The Distribution of Means of Samples from an Exponential Population Distribution

Jonathan Owen

Saturday, May 23, 2015

Introduction

A practical interpretation of the *central limit theorem* is that for large sample sizes the means of independent identically-distributed random variables are approximately normally distributed regardless of the underlying distribution. The R environment includes functions for generating random variables from many distributions, which will be used here to explore the distribution of means of a large number of samples from an exponential population. The mean and variance of the samples will be compared with theoretical values obtained by application of the central limit theorem.

Simulation

An exponential distribution of variable x has a density function $f(x)$

$$f(x) = \lambda e^{-\lambda x}$$

where mean $\mu = \frac{1}{\lambda}$ and standard deviation $\sigma = \frac{1}{\lambda}$ so that the exponential distribution is described by λ. For the current exploration, $\lambda = 0.2$ and this is assigned in R as variable `lambda` to be used in the simulation.

```r
lambda <- 0.2
```

The approximation of a sample mean distribution to a normal distribution becomes better for larger sample sizes and more repetitions of the sampling. The central limit theorem will be tested with a sample size (`sample_size`) of 40 and 1000 replicates (`replicates`) of 1000.

```r
replicates <- 1000
sample_size <- 40
```

The `rexp` function is used to create `replicates`×`sample_size` random variables, which are organized into a matrix `exp_dist` of `replicates` rows by `sample_size` columns. A seed is used to allow the random sampling to be reproduced by others.

```r
set.seed(51815)
exp_dist <- matrix(rexp(replicates * sample_size, rate = lambda), nrow = replicates)
```

The means of each sample of 40 numbers can be calculated by repeatedly applying the `mean` function to each row of the matrix `exp_dist`.

```r
means_dist <- apply(exp_dist, 1, mean)
```

In order to use the ggplot2 package for plotting the data need to be typed as a data frame. A column `distribution` indicating whether the data are samples from the exponential distribution or distribution of sample means is also added to help with the plots.

```
exp_df <- data.frame(distribution = c(rep("exponential", 40000),
                                       rep("sample means", 1000)),
                                 X = c(c(exp_dist), c(means_dist)))
```

Comparison of Sample and Theoretical Properties

The means and variances of the exponential and sample means distribution are summarized in the table below.

Summary of simulated and theoretical mean and variance for exponential and sample means distributions

	simulated	theoretical
exponential distribution, mean	5.027	5.000
exponential distribution, variance	25.075	25.000
sample mean distribution, mean	5.027	5.000
sample mean distribution, variance	0.670	0.625

Theoretical values for the exponential distribution are based on the property that the mean $\mu = \frac{1}{\lambda}$ and standard deviation $\sigma = \frac{1}{\lambda}$. Variance is the square of the standard deviation.

The central limit theorem is applied so that the sample means distribution is treated as being normal with the mean $E(\bar{X})$ being equal to that of the underlying, in this case exponential, population. The variance S^2 of the sample means distribution is related to that of the exponential population by

$$S^2 = \frac{\sigma^2}{n}$$

Using the properties above the sample mean and variance can be compared to their theoretical values.

How does the sample mean compare to the theoretical mean of the distribution?
The tabulated mean of the simulated variables is 5.027, which differs by 0.5% from the expected theoretical mean of exactly 5.

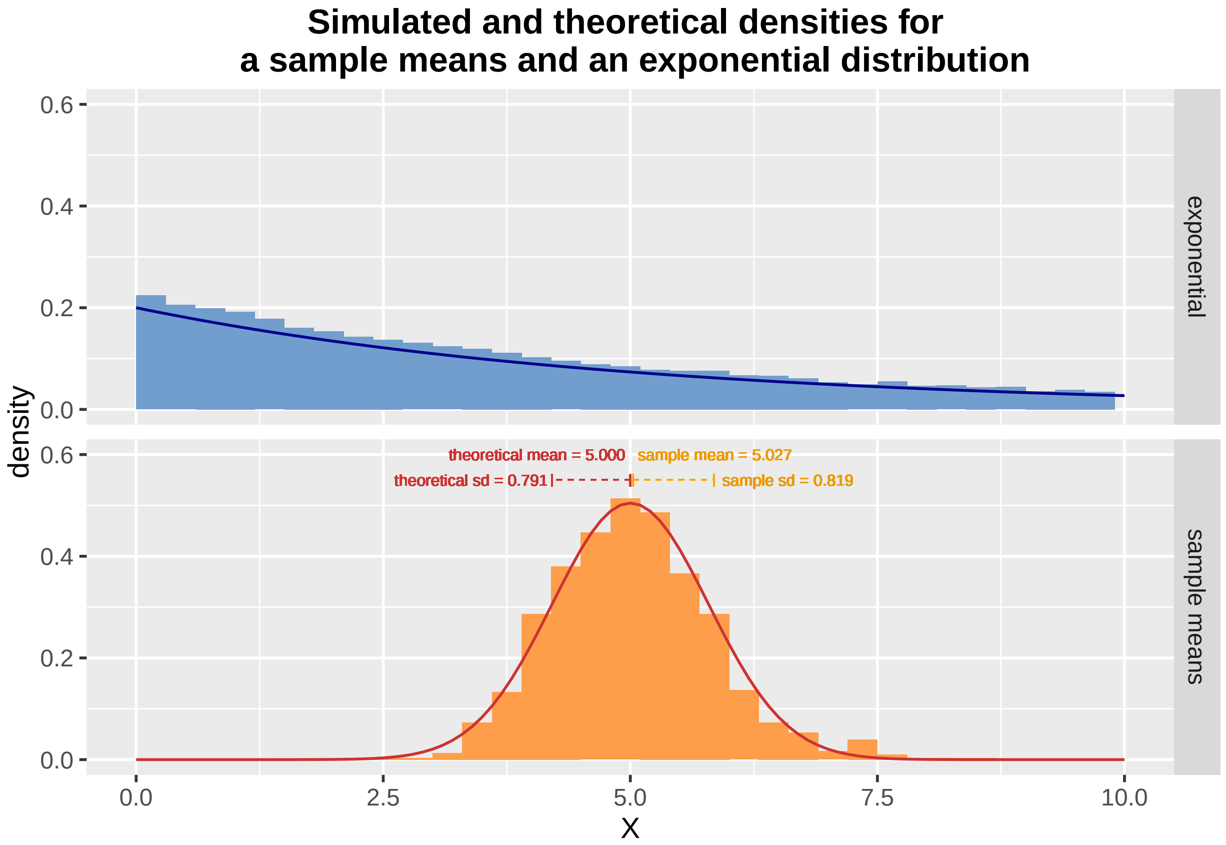
How variable are the samples and how does the variance compare to the theoretical variance of the distribution?
The variance of the samples is 0.670, which is close to the expected variance of 0.625 but higher by about 7%.

For completeness: The variance when the simulated variables are treated as 40,000 samples of size 1 is 25.075, which is only 0.3% from the expected variance of exactly 25.

Assessing whether the sample distribution is approximately normal can be done by looking at plots of the simulated variables and comparing them to theoretical density curves.

```
## Warning: Removed 5420 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 6 rows containing missing values (geom_bar).
```



The plots include histograms for the 40,000 random variables from the exponential distribution (top in blue) and the means of 1000 samples of size 40 (bottom in orange). The plot for the exponential distribution is truncated at $X = 10$ to allow comparison with the sample means distribution using the same scale.

Curves for the theoretical distributions are overlayed on each plot.

Is the sample distribution approximately normal?
In the lower plot, the (brown) line corresponding to the theoretical density curve for a normal distribution $N(\mu, \frac{\sigma^2}{n})$ lies close to the (orange) histogram of the sample means distribution. Based on this apparent fit, the sample means distribution is approximately a normal distribution.

Appendix

This is the full code for creating variables, analyzing, and plotting the results.

```
lambda <- 0.2
```

```
replicates <- 1000
sample_size <- 40
```

```
set.seed(51815)
exp_dist <- matrix(rexp(replicates * sample_size, rate = lambda), nrow = replicates)
```

```
means_dist <- apply(exp_dist, 1, mean)
```

```
exp_df <- data.frame(distribution = c(rep("exponential", 40000),
                                      rep("sample means", 1000)),
                                X = c(c(exp_dist), c(means_dist)))
```

```
exp_mean <- mean(c(exp_dist))
exp_sd <- sd(c(exp_dist))
means_mean <- mean(means_dist)
means_sd <- sd(means_dist)
theor_exp_mean <- 1/lambda
theor_exp_sd <- 1/lambda
theor_means_mean <- 1/lambda
theor_means_sd <- 1/(lambda * sqrt(sample_size))
comparison <- data.frame(matrix(c(exp_mean, exp_sd^2, means_mean, means_sd^2,
                         theor_exp_mean, theor_exp_sd^2, theor_means_mean,
                         theor_means_sd^2), nrow = 4),
                       row.names = c("exponential distribution, mean ",
                                     "exponential distribution, variance ",
                                     "sample mean distribution, mean",
                                     "sample mean distribution, variance"))
```

```
library(knitr)
kable(comparison, digits = 3, row.names = TRUE,
      col.names = c("simulated", "theoretical"))
```

```
library(ggplot2)
library(ggthemes)
ggplot(data=exp_df, aes(X, fill = factor(distribution))) +
  theme(legend.position = "none") +
  scale_fill_tableau("tableau10medium") +
  geom_histogram(aes(y = ..density..), binwidth = 0.3) +
  xlim(0, 10) + ylim(0, 0.6) +
  facet_grid(distribution ~ .,) +
  stat_function(data = subset(exp_df, distribution == "sample means"),
                fun = dnorm, color = "brown3",
                args=list(mean = theor_means_mean, sd = theor_means_sd)) +
  stat_function(data = subset(exp_df, distribution == "exponential"),
                fun = dexp, color="darkblue",
              args = list(rate = lambda)) +
  ggtitle("Simulated and theoretical densities for \n a sample means and an exponential distribution") +
  theme(plot.title = element_text(face = "bold")) +
  geom_point(data = subset(exp_df, distribution == "sample means"),
             aes(x = means_mean, y = 0.55), shape = "|", color = "orange") +
  geom_point(data = subset(exp_df, distribution == "sample means"),
             aes(x = means_mean + means_sd, y = 0.55), shape = "|", color = "orange") +
  geom_text(data = subset(exp_df, distribution == "sample means"),
            aes(x = means_mean + 0.05, y = 0.6), label = "sample mean = 5.027",
            size = 2, color = "orange2", fontface = "plain", family = "sans",
            hjust = 0) +
  geom_segment(data = subset(exp_df, distribution == "sample means"),
               aes(x = means_mean, y = 0.55, xend = means_mean + means_sd,
                   yend = 0.55), color = "orange", size = 0.25, linetype = "dashed") +
  geom_text(data = subset(exp_df, distribution == "sample means"),
            aes(x = means_mean + means_sd + 0.75, y = 0.55), label = "sample sd = 0.819",
            size = 2, color = "orange2", fontface = "plain", family = "sans") +
  geom_point(data = subset(exp_df, distribution == "sample means"),
             aes(x = theor_means_mean, y = 0.55), shape = "|", color = "brown3") +
  geom_point(data = subset(exp_df, distribution == "sample means"),
             aes(x = theor_means_mean - theor_means_sd, y = 0.55), shape = "|", color = "brown3") +
  geom_text(data = subset(exp_df, distribution == "sample means"),
            aes(x = theor_means_mean -0.05, y = 0.6), label = "theoretical mean = 5.000",
            size = 2, color = "brown3", fontface = "plain", family = "sans",
            hjust = 1) +
  geom_segment(data = subset(exp_df, distribution == "sample means"),
               aes(x = theor_means_mean, y = 0.55, xend = theor_means_mean - theor_means_sd,
                   yend = 0.55), color = "brown3", size = 0.25, linetype = "dashed") +
  geom_text(data = subset(exp_df, distribution == "sample means"),
            aes(x = theor_means_mean - theor_means_sd - 0.82, y = 0.55), label = "theoretical sd = 0.79
            size = 2, color = "brown3", fontface = "plain", family = "sans")
```