Application of Central Limit Theorem to the exponential distribution

Alex Gaggin

Overview

Central Limit Theorem (CLT) states that if we take sufficiently large number of samples of some distribution, then means of these samples will be distributed in a different way - they will follow normal distribution, and their average (mean of means) will be centered on theoretical mean of the original distribution. The following simulation demonstrates how this applies to exponential distribution. Exponential distribution is described by a rate parameter lambda, and then its theoretical mean and standard deviation will both be 1/lambda.

Simulations

We use rexp() function in R to simulate random observations.

Let's set simulation parameters and calculate theoretical values for them.

```r
lambda         <- 0.2
mu             <- 1/lambda                # Theoretical mean
sigma          <- 1/lambda                # Theoretical standard deviation
variance       <- sigma^2                 # Theoretical variance
sample.size    <- 40                      # Sample size
numsim         <- 1000                    # Number of simulations
SEM            <- sigma/sqrt(sample.size) # Theoretical standard error of the mean
var.of.mean    <- SEM^2                   # Theoretical variance of the mean
```

Let's generate set number of samples and calculate properties of the distrinution of their means.

```r
dat <- list(); set.seed(1)
for(i in 1:numsim) dat[[i]] <- rexp(sample.size, lambda)
means <- sapply(dat, mean)
sample.mean <- mean(means); sample.var <- var(means); sample.sd <- sd(means)
```

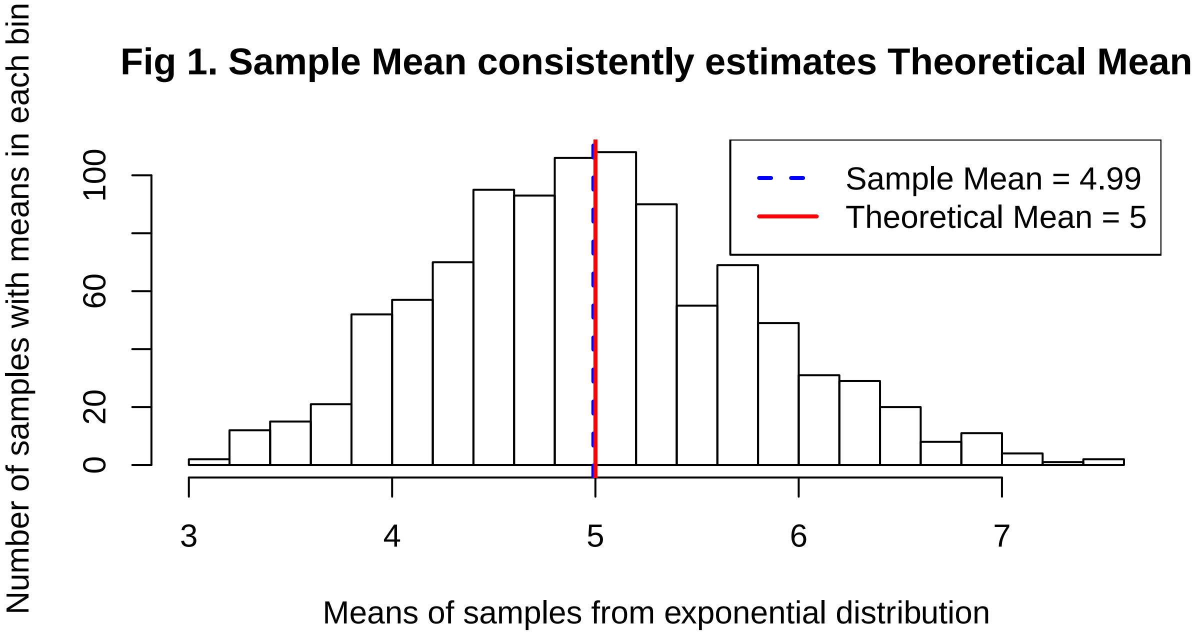
Sample Mean versus Theoretical Mean

The CLT suggests that mean of means of sufficiently large number of samples should converge to theoretical mean of the original distribution.

```
cat(paste("Mean is",sample.mean,"in the simulation and", mu, "in the theory."))
```

```
## Mean is 4.99002520077716 in the simulation and 5 in the theory.
```

Sample and Theoretical means are almost indistinguishable in the histogram of the sample means (see the R code for the plot in the Appendix 3).



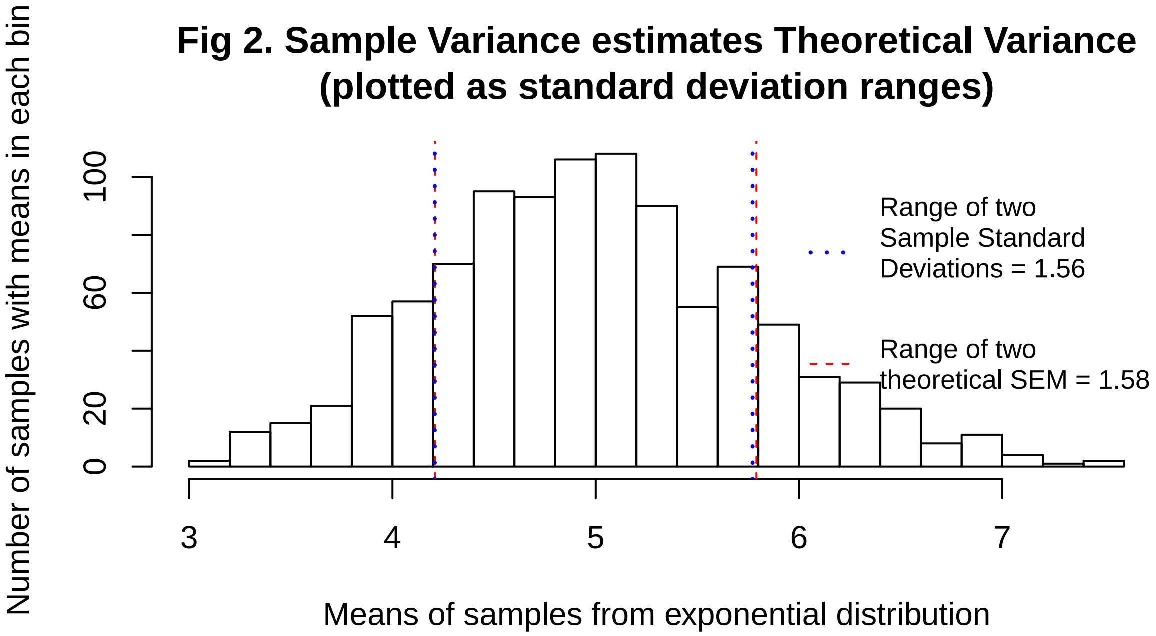
Sample Variance versus Theoretical Variance

Variance of sample means for sufficiently large number of samples should converge to squared theoretical Standard Error of the Mean. Their distribution is more focused than the original distribution, because the more samples we collect, the less error and less variability is in the collected data. At infinite number of samples, n would be infinity, SEM and theoretical variance of the sample means distribution would be zero, and sample mean will be exactly equal to theoretical mean of the original distribution. But we have limited number of samples, so let's see how close sample and theoretical variances of the means are.

```
cat(paste("Variance is",sample.var,"in the simulation and", var.of.mean, "in the theory."))
```

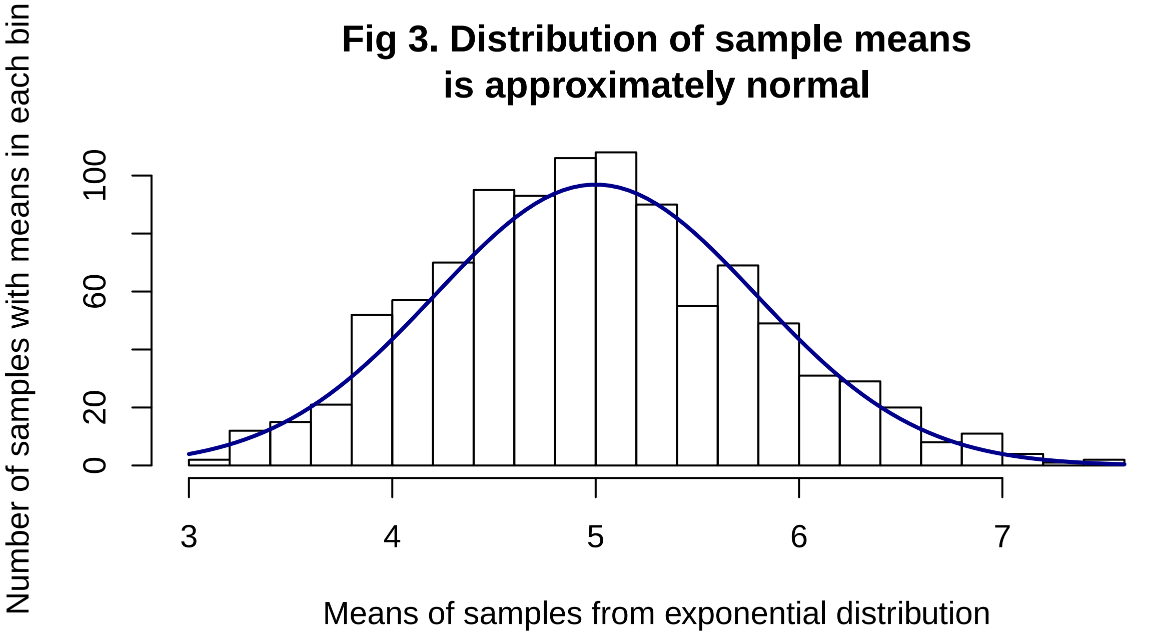
```
## Variance is 0.611116466559575 in the simulation and 0.625 in the theory.
```

For the chart, let's plot range of two standard deviation, because the variance is in squared units and can't be visualized directly (see the R code for the plot in the Appendix 3).



Distribution

Histogram's shape is bell-like, let's draw a normal distribution's bell (with the theoretical mu and SEM) on top of it to illustrate this (see the R code for the plot in the Appendix 3).

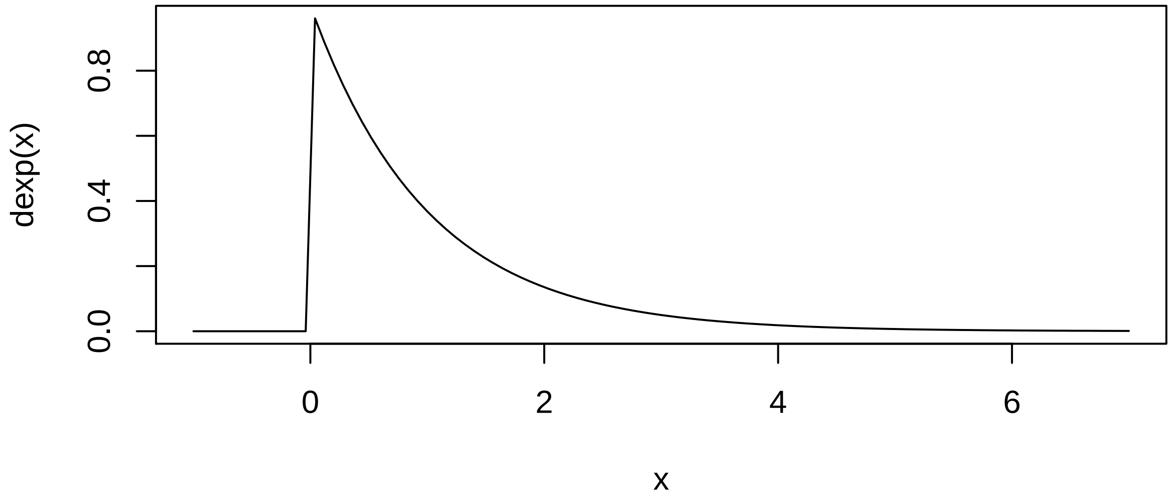


Appendices

Appendix 1. Shape of exponential distribution

This is the shape of exponential distribution. We can look at it by plotting its density function.

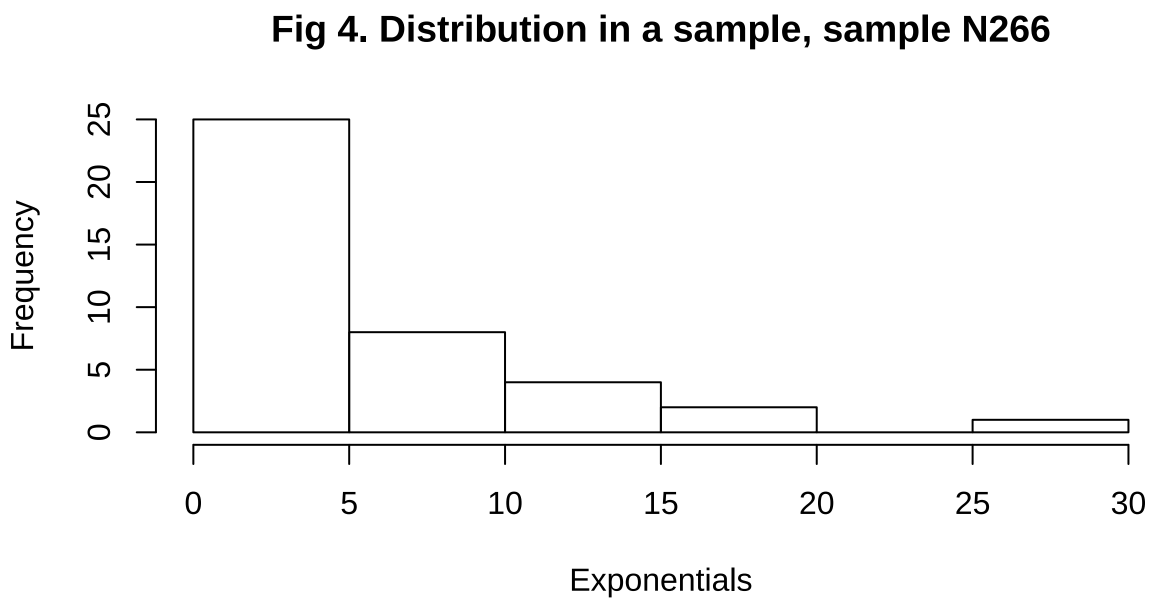
```r
curve(dexp(x),-1,7)
```



Appendix 2. Look at a random sample

If we take a random sample and see it, we should see something similar to the previous plot of the distribution function for the exponential distribution. And its mean and standard deviation should be close to 5 both.

```r
set.seed(1)
random.sample <- sample(1:numsim,1)
hist(dat[[random.sample]],
     main=paste0("Fig 4. Distribution in a sample, sample N", random.sample),
     xlab="Exponentials")
```



```
(random.mean <- mean(dat[[random.sample]]))
```

```
## [1] 5.232086
```

```
(random.sd <- sd(dat[[random.sample]]))
```

```
## [1] 5.491864
```

Appendix 3. R code for plots

R code for Fig 1.

```
breaks <- 20
hist(means, breaks=breaks,
     main="Fig 1. Sample Mean is a consistent estimator of the Theoretical Mean",
     xlab="Means of samples from exponential distribution",
     ylab="Number of samples with means in each bin")
abline(v=sample.mean, lwd=2, lty=2, col="blue")
abline(v=mu, lwd=2, col="red")
legend("topright", c(paste0("Sample Mean = ", round(sample.mean, 2)),
                     paste0("Theoretical Mean = ", mu)),
       lty=c(2,1), lwd=c(2,2),col=c("blue","red"))
```

R code for Fig 2.

```
hist(means, breaks=breaks,
     main="Fig 2. Sample Variance estimates Theoretical Variance
(plotted as standard deviation ranges)",
     xlab="Means of samples from exponential distribution",
     ylab="Number of samples with means in each bin")
abline(v=mu - SEM, lwd=1, lty=2, col="red")
abline(v=mu + SEM, lwd=1, lty=2, col="red")
abline(v=sample.mean - sample.sd, lwd=2, lty=3, col="blue")
abline(v=sample.mean + sample.sd, lwd=2, lty=3, col="blue")
legend("topright", c(paste0("Range of two\nSample Standard\nDeviations = ",
                            round(2 * sample.sd, 2), "\n"),
                     paste0("Range of two\ntheoretical SEM = ",
                            round(2 * SEM, 2))),
       lty=c(3,2), lwd=c(2,1),col=c("blue","red"), bty="n", cex=0.8)
```

R code for Fig 3.

```
hist(means, breaks=breaks, main="Fig 3. Distribution of sample means
is approximately normal",
     xlab="Means of samples from exponential distribution",
     ylab="Number of samples with means in each bin")
curve(192*dnorm(x, mean=mu, sd=SEM),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

Because normal distribution line at the figure above is drawn on the same vertical scale as the histogram, it had to be vertically scaled. Let's use another simulation to determine the scale, assuming that hist()'s plot function algorithm isn't transparent to us, and so we don't know a way to calculate this theoretically. In order to make it more symmetrical, number of observations should be largely increased.

```
larger <- 1000
round(max(hist(rnorm(numsim * larger), breaks=breaks, plot=FALSE)$counts)/larger)
```

```
## [1] 192
```