Simulated distribution of the mean of exponentially distributed random variables

Timo Voipio

11 Aug 2016

Overview

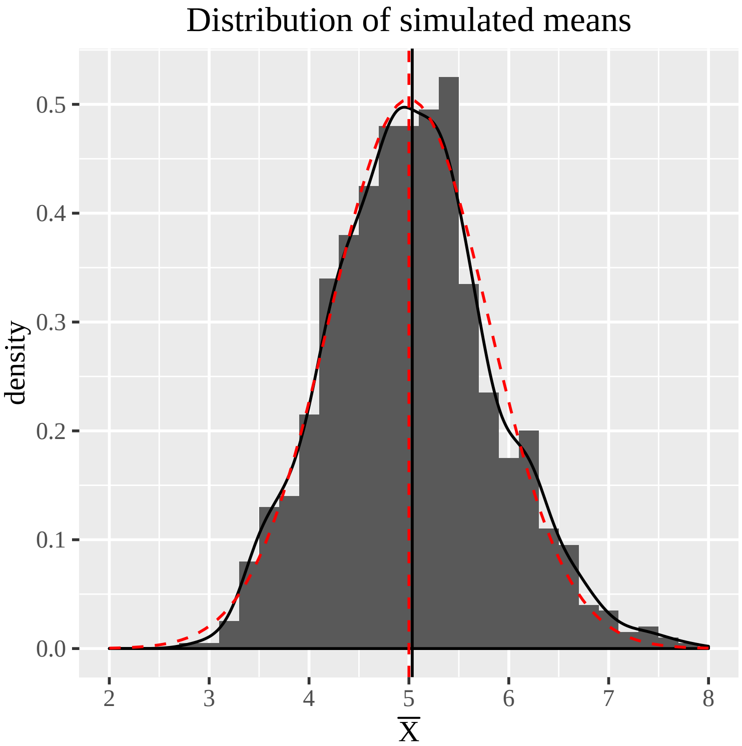
This document explores the distribution of the mean of 40 exponentially distributed independent random variables via computer simulation. The empirically determined distribution is found to be in good agreement with the central limit theorem.

Simulation of the mean

We investigate the distribution of the random variable $\bar{\mathrm{X}} = \sum_{\mathrm{i}=1}^{40} \mathrm{Y}_\mathrm{i}$, where Y_i are *iid*, exponentially distributed random variables with the parameter $\lambda = 0.2$. The simulation is carried out by drawing 40000 numbers from an exponential distribution with rate $= \lambda$, grouping these into 1000 groups of 40 each, and calculating the arithmetic mean for each group.

Simulation results

The distribution of the averages obtained from the simulation is shown in the following figure.



The vertical bars show the (density) histogram of the distribution of $\bar{X}$, and the black curve shows the density. The black vertical line indicates the mean of $\bar{X}$, i.e., the empirical mean. The red dashed curve shows the (theoretical) distribution of $\bar{X}$, while the dashed vertical line presents the population mean of $\bar{X}$.

Mean of averages

The sample mean of $\bar{X}$ is found to be 5.033. Due to the linearity of the expectation value, the expected value of $\bar{X}$ is equal to the expected value of an exponentially distributed random variable with rate parameter λ, i.e. $1/\lambda = 1/0.2 = 5$. The empirical mean deviates less than 1 ‰ from the theoretical prediction.

Variance and distribution of averages

The sample variance of $\bar{X}$ is 0.654. The central limit theorem (CLT) predicts that, for sufficiently large n, the average of n *iid* random variables has the variance σ^2/n, where σ^2 is the variance of the distrubution which the random variables follow. In the case of the exponential distribution, $\sigma^2 = 1/\lambda^2$, and therefore the CLT predicts that the variance of $\bar{X}$ is $1/\lambda^2 n = 1/(0.2^2 \times 40) = 0.625$. The empirical result agrees very well with the prediction given by CLT, so it seems that, at least in this case, $n = 40$ is large enough for the central limit theorem to be applicable.

Distribution of sample means

The central limit theorem predicts that the mean of a large number of *iid* random variables follows the normal distribution. In the figure above, the dashed red curve indicates the probability density of random variables following the distribution $N(5, 25/40)$. This distrubution is the one predicted by CLT for the mean of $n = 40$ *iid* random variables with the mean of $1/\lambda = 5$ and variance of $1/\lambda^2 = 25$. Compared to the empirically determined probability density, we see that the theoretical prediction works quite well. The number of simulations used is not very large, so there is bound to be some Monte Carlo "air" in the result. The appendix presents and discusses the results of a comparable simulation conducted with a significantly larger sample size.

Conclusions

We determined empirically the distribution of the average of 40 *iid*, exponentially distributed random variables. The resulting distribution and its mean and variance was compared to the normal distribution predicted by the central limit theorem (CLT). The empirical results agreed very well with the prediction given by CLT, and we conclude that, at least in this particular case, 40 may be considered to be a "large" number.

Appendix

This appendix includes the R code used to conduct the simulation experiment and to format the results. The simulation was performed using R version 3.2.3 (2015-12-10) on OS X 10.10.5 (Yosemite) [x86_64-apple-darwin13.4.0 (64-bit)].

Source code

```
library(ggplot2)

# Set random seed for repeatability
set.seed(160808)
Nmean <- 40
Nsim <- 1000
rate <- 0.2

# Generate the requisite amount of random numbers from the exponential
# distribution, wrape them into a Nsim x Nmean matrix, calculate
# row means
expmeans <- apply(matrix(rexp(Nmean*Nsim, rate=rate), nrow=Nsim),
                  1, mean)
# Sample mean of Xbar
samplemeanavg <- mean(expmeans)
# Population mean of Xbar
popmeanavg <- 1/rate

# Sample variance of Xbar
samplevaravg <- var(expmeans)
# Population variance of Xbar
popvaravg <- (1/rate^2)/Nmean

# A function for constructing a histogram+density plot of the
# simulated averages, along with the CLT prediction.
# (Placing the code inside a function enables reuse in the Appendix)
meandistplot <- function(means, popmean, popvar, samplemean)
{
    # Calculate suitable bin widths
    binwidth <- ceiling(sqrt(popvar)*2)/10
    # Shift the bin origin such that integers values coincide with
    # bin centers
    binorigin <- binwidth*floor(min(means)/binwidth)-binwidth/2
    # Calculate the limits of the x axis such that at least 3
    # standard deviations (of the CLT normal distribution)
    # are included; mean predicted by the CLT is at the center of
    # the x axis
    xlims <- ceiling(3*sqrt(popvar))*c(-1, 1) + popmean

    # Construct the plot using ggplot2
    g <- ggplot(data.frame(expmean = means), aes(x = expmean))
    # Histogram and density of the simulated data
    ghist <- g + geom_histogram(aes(y = ..density..),
                                binwidth = binwidth, origin = binorigin) +
        geom_density() +
        geom_vline(xintercept = mean(means), color = "black") +
        labs(x = expression(bar(X)),
             title = "Distribution of simulated means") +
        scale_x_continuous(breaks = seq(xlims[1], xlims[2]),
                           limits = xlims) +
        theme(text = element_text(family = "serif"))

    # Determine the shape of the distribution of Xbar predicted by
    # CLT
    densx <- seq(xlims[1], xlims[2], length.out = 101)
    densy <- dnorm(densx, mean = popmean, sd = sqrt(popvar))
    densdata <- data.frame(densx = densx, densy = densy)

    # Add the mean and density predicted by CLT to the plot
    gcomp <- ghist + geom_line(aes(x = densx, y = densy), data = densdata,
                               color = "red", lty = "dashed") +
        geom_vline(xintercept = popmean, color = "red", lty = "dashed")

    return(gcomp)
}

# tidyprint suppresses warnings issued by ggplot2 in the case a bin
# does not contain any values; this may happen since the limits have
# been set manually
# (further simulation with Nsim = 1e6 predicts that, for lambda used
# here, the 1/1000th quantile is at approximately 2.9; for Nsim = 1000,
# it is thus rather improbable that all of the bins 2.1-2.3, 2.3-2.5,
# 2.5-2.7, 2.5-2.9 would contain at least one observation)
tidyprint <- function(obj, tidy)
{
    # If tidy output is desired, disable warnings (caused by xlim
    # containing bins with no values falling in them)
    if (tidy)
    {
        oldw <- getOption("warn")
        options(warn = -1)
    }

    # Print the object using the default print method
    print(obj)

    # Restore previous state of warnings
    if(tidy) { options(warn = oldw) }
```

```
}

# Construct and print a plot of the simulated averages along
# with the CLT prediction
tidyprint(meandistplot(expmeans, popmeanavg, popvaravg, samplemeanavg),
          tidydoc)
```

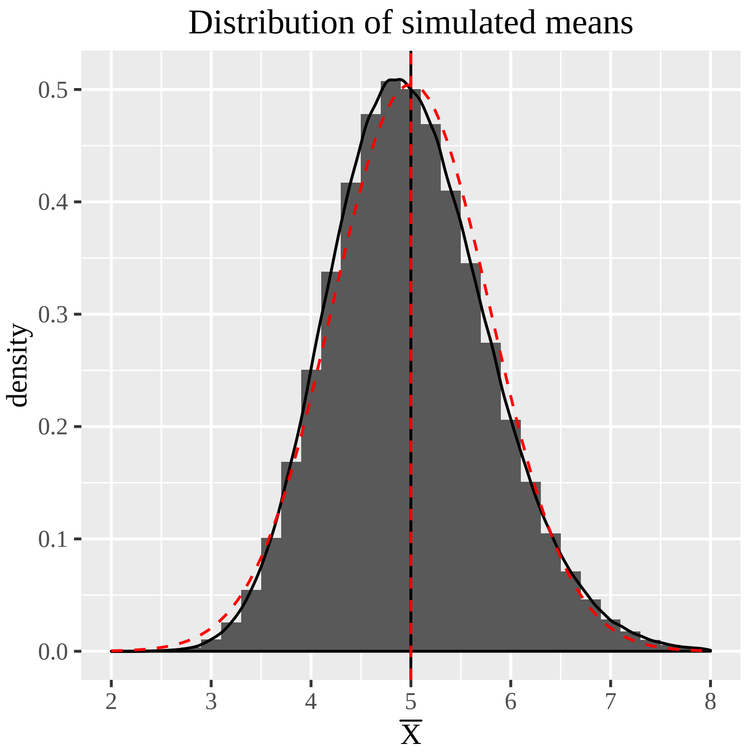
The complete R Markdown source file may be found at https://github.com/tvoipio/JHU_SIProject

Simulation using n = 1000000

Out of interest (and because computing is cheap), it was decided to carry out the simulation of the average of 40 *iid* exponentially distributed random variables using a million simulations, instead of one thousand.

```
Nsim2 <- 1e6
expmeans2 <- apply(matrix(rexp(Nmean*Nsim2, rate=rate), nrow=Nsim2),
                   1, mean)
```

```
# Calculate sample mean and sample variance of the averages calculated
# using Nsim2 simulations and print the resulting distribution
samplemeanavg2 <- mean(expmeans2)
samplevaravg2 <- var(expmeans2)
tidyprint(meandistplot(expmeans2, popmeanavg, popvaravg, samplemeanavg2),
          tidydoc)
```



The sample mean of the 1000000 simulated averages is 4.999539 (CLT prediction 5.00000), and the sample variance is 0.6254257 (CLT prediction 0.62500). We observe that, by increasing the number of simulations the sample mean is, for most practical purposes, exactly the same the as the prediction given by the linearity of the expected value. The variance is also closer to the CLT prediction than the one obtained with 1000 simulations (0.6541418), but not as close as in the case of the mean. It should be kept in mind that the CLT is exact only as n approaches infinity; in our case, n = 40, so it is not surprising that the predicted and empirically determined variances differ.

The shape of the distribution obtained with 1000000 simulations is also clearly not normal; the distribution is positively skewed, the lower tail is thinner than normal, and the upper tail is fatter than normal. Determining the exact distribution, while possibly an interesting exercise, is way, way outside the scope of this project.