A Simple Statistical Exploration of the Exponential Distribution

Wayne Witzke

Synopsis

This analysis explores the relationship between theoretical and sample statistics for the exponential distribution. It simulates sets of experimental samples drawn from the exponential distribution, and calculates statistics for those sets. Finally, it shows that the distribution of sample means of those sets is Gaussian, and thus the Central Limit Theorem applies.

Preliminaries

This section prepares the analysis to run. For more information on replicating this analysis, see the appendix.

```
knitr::opts_chunk$set( echo = TRUE );
library(ggplot2);
set.seed(397413653);
```

Simulation

This analysis generates 1000 sets of 40 random numbers taken from the exponential distribution. The chosen rate (λ) is 0.2, giving the distribution a theoretical mean and standard deviation of $\mu = \sigma = \frac{1}{\lambda} = 5$.

```
sim.matrix = matrix( rexp( 40000, 0.2 ), 1000, 40 );
raw.mean = mean( as.vector( sim.matrix ) );
raw.sd = sd( as.vector( sim.matrix ) );
```

Note that, over all generated values, the mean is $\bar{x} = 4.9904507$ and the standard deviation is $s = 4.9631457$.

Results

Sample Mean vs. Theoretical Mean

```
sim.means = apply( sim.matrix, 1, mean );
sim.mean = mean( sim.means );
```

Once again, the theoretical mean of the distribution is given by $\mu = \frac{1}{\lambda} = 5$. Taking the average of the means of each simulation, we see that the simulated sample mean is $\bar{X} = 4.9904507$.

Sample Variance vs. Theoretical Variance

```
sim.variance = var( sim.means );
sim.sd = sd( sim.means );
```

The theoretical variance of the distribution is given by $\sigma^2 = (\frac{1}{\lambda})^2 = 25$. The variance of the sample mean, however, is a much smaller $s_m^2 = 0.6565609$. Note that there was no expectation that the sample mean variance and theoretical variance would be similar, as they describe different distributions. However, note that $s_m^2 \approx \frac{s^2}{n} = 0.6158204 \approx \frac{\sigma^2}{n} = 0.625$, where $n = 40$ is the sample size, $\frac{\sigma}{\sqrt{n}}$ is the true standard error, and $\frac{s}{\sqrt{n}}$ is the sample standard error.

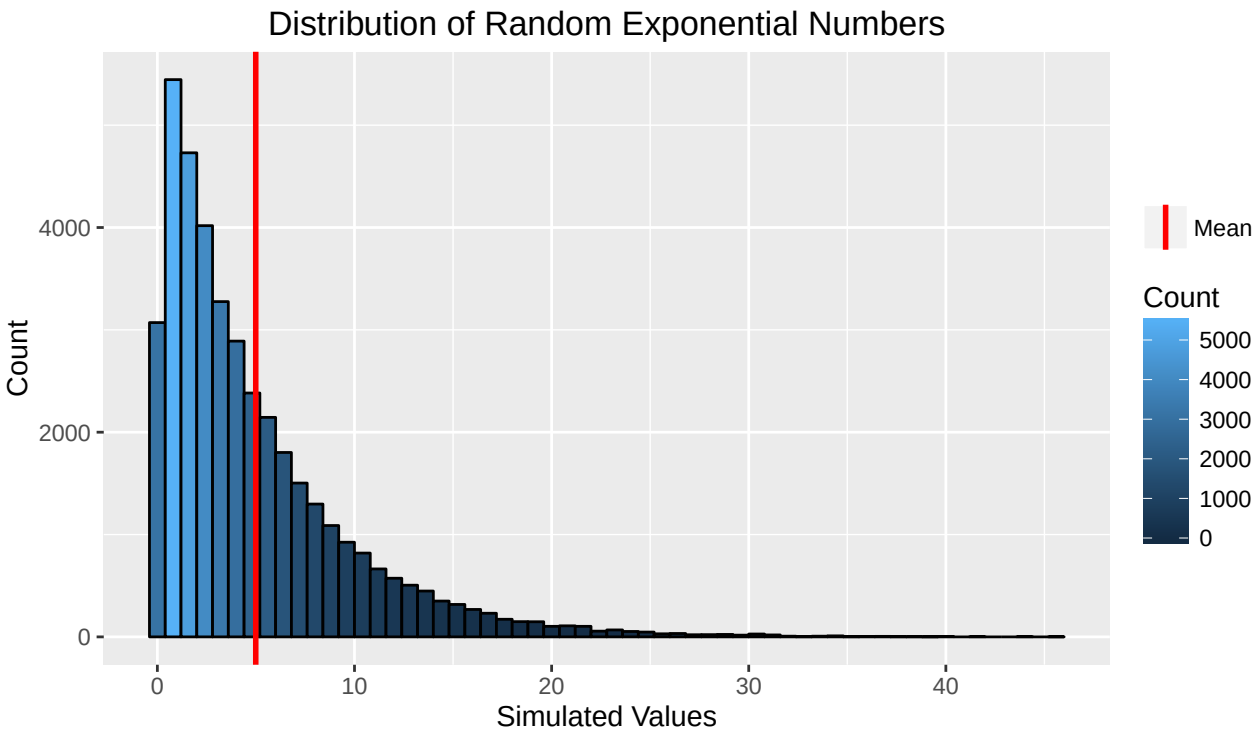
The statistics for the theoretical distribution and sample means are summarized in the following table.

Theoretical Distribution vs. Simulated Sample Mean Statistics		
	Mean	Variance
Theoretical Exponential Distribution	$\frac{1}{\lambda} = 5$	$\left(\frac{1}{\lambda}\right)^2 = 25$
Simulated Sample Means	4.9904507	0.6565609

Distribution

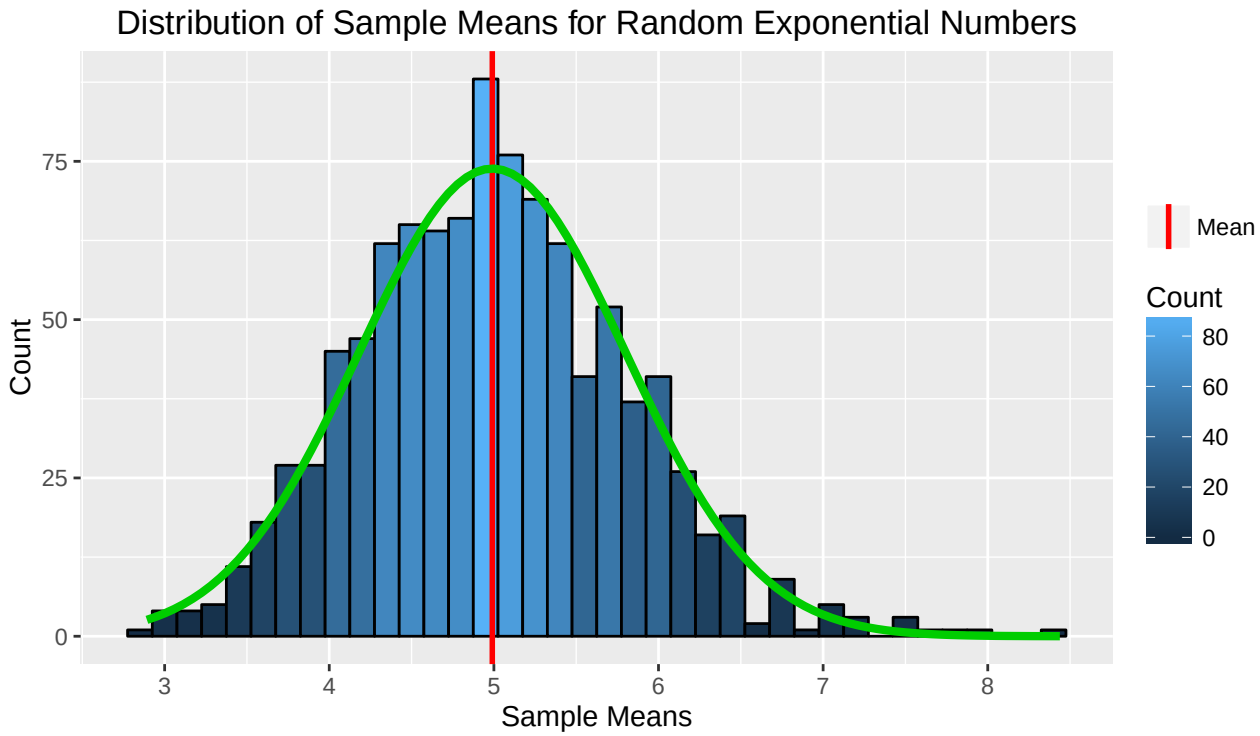
As can be seen from the following histrogram, random values drawn from the exponential distribution do not have a normal distribution. For example, the mean does not occur at the peak value.

```
raw.histogram =
    ggplot( show.legend = TRUE ) + aes( as.vector( sim.matrix ) ) +
    geom_histogram( binwidth = 0.8, col = "black", aes( fill = ..count.. ) ) +
    xlab( "Simulated Values" ) + ylab("Count") +
    ggtitle( "Distribution of Random Exponential Numbers" ) +
    geom_vline( aes( xintercept = raw.mean, color = "Mean" ), size = 1 ) +
    scale_color_manual( name = "", values = c( Mean = "red" ) ) +
    scale_fill_continuous( name = "Count" );
print( raw.histogram );
```



However, the distribution of the sample means calculated from the individual simulations *is* normal. The following histogram shows the distribution of the sample means and fits a normal curve with the same statistics, clearly showing that the sample means follows a Gaussian distribution.

```
sim.histogram =
    ggplot( show.legend = TRUE ) + aes( sim.means ) +
    geom_histogram( binwidth = 0.15, col = "black", aes( fill = ..count.. ) ) +
    xlab( "Sample Means" ) + ylab("Count") +
    ggtitle(
        "Distribution of Sample Means for Random Exponential Numbers"
    ) +
    geom_vline( aes( xintercept = sim.mean, color = "Mean" ), size = 1 ) +
    scale_color_manual( name = "", values = c( Mean = "red" ) ) +
    scale_fill_continuous( name = "Count" ) +
    stat_function(
        fun = function(x, mean, sd, n, bw) dnorm(x=x,mean=mean,sd=sd)*n*bw,
        color = "green3",
        size = 1.5,
        args = list( mean = sim.mean, sd = sim.sd, n = 1000, bw = 0.15 )
    );
print( sim.histogram );
```



The Central Limit Theorem states that

> The distribution of averages of independent and identically distributed variables becomes that of a standard normal as the sample size increases.

In this case, despite the underlying shape of the exponential distribution, the distribution of the sample means clearly approaches a normal distribution, centered about the mean of the exponential distribution, as predicted by the Central Limit Theorem. Note that the variance on the sample mean is controlled by the sample size, $n = 40$.

Appendix

System Information

This analysis was performed using the hardware and software specified in this section.

```
sessionInfo();
```

```
## R version 3.2.4 Revised (2016-03-16 r70336)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 15.10
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_2.1.0   rmarkdown_0.9.5
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.4        digest_0.6.9       plyr_1.8.3
##  [4] grid_3.2.4         gtable_0.2.0       formatR_1.3
##  [7] magrittr_1.5       evaluate_0.8.3     scales_0.4.0
## [10] stringi_1.0-1      labeling_0.3       RColorBrewer_1.1-2
## [13] tools_3.2.4        stringr_1.0.0      munsell_0.4.3
## [16] yaml_2.1.13        colorspace_1.2-6   htmltools_0.3.5
## [19] knitr_1.12.3
```