Exponential Distribution Course Project

mjfii

2012-01-20

Overview:

This course project report looks at a series of exponential distribution iterations, and compares the result set to a normal distribution and the theoretical mean and variance.

This report is for a Coursera Class project - Statistical Inference (Part 1). Per the project requirements, the below solution maintains the following:

> Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should:
> 1. Show the sample mean and compare it to the theoretical mean of the distribution.
> 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
> 3. Show that the distribution is approximately normal.

Simulations:

The simulation can be built in R using the `rexp(n,l)` function, where `n` is the number of exponentials per set, and `l`, or lambda is the rate (a constant of these purposes of 0.2). We need to build a single column data frame with the average of each of the sets. We iterate 1k times, or 1k simulations, or `s`, and use the `mean` function to calculate the average. The code is as follows:

```
l<-0.2          # lambda
n<-40           # number of exponentials
s<-1000         # number of simulations

# initialize a data frame with a row count of the number of simulations
df<-data.frame(mean=numeric(s))

# iterate 1 to the number of simulations variable
for (i in 1:s) {
  ss<-rexp(n,l)      # simulation set of n exponential with l lambda
  df[i,1]<-mean(ss) # mean of simulation set
}
```

Sample Mean versus Theoretical Mean:

With our sample data in place, we will compare the sample mean to the theoretical mean. The latter is defined as 1/lambda; we will define `tm` in this manner.

```
tm<-1/l             # theoretical mean
```

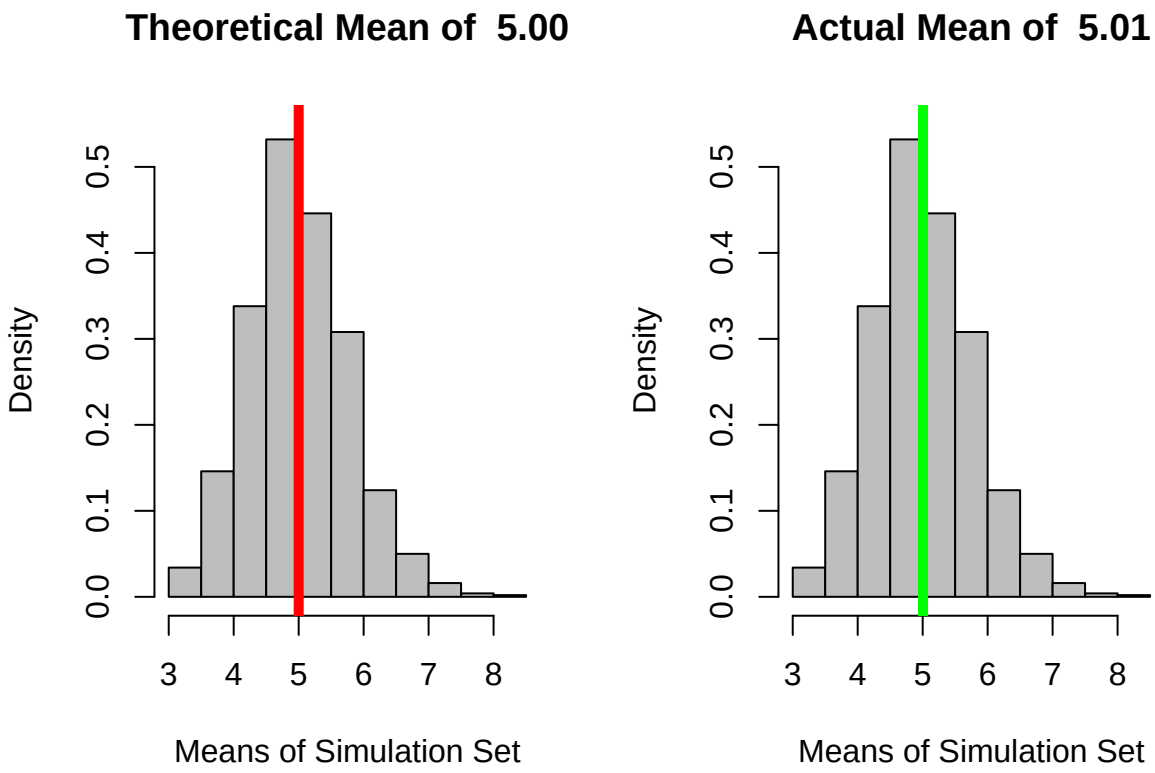
```
## [1] "5.00"
```

We can define the sample mean as the average of the interated set; `am` is defined in the following manner.

```
am<-mean(df$mean)    # actual mean
```

```
## [1] "5.01"
```

As you can see from the above two results, the sample mean (or actual mean) of 5.1 is very close to the theoretical mean of 5.0. When we plot the sample set below and place a vertical line on both mean results, the histograms look almost identical.



Sample Variance versus Theoretical Variance:

We can futher compare by defining the theoretical variance as 1/lambda squared, divided by the number of exponential observations, or `n`. We will define `tv` in this way as seen below.

```
tv<-((1/l)^2)/n      # theoretical variance
```

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

We can use the R `var` function to determine the actual variance of the observations. We can define `av` as follows.

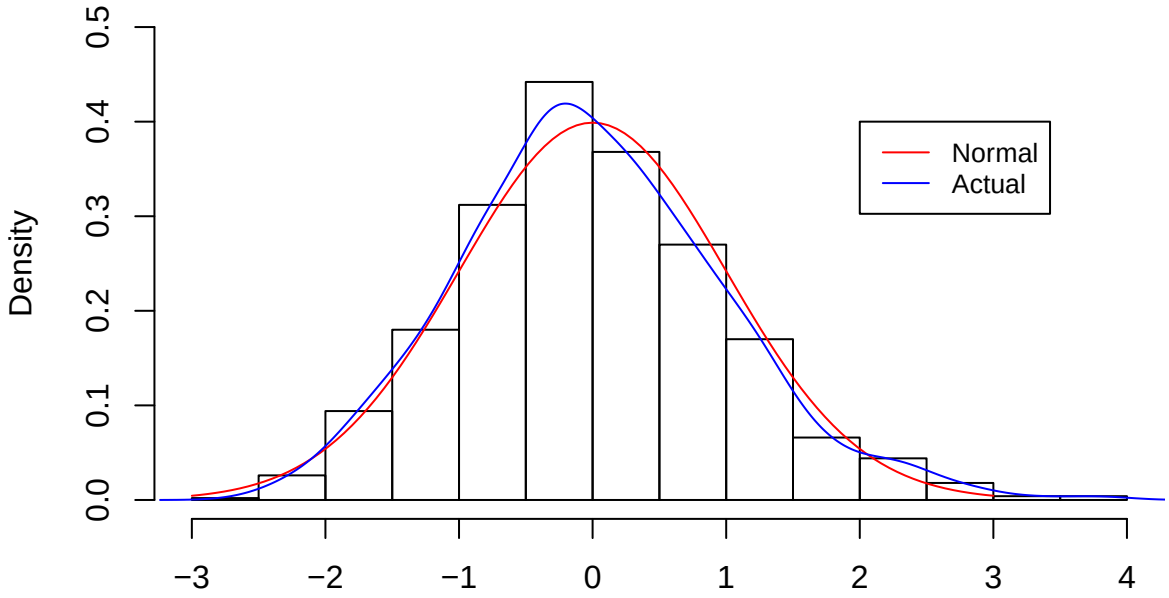
```
av<-var(df$mean)     # actual variance
```

```
## [1] "0.601"
```

The results of the theoretical variance as compared to actual variance are very close, as noted above. So, the sample data is 'spread-out' in a similar fashion to what is expected.

Distribution:

Finally, if we plotted the a normal curve and overlayed to the sample set along with the actual distribution curve, we can see very clearly that they are very similar; in turn, the sample set can be defined as normal, in terms of distribution.



Appendix:

A complete version of the R code is provided below, and the source code can be found here: http://github.com/mjfii/Statistical-Inference.

```
set.seed(123456789);

l<-0.2          # lambda
n<-40           # number of exponentials
s<-1000         # number of simulations

# initialize a data frame with a row count of the number of simulations
df<-data.frame(mean=numeric(s))

# iterate 1 to the number of simulations variable
for (i in 1:s) {
  ss<-rexp(n,l)      # simulation set of n exponential with l lambda
  df[i,1]<-mean(ss) # mean of simulation set
}
```

```r
# calculate the means
tm<-1/l             # theoretical mean
am<-mean(df$mean)   # actual mean

# plot the means
par(mfrow=c(1,2))
hist(df$mean,probability=T,main=paste('Theoretical Mean of ',format(round(tm,2),nsmall=2)),ylim=c(0,0.5
abline(v=tm,col='red',lwd=5)
hist(df$mean,probability=T,main=paste('Actual Mean of ',format(round(am,2),nsmall=2)),ylim=c(0,0.55),col
abline(v=am,col='green',lwd=5)

# calulate the variances
tv<-((1/l)^2)/n     # theoretical variance
av<-var(df$mean)

# add distribution curve
par(mfrow=c(1,1))
hist(scale(df$mean),probability=T,main='',ylim=c(0,0.5),xlab='')
curve(dnorm(x,0,1),-3,3, col='red',add=T) # normal distribution
lines(density(scale(df$mean)),col='blue') # actual distribution
legend(2,0.4,c('Normal','Actual'),cex=0.8,col=c('red','blue'),lty=1)
```