Averages of exponential distribution vs. CLT

Mariusz Musiał

21 listopada 2015

Overview

In this report we investigate the properties of the averages of samples drawn from exponential distribution. Those properties are compared to the theoretical properies forseen for the distribution of such averages by the Central Limit Theorem (CLT). By comparing the values of mean and variance of this distribution to their theoretical values and the distribution itself to the relevant normal distribution we showed that our investigated disribution follows the CLT.

Simulations

In the present report we'll simulate a series of a random drawings of 40 samples (i.e. n = 40) from the population with exponential distribution with the parameter $\lambda = 0.2$. That means, that the mean of this distribution is $\mu = \frac{1}{\lambda}$ and its standard deviation is $\sigma = \frac{1}{\lambda}$. So, given the λ we have:

- $\mu = 5$
- $\sigma = 5$

In our simulation we'll generate a thousand sets of n = 40 samples of random variables from the exponential distribution with given λ using R `rexp` function. The samples generated this way will form a matrix `m` of 40 columns and 1000 rows. For each row of 40 samples we'll calculate a mean, using `apply` function, thus creating a vector of a 1000 averages $\bar{X}_i$, appropriately called `means`. For the purpose of the following analysis, a data frame called `df` based on `means` is also created. This data frame consists of one column only.

```
set.seed(1901)      # setting the seed for reproducibility
lambda <- 0.2       # given parameter of the exp distribution to simulate
mu <- 1/lambda      # value of our expected mean for the poulation
sigma <- 1/lambda   # value of our expected standard deviation for the population
n <- 40             # number of samples "drawn" from the population
nsim <- 1000        # number of simulations

m <- matrix(rexp(n*nsim,lambda),ncol = n)  # matrix of simulated data, 1000 x 40
means <- apply(m, 1, function(x) mean(x))  # vector of sample means
df <- data.frame(means = means)
```

Sample Mean versus Theoretical Mean

According to CLT, as number of samples approaches infinity, the distribution of averages $\bar{X}_i$ of iid variables approaches that of the normal distribution $N(\mu, \frac{\sigma^2}{n})$.

As a consequence, the sample mean $E[\bar{X}]$, that means, the mean of 1000 averages of 40 samples each, is approaching the population mean μ. That further means, that μ is the Theoretical Mean of the sample mean distribution. In our case, the Theorethical Mean μ is equal to $\frac{1}{\lambda} = 5$. This is also theoretical center of the distribution.

If we look at the summary of our `means` vector we see the following:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   4.420   4.980   5.008   5.545   8.727
```

Our avrerages of exponential samples range from 2.8798418 to 8.7270329 with the mean of **5.0082566**, which is very close to the Theoretical Mean $\mu = 5$. This is the value around which the disribution of the sample mean is centered, as can be seen in Fig. 1 in chapter "Distribution", marked by the blue line.

The relative error of the estimation of population mean, $\frac{\bar{X} - \mu}{\mu} * 100\%$, is 0.17%.

Sample Variance versus Theoretical Variance

According to CLT we expect that the sample mean $\bar{X}$ follows the normal distribution characterized by the variance $Var(\bar{X}) = \sigma^2/n$, where σ^2 is the variance of the population the samples were drawn from. So, the Theoretical Variance for sample mean distribution is σ^2/n.

The actual variance of the sample mean distribution in our simulation can be easily calculated using R function `var` (see code chunk "variances" in Appendix):

$Var(\bar{X}) = 0.694$

The Theoretical Variance value is:

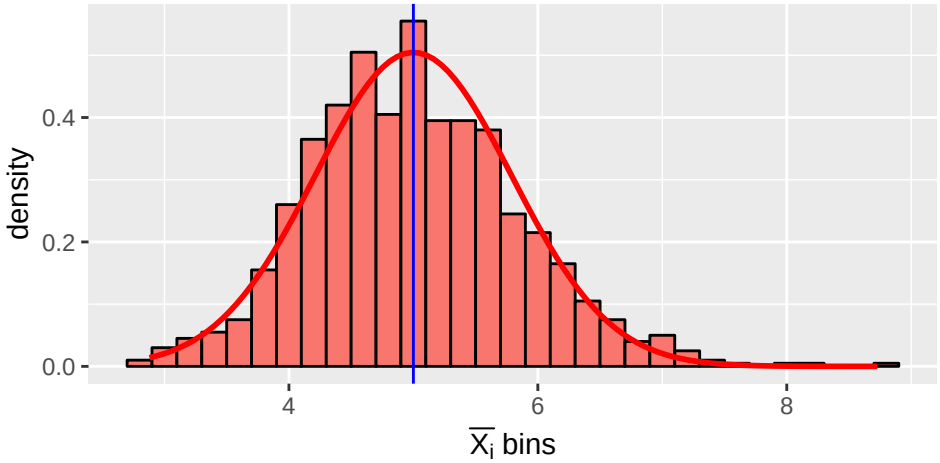
$\frac{\sigma^2}{n} = 0.625$

The variance of the sample distribution is quite close to the Theoretical Variance value, but with the relative error of $\frac{Var(\bar{X}) - \sigma^2/n}{\sigma^2/n} * 100\% = 11\%$ we see a bigger deviation than in case of Sample Mean. We attribute this error to the relatively small sample size (40 samples). We also hypothesize, that sample mean converges faster to μ than sample mean variance to $\frac{\sigma^2}{n}$ for the given increase of n.

Distribution

We compare graphically the distribution of averages $\bar{X}_i$, stored in the vector `means`, with the normal distribution $N(\mu, \frac{\sigma^2}{n})$ of the population from which the samples were drawn. For that purpose we plot the density histogram of $\bar{X}_i$ and we will overlay the plot of the relevant normal distribution (in our case it is N(5,0.625)). Code for generating this plot is provided in the Appendix as code chunk 'hist'.

Fig. 1
Histogram of averages vs. theoretical distribution

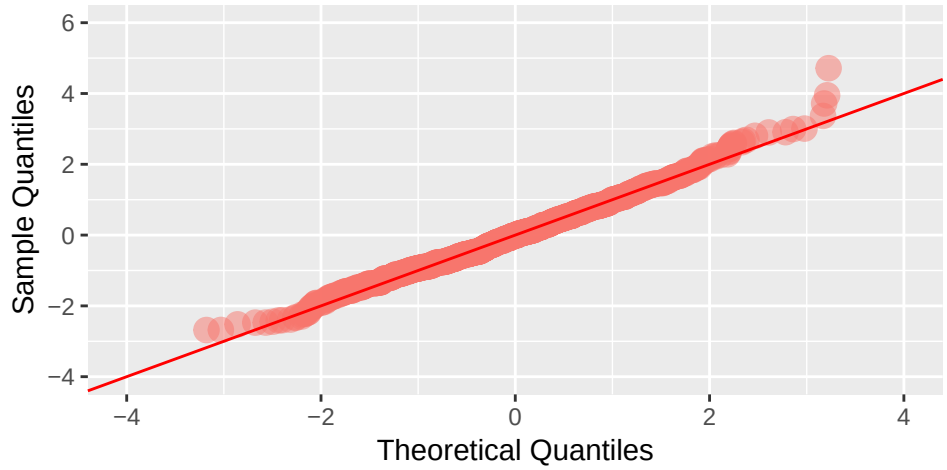


As it can be seen, the density histogram of averages $\bar{X}_i$ follows very closely the distribution forseen by the CLT (red curve), and is centered around the Theoretical Mean (blue line). From this we can conclude that the distribution of averages of 40 samples from exponential distribution, as simulated in this report, follows the normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 0.625$. (For discussion related to variance see the previous chapter.)

Another way to compare two distributions is the so-called Q-Q plot. Here we plot theoretical quantiles of the distribution we are comparing to versus the sample quantiles from our sample distribution. If both theoretical and sample distributions are close to each other, the points on the Q-Q plot will lie on the line with intercept equal to μ and slope equal to 1.

Figure 2 presents Q-Q plot of our sample means distribution stemming from sampling exponential distribution vs. $N(0, 1)$ standard normal distribution. Please note that as we decided to compare with N(0,1), we need to normalize our averages distribution to become $\frac{\bar{X}_i - \mu}{\sigma/\sqrt{n}}$. Code for generating this plot is provided in the Appendix as code chunk 'qqplot'.

Fig. 2
Q−Q plot of sample mean distribution



Vast majority of our sample quantiles lie on the reference line, with only far ends of the tails showing more error. Therefore we conclude that our sample distribution follows the relevant normal distribution, as stipulated by CLT.

More formal test of the distribution can be realized by means of Kolmogorov-Smirnov test, which is however outside the scope of this report.

Conclusions

In this report we have shown that the sample mean of the distribution of averages of 40 samples from exponential distribution is very close to the Theoretical Mean (error: 0.17 %). We have also shown, that the variance of the sample mean is reasonably close to the Theoretical Variance (error: 11 %). We have shown as well, by means of density plots and Q-Q plots, that the sample mean distribution is following almost exactly the theoretical distribution in the range $\pm 2\frac{\sigma}{n}$. From the above we conclude that sample mean distribution of averages of 40 samples from exponential distribution behaves as forseen by CLT.

Appendix

Code for calculating variances

```
# Code chunk 'variances'
mvar <- var(means) # variance of sample mean
s2 <- sigma^2/n    # theoretical variance of sample mean
```

```
mvar
```

```
## [1] 0.6941592
```

```
s2
```

```
## [1] 0.625
```

Code for generating histogram plot

```
# Code chunk 'hist'
f <- f + 1 # for numering the figures, initialized in hidden code chunk
gg <- ggplot(df, aes(x = means, fill = as.factor(1)))
gg <- gg + geom_histogram(binwidth = .2, aes(y = ..density..), color = "black")
gg <- gg + stat_function(fun = dnorm, args = list(mean = mu, sd = sigma/sqrt(n)), size = 1,
                         color = "red")
gg <- gg + geom_vline(xintercept = mu, color = "blue")
gg <- gg + ggtitle(paste("Fig. ", f,
                         "\nHistogram of averages vs. theoretical distribution"))
gg <- gg + xlab(substitute(paste(bar(X[i]),' bins')))
gg <- gg + theme(legend.position="none")
```

Code for generating the Q-Q plot

```
# Code chunk 'qqplot'
sortmeans <- sort((means-mu)/(sigma/sqrt(n))) # Normalize sample distribution
sortref <- sort(rnorm(1000))                  # generate 1000 samples of N(0,1)

f <- f + 1 # for numering the figures, initialized in hidden code chunk
ggqq <- ggplot()
ggqq <- ggqq + geom_point(aes(sortref, sortmeans, color = as.factor(1)), alpha = .5,
                          size = 4)
ggqq <- ggqq + geom_abline(intercept = 0, slope = 1, color = "red")
ggqq <- ggqq + ggtitle(paste("Fig. ", f, "\nQ-Q plot of sample mean distribution"))
ggqq <- ggqq + labs(x = "Theoretical Quantiles", y = "Sample Quantiles")
ggqq <- ggqq + theme(legend.position = "none") + xlim(-4,4) + ylim(-4,6)
```