Simulation and Comparison against Normal Distributions

thecapacity

January 24, 2015

Overview

This is the project file for my Peer Assignment of the Statistical Inference class.

In this assignment, I will use simulation to explore inference and do some simple inferential data analysis.

The project consists of two parts:

1. Data Simulation and results.
2. Distribution inferential analysis.

I will create a report to answer the questions. Given the nature of the series, ideally knitr will be used to create the reports and convert to a pdf. The pdf report will be **no more than 6 pages total including supporting material** if needed (code, figures, etcetera).

This will also be published at: http://rpubs.com/thecapacity/StatInf_Proj1 (http://rpubs.com/thecapacity/StatInf_Proj1)

I will:

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; focusing on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of exponentials.

Approach

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem.

Some informational elements used are:

- The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter.
- The defined **mean** of exponential distribution is `1/lambda` and
- The defined **standard deviation** is also `1/lambda`, which means the **variance** will be `(1/lambda)^2`
- I will set `lambda = 0.2` for all of the simulations.
- I will investigate the distribution of averages of **40 exponentials**, i.e. `n = 40`.
- Note that I will do **one thousand** simulations, i.e. `iterations = 1000`.

Results

The following illustrates, via simulation and associated explanatory text, the properties of the exponentials outline above:

Simulations

```
set.seed(1357) # for reproducable results

# Given Variables, from assignment (see above)
lambda <- .02
n <- 40
iterations <- 1000

theoretical_mean <- 1/lambda
theoretical_var <- (1/lambda)^2

results <- matrix( rexp(n*iterations, lambda), nrow=iterations )
# Should result in 1000x40 matrix, with each row rexp() for n=40
dim(results)
```

```
## [1] 1000   40
```

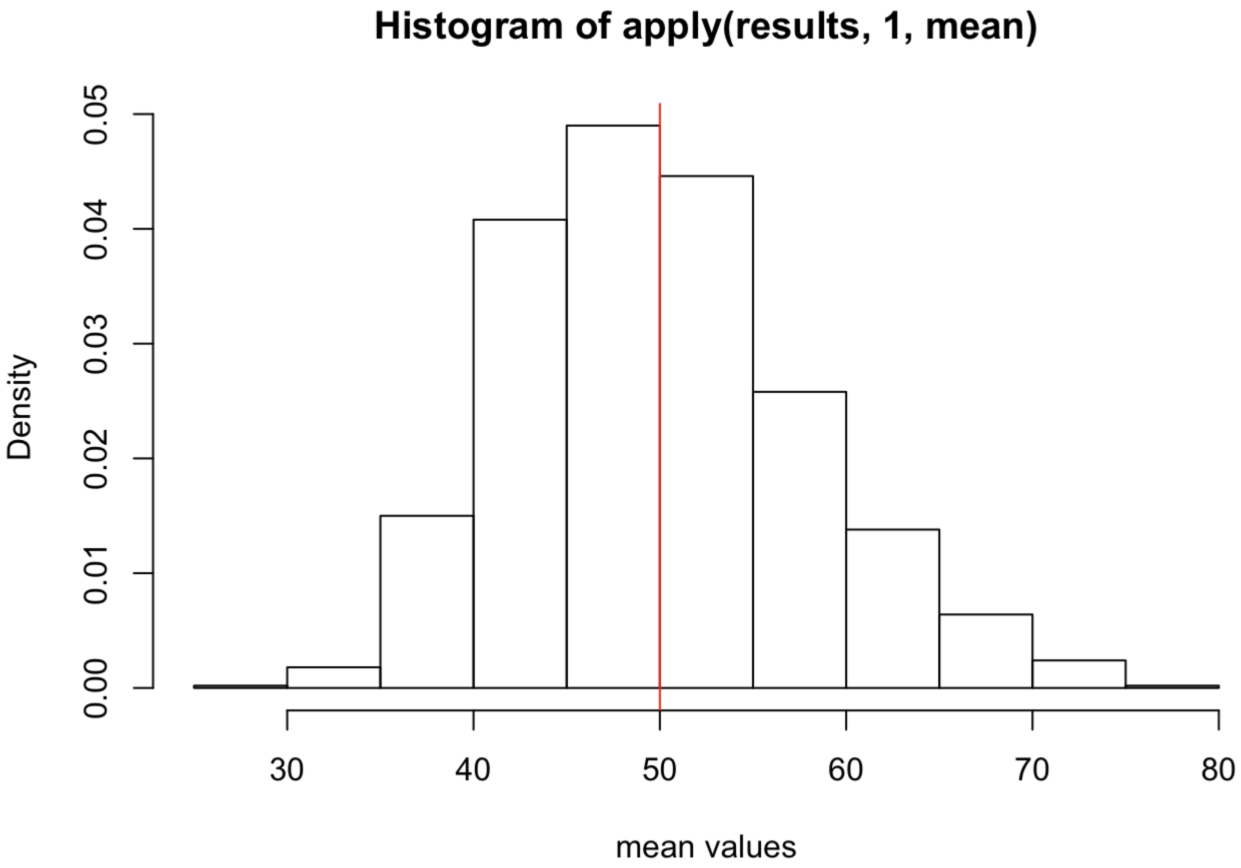
Mean

- The **sample mean** is `50.0124554`.
- The **theoretical mean** (given in the introduction above as `1/lambda`) is: `50`.

In our simulation, the sample differs from the theoretical value by only `0.0124554`, which is to be expected given the relatively high values of `n` and (especially) `iterations`.

A graph of the sample results, relative to the theorical mean is shown in the figure below:

```
hist(apply(results,1,mean), xlab="mean values", prob=TRUE)
abline(v=theoretical_mean, col="red")
```



The distribution of averages for `1000` *iterations of* `40` *exponential random variables with lambda =* `0.02`*.*
The red vertical line indicates the theoretical mean.

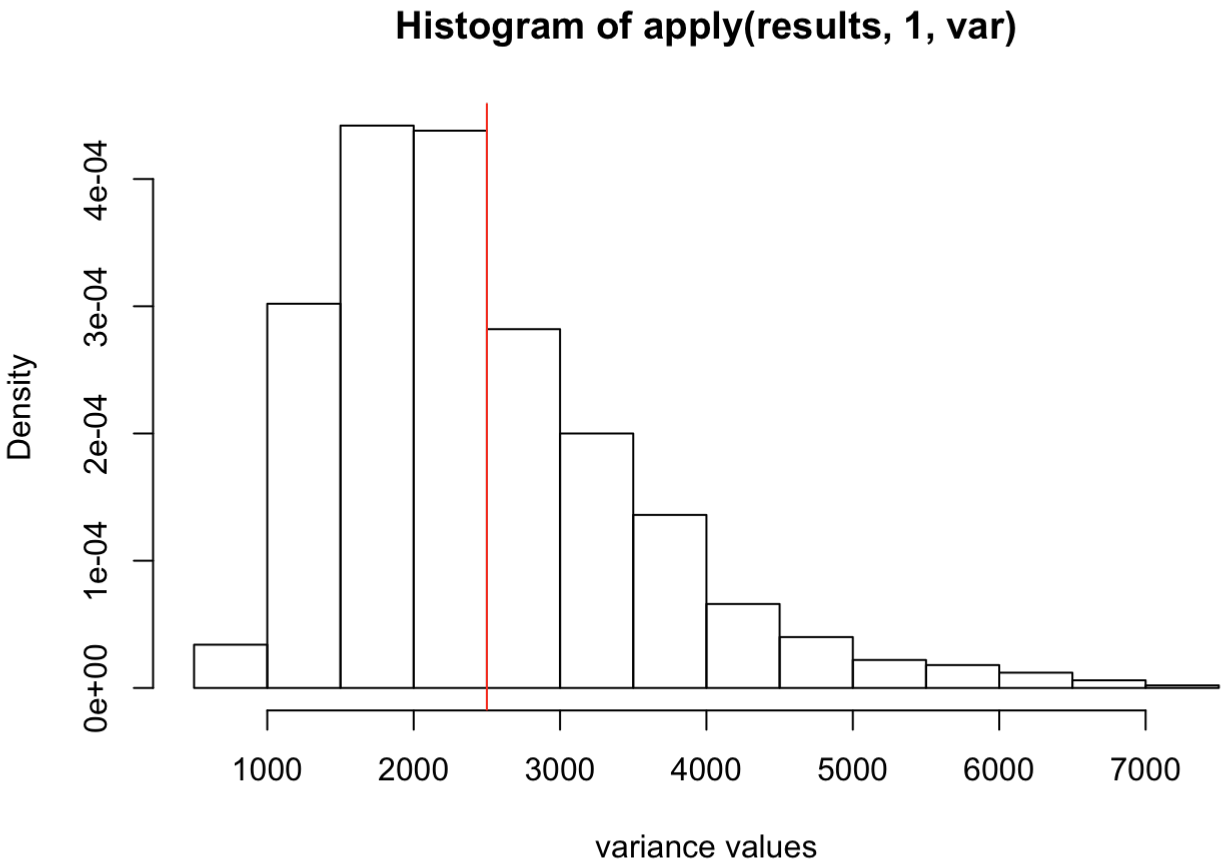
Variance

- The **sample variance** is `2454.7027252`.
- And the **theoretical variance** (predefined above) is `2500`.

In our simulation, the sample differs from the theoretical value by `12.2347594`. This similarity was largly expected, given the values used in the simulations that help converge to the expected values.

A graph of the sample results, relative to the theorical variance is shown in the figure below:

```
hist(apply(results,1,var), xlab="variance values", prob=TRUE)
abline(v=theoretical_var, col="red")
```



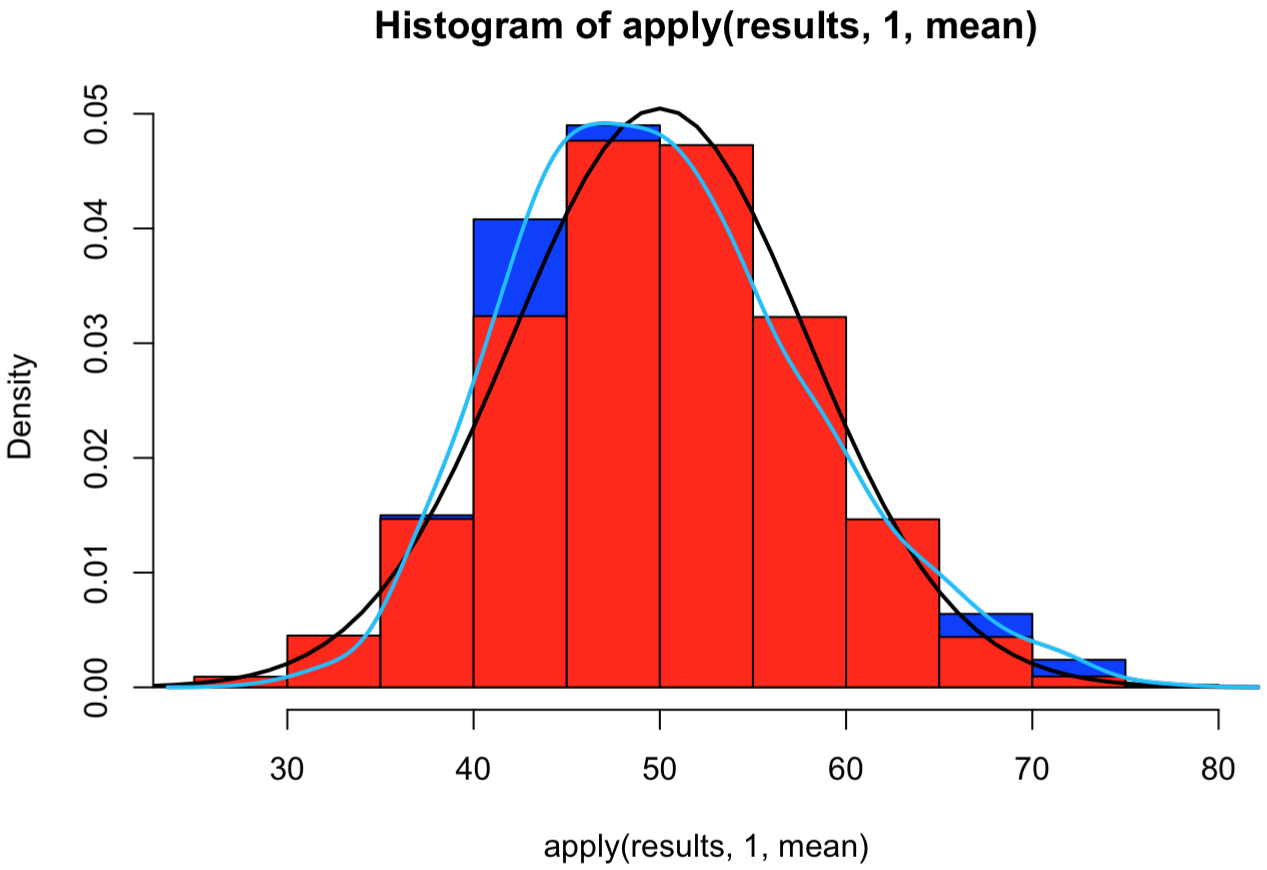
The distribution of variance for `1000` iterations of `40` exponential random variables with lambda = `0.02`. The red vertical line indicates the theoretical variance.

Distribution

Here we show that the distribution is approximately normal:

```
hist( apply(results,1,mean), prob=TRUE, col="blue")
hist(rnorm(results, mean=1/lambda, sd=(1/lambda)/sqrt(n)), col="red", prob=TRUE, add=TRUE)

dx <- seq(min(results), max(results))
dy <- dnorm(dx, mean=1/lambda, sd=(1/lambda)/sqrt(n))
lines(dx, dy, pch=22, col="black", lwd=2)
lines(density(apply(results,1,mean)), col="deepskyblue", lwd=2)
```



The distribution of averages in blue for our sample iteratations, with the density line in a lighter blue, is graphed against the random normal distribution in red, with the black line representing the density of our normal distribution.

Here we can see large similarity between the emperical and theoretical datasets, largely due to the significant number of iterations used.