Statistical Inference - Course Project pt 1

Richard Hardy

2019-12-15

Part 1 - Simulation Exercise

Overview

This exercise is based on the exponential distribution, demonstrating an application of the Central Limit Theorem. The mean and variance of 1,000 means of 40 random draws from the exponential distribution, is shown to approximate the theoretical population mean and variance given by the exponential distribution pdf.

With $X \sim Exp(\lambda)$:

- $E[X] = 1/\lambda$
- $Var[X] = 1/\lambda^2$.

Simulations

Let: x_i be the ith draw of an iid random variable X where:

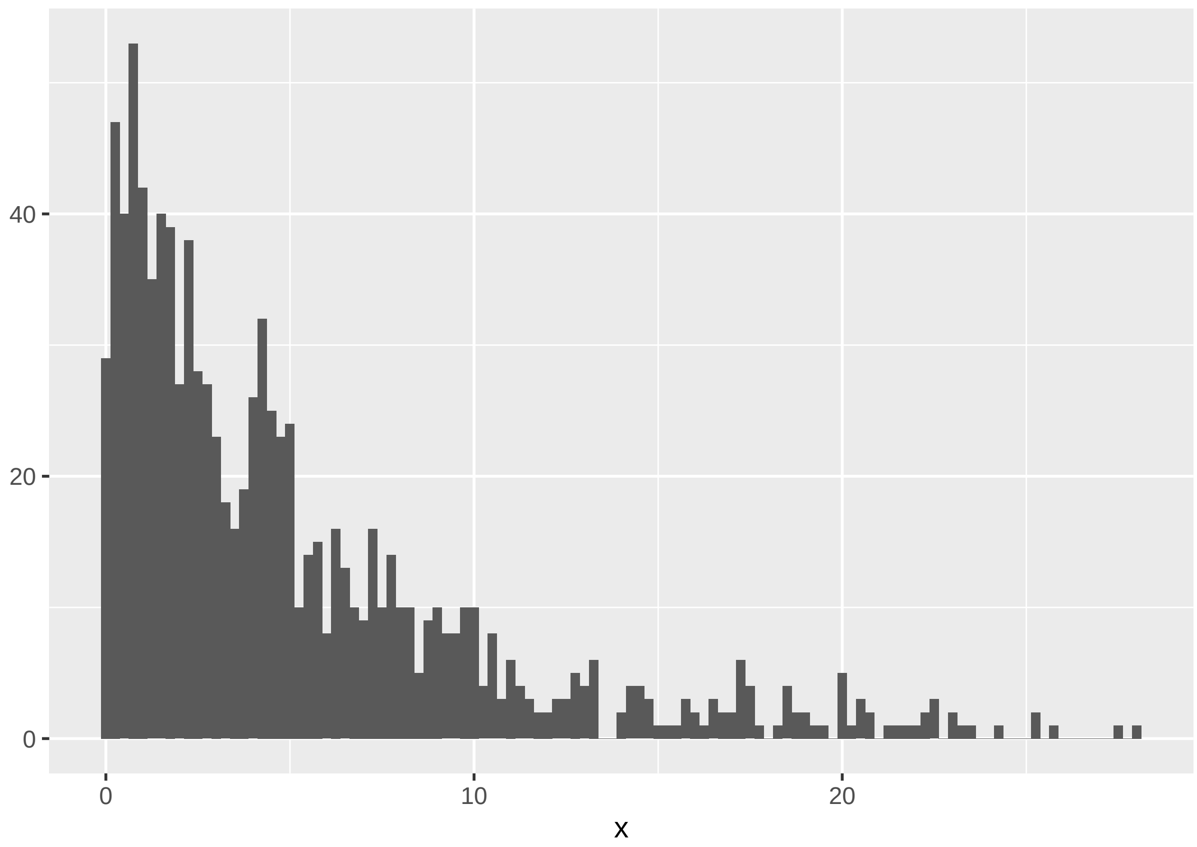
$X \sim Exp(\lambda)$

Simulating 1,000 random draws from the exponential function:

With $\lambda = 0.2$, and $n_x = 1,000$ a histogram, mean and variance is given by the following code:

```
set.seed(56789)
x <- rexp(1000,0.2)

library(ggplot2)
qplot(x, geom = "histogram", binwidth = 0.25)
```



```
xmn <- mean(x)
xvar <- var(x)
```

This represents 1,000 randomly generated values, drawn with the probability density function (pdf) given by:

$f_x(x) = 1/\lambda \times e^{-x/\lambda}, x > 0, \lambda > 0$

This set of 1,000 random draws has mean 5.14 and variance 26.98.

Simulating 1,000 means and variances of random draws of size 40

1,000 Means and variances of 40 exponentials with $\lambda = 0.2$, is given by:

```
set.seed(56789)

mns = NULL
vars = NULL
for (i in 1 : 1000) {
        mns = c(mns, mean(rexp(40,0.2)))
        vars = c(vars, var(rexp(40,0.2)))
}
```

The expected values are saved as follows:

```
expmn <- mean(mns)
expvar <- mean(vars)
```

Sample Mean versus Theoretical Mean

The theoretical mean of means of 1,000 simulations (a large number) of 40 exponentials is an approximation of the sample mean of an iid random variable $X \sim Exp(\lambda)$, representing a large collection of iid random exponentials. This is consistent with the Law of Large Numbers and the Central Limit Theorem.

This is given by $1/\lambda$.

With $\lambda = 0.2$ the theoretical mean is given as:

```
1 / 0.2
```

```
## [1] 5
```

From the sample of 1,000 simulations of 40 draws, the observed mean is:

```
expmn
```

```
## [1] 4.985121
```

From the single sample of 1,000 draws from the pdf, the observed mean is:

```
xmn
```

```
## [1] 5.137991
```

The expected sample mean of 1,000 simulations of 40 draws: 4.99 is shown to be a close approximation of the theoretical mean 5.

1,000 draws from the exponential pdf gives 5.14.

Sample Variance versus Theoretical Variance

Similarly, the expected variance of 1,000 simulations (a large number) of 40 exponentials is an approximation of the sample variance of an iid random variable $X \sim Exp(\lambda)$, representing a large collection of iid random exponentials. This is consistent with the Law of Large Numbers and the Central Limit Theorem.

The expected variance of 1,000 simulations of 40 exponentials is given by $1/\lambda^2$

```
1 / (0.2^2)
```

```
## [1] 25
```

From the sample of 1,000 simulations of 40 exponentials, the variance of the mean is:

```
expvar
```

```
## [1] 24.87157
```

From the sample of 1,000 draws from the pdf, the variance of X is:

```
xvar
```

```
## [1] 26.979
```

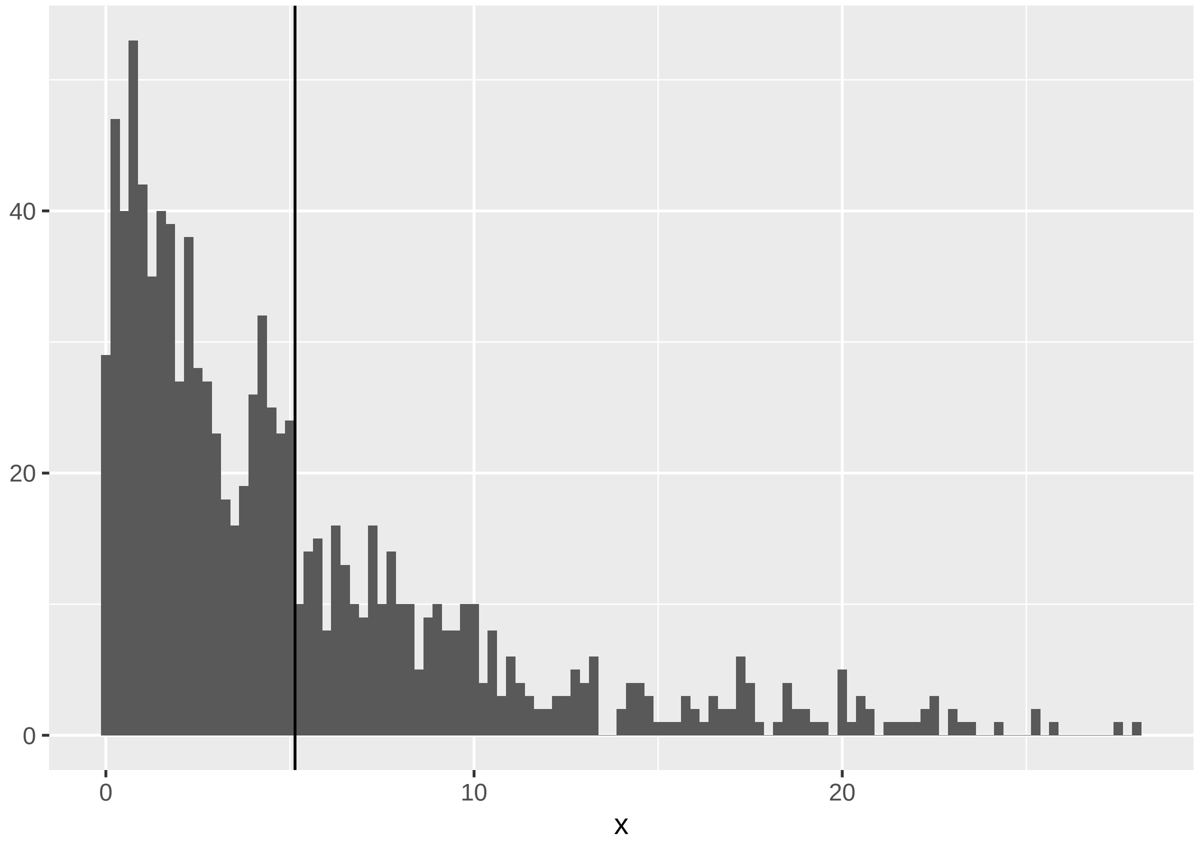
The expected sample variance of 1,000 simulations of 40 draws 24.87 is shown to be a close approximation of the theoretical variance 25.

1,000 draws from the exponential pdf has variance 26.98.

Distributions

The distribution of 1,000 randomly drawn exponentials is again, given by:

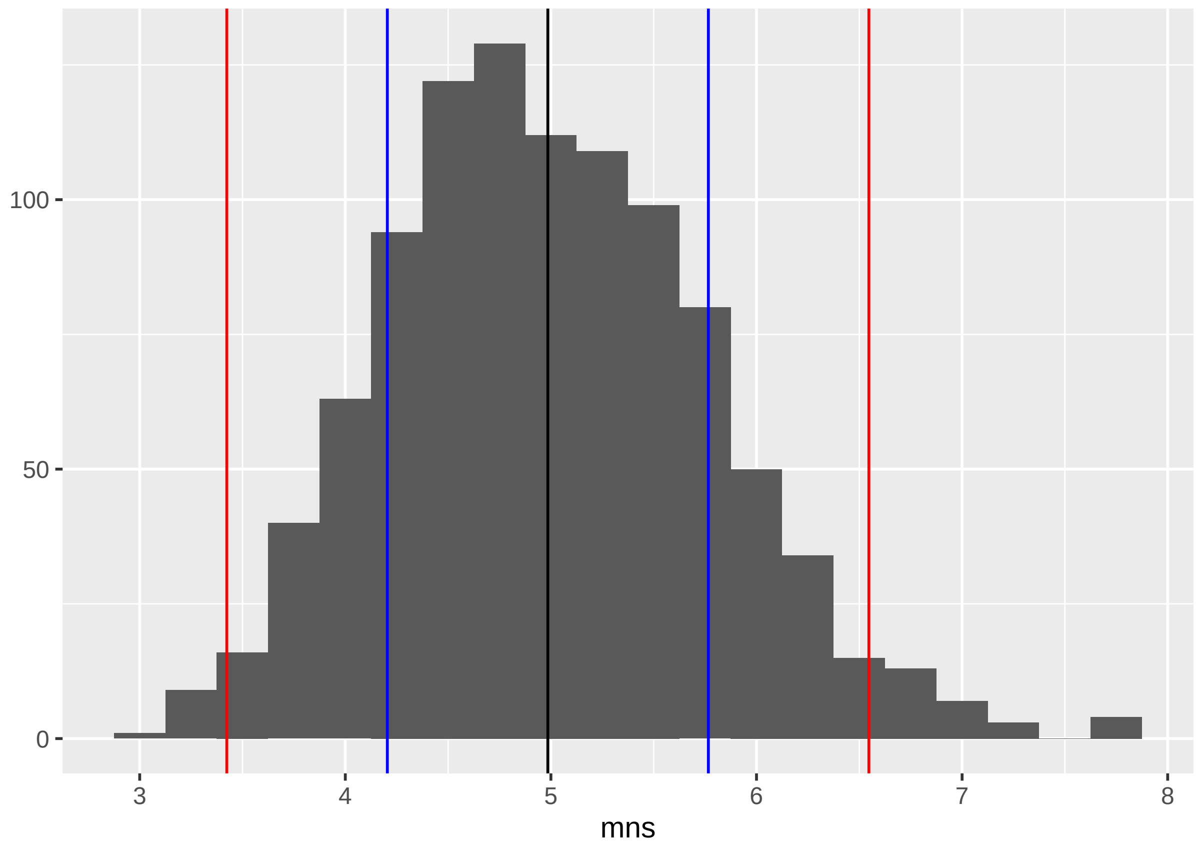
```
l <- qplot(x, geom = "histogram", binwidth = 0.25)
l  + geom_vline(xintercept = xmn)
```



The 1,000 randomly drawn values are distributed according to the exponential pdf around a mean of 5.14.

The distribution of the mean of 1,000 simulations of 40 randomly drawn exponentials is given as:

```
q <- qplot(mns, geom = "histogram", binwidth = 0.25)
q + geom_vline(xintercept = expmn) +
  geom_vline(xintercept = expmn + sd(mns), col = "blue") +
  geom_vline(xintercept = expmn - sd(mns), col = "blue") +
  geom_vline(xintercept = expmn +2*sd(mns), col = "red") +
  geom_vline(xintercept = expmn -2*sd(mns), col = "red")
```



The means of 1,000 simulations of size 40 are approximately normally distributed around an expected value of 4.99. Blue and red lines drawn at 1 and 2 standard deviations above and below the expected value illustrate that approximately 68% and 95% of the 1,000 observations of means of 40 draws will lie, respectively, within 1 and 2 standard deviations of the mean of the 1,000 means of 40 draws.

Taken together, these plots show the CLT in action. Strictly speaking, as the number of simulations approaches infinity, the expected value of the simulated means of 40 draws will approach the expected value of a large collection of draws from the underlying exponential probability density function: i.e, will approach $1/\lambda$

Slight deviations from normality might be observed at this simulation sample size given the extreme right skew of the exponential distribution. This is evident in the slightly wider right tail in the plot of the means of the 40 draws.