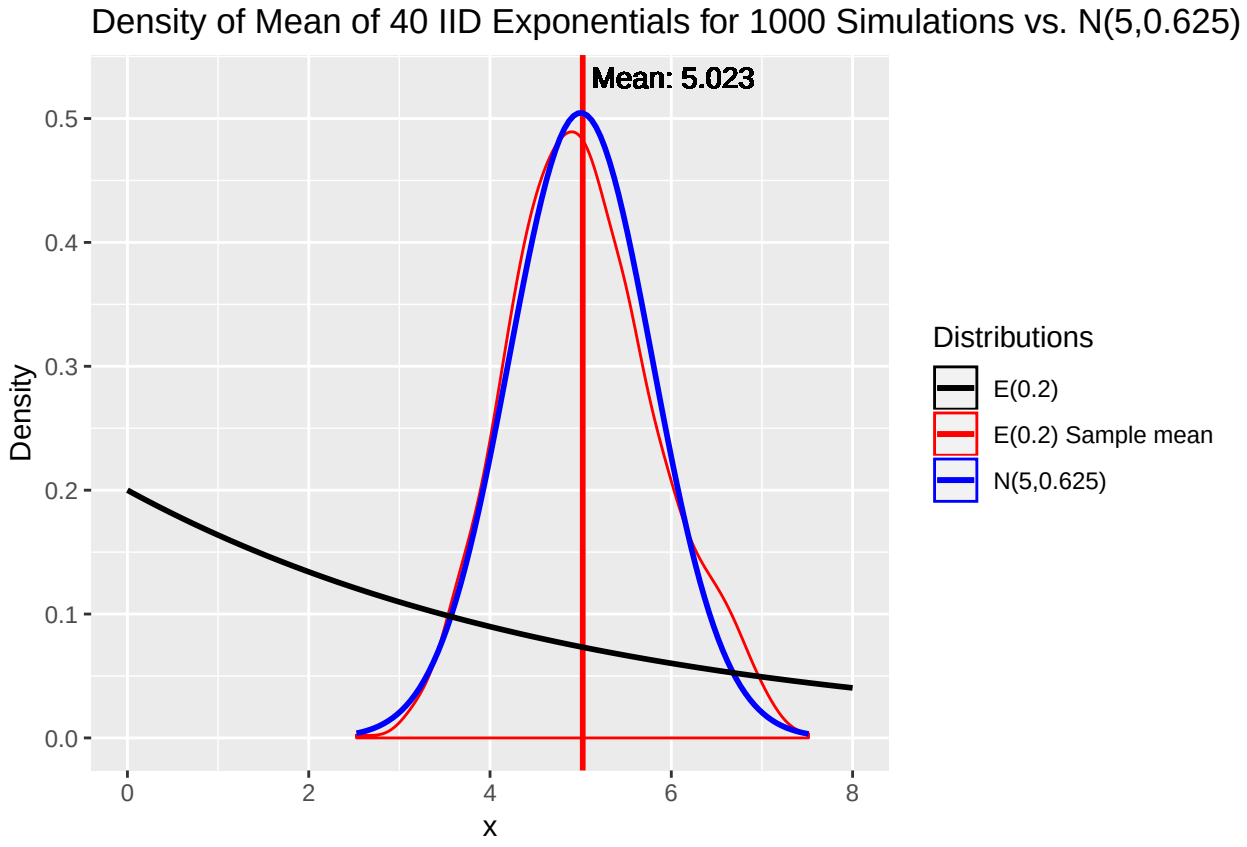
Summary of Mean and Variance:

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 initial plot of the histogram for this simulation certainly appears to fit a Gaussian model. We can replot the distribution of exponential sample means, as a density histogram and overlay the above theoretical Normal distribution per the CLT to get a clearer picture of how well this distribution fits to this Normal. We also plot this vs. the original expoential distribution which shows these are clearly different.

```
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 + labs(title=paste0("Density of Mean of ",num_exps," IID Exponentials for ",
                           num_sims," Simulations vs. ",norm_str))
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))
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)
```



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.

Appendix: 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, taking the difference in means, to see if there is any statistical difference between them. Our null hypothesis is

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

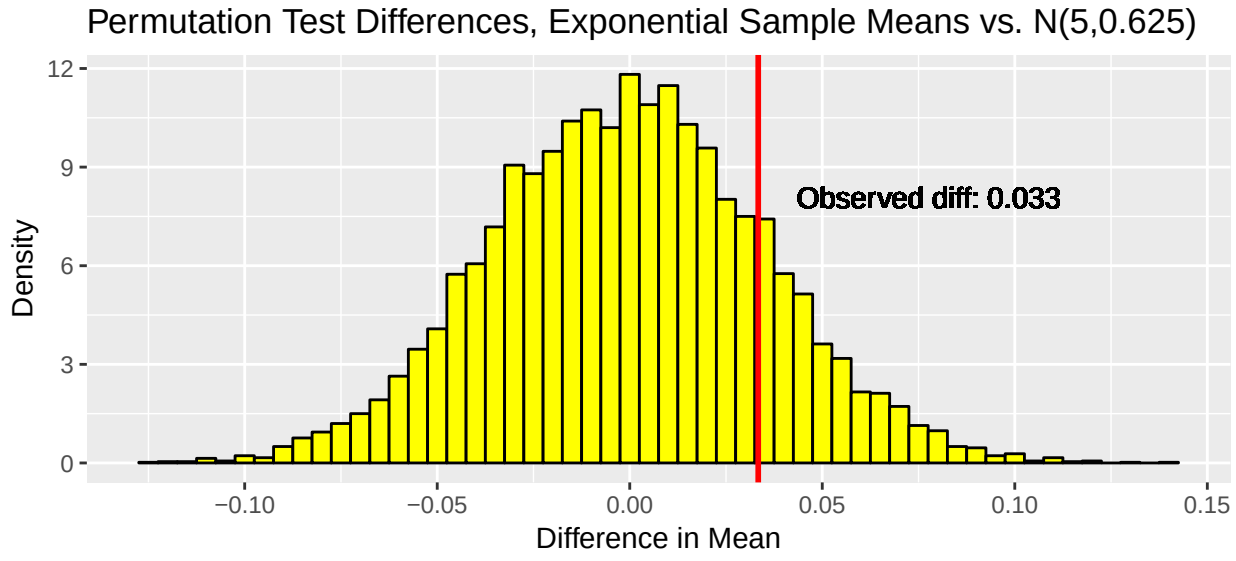
```
norms<-rnorm(num_sims,mean=theor_mean,sd=sqrt(theor_var))     # Draw {num_sims} normals
labels<-c(rep("exp",num_sims),rep("norm",num_sims))           # Set of labels for full data set
data<-c(mns,norms)                                            # Combine the exponential means and norms

# 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 of 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")
observedStat <- get_mean_diff(data,labels)      # get the stats for the original datasets
c(observedStat,sim_mean,mean(norms))
```

```
## [1] 0.03336976 5.02291512 4.98954536
```

```
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 + labs(title=paste0("Permutation Test Differences, Exponential Sample Means vs. ",norm_str))
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)
```



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