Using simulation to show exponentional distribution paramaters

Natalie Phillips

24 July 2015

Overview

In this paper we are investigating samples and their ability to estimate population paramaters such as the mean. We are going to use the exponential distribution for our simulations. The exponential distribution describes the probability distribution of the time between events in a Poisson process. For example the exponential distribution may be used to describe the time between people arriving at a bus stop. We are going to show that averages of samples from this distribution are roughly normal distributed and centered at the population mean.

Simulations

One sample of 40 iid values

The exponential distribution follows the formula, $f(x;\lambda) = \lambda \exp^{-\lambda x}$ for $x \geq 0$ where λ is the rate of occurances per unit time. The mean of the exponential distribution, which is also the average time between events, has the formula $\mu = 1/\lambda$. We want to show that distribution of means of random samples from the exponential distribution in a mound shape around the population mean. We will demonstrate that the distribution of sample means is roughly the exponential distribution.

In our experiment $\lambda = 0.2$. A rate of 0.2 tell us that there are an average of .2 events in each period. The average time between events is 5. Each sample contains 40 values randomly drawn from the exponential distribution. Here is one sample of 40 independent and identically distributed, iid, values generated using the R command for random expondentials `rexp`

```r
lambda <- .2
sampleSize <- 40
set.seed(12)
sample <- rexp(sampleSize, rate = lambda)
samplemean <- mean(sample)
```

Here are the first few of our 40 sampled values 10.946081, 3.177752, 0.5869159. The average of these 40 values is 6.0882218 compared to a true population mean of 5. A histogram of this random sample plotted against the exponential distribution showing both the sample mean and the population mean appears in Apendix 1.

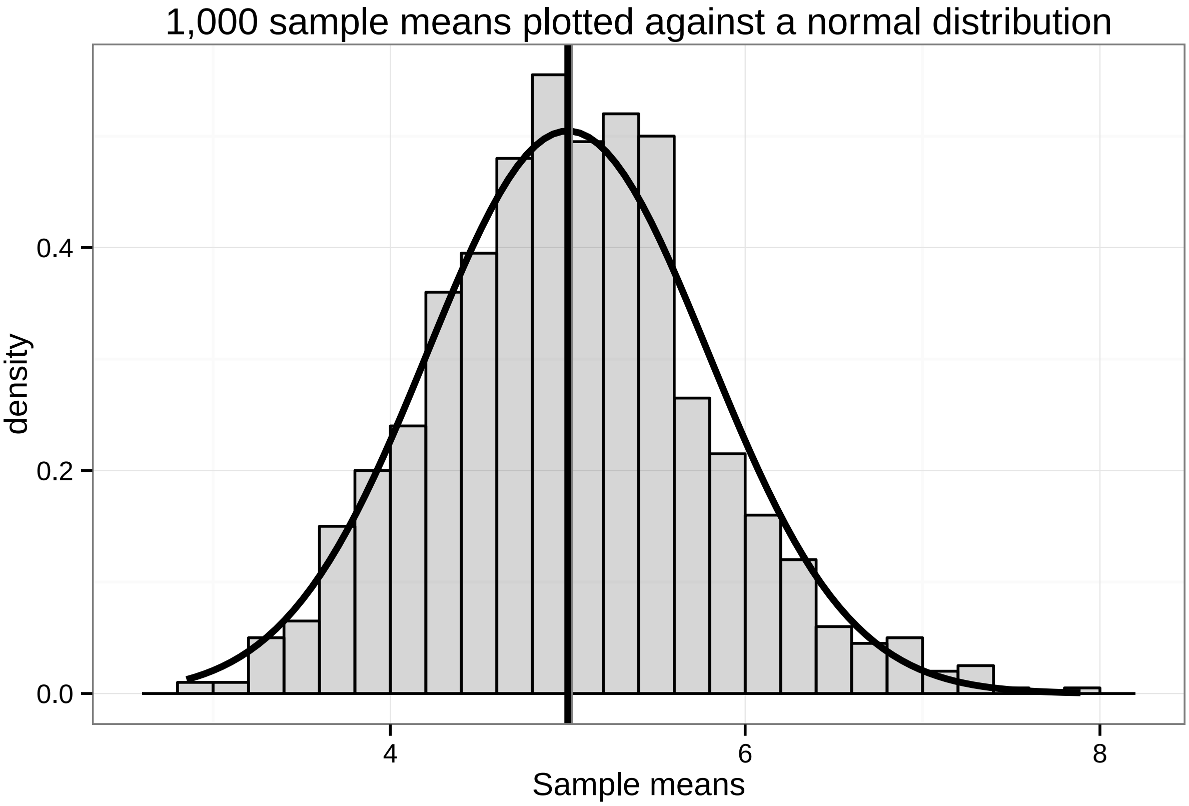
1,000 samples of 40 values

The following simulation takes 1,000 samples like the one above and uses them to generate 1,000 sample means. We are primarily interested in the distribution of these sample means. The following code performs this task. The averages, or means, of all these 1,000 samples is saved in the variable called *mns*. We also collect the sample variance in *vrs*.

```
mns = NULL  #initialise variable for means
vrs = NULL
for (i in 1 : 1000) {
    expdist <- rexp(sampleSize, rate = lambda)
    mns = c(mns, mean(expdist))
    vrs = c(vrs, var(expdist))
}
avgAllMeans <- mean(mns) #Average of all 1000 sample means
sdMean <- 1/(sqrt(40) * lambda)  #Standard error of sample means
```

Sample Mean versus Theoretical Mean

The average of the 1,000 sample means is 5.00924 while the theoretical average, $1/\lambda$, is 5. While the first sample didn't give us a very good approximation to the mean of the exponential distribution once we have taken many such samples and take their average we get closer and closer to the true mean. We know the distribution of sample averages gets closer and closer to a normal distribution by the central limit theroem. We can see from the graph below that the distribution of sample means does indead look close to a normal distribution. The sample mean distribution looks symetric and centred around the population mean. The normal distribution with $\mu = 5$ and $sd = \sigma/\sqrt{40}$ is drawn on the graph for comparison. The normal distribution and the distribution of averages resemble each other closely.



Confidence Intervals

It is by appling the central limit theorem and understanding that the distribution of the means tends to a normal distribution that we are able to produce confidence intervals for the population mean from large samples. Assuming a normal distribution our formula for a 95% confidence interval for μ is

$$\bar{X} \pm Z(.975) \times \sigma/\sqrt{n}$$

where Z is a normal distribution with $\mu = 0$ and $\sigma = 1$. This population average $\bar{X}$ is taken over 40×1000 values. This gives us a sample population of 40,000. Plugging into the formula above give us a confidence interval of 4.96, 5.06. This confidence interval has the interpretation that if we took 100 samples of 40,000 each that the true population mean would lie with the confidence range generated about 95 times out of 100 and outside the range around 5 times out of 100.

Sample Variance versus Theoretical Variance

We don't know the actual distribution of the sample mean variance but we do know that as the number of samples tends to infinity the average sample mean variance will tend to the true theoritical sample mean variance of σ^2/n. The standard error can be found by taking the square root of the sample mean variance. The standard error can be easier to work with as it is in the same units as the original distribution and is less skewed. A histogram of standard errors estimated from 1,000 samples can be seen in Appendix 2. This graph is a little more skewed than we saw with the sample means however it is still mound shape and it's average is near the expected standard error.

```
stderror <- sqrt(vrs/40)
se <- 5/sqrt(40)
avese <- mean(stderror)
```

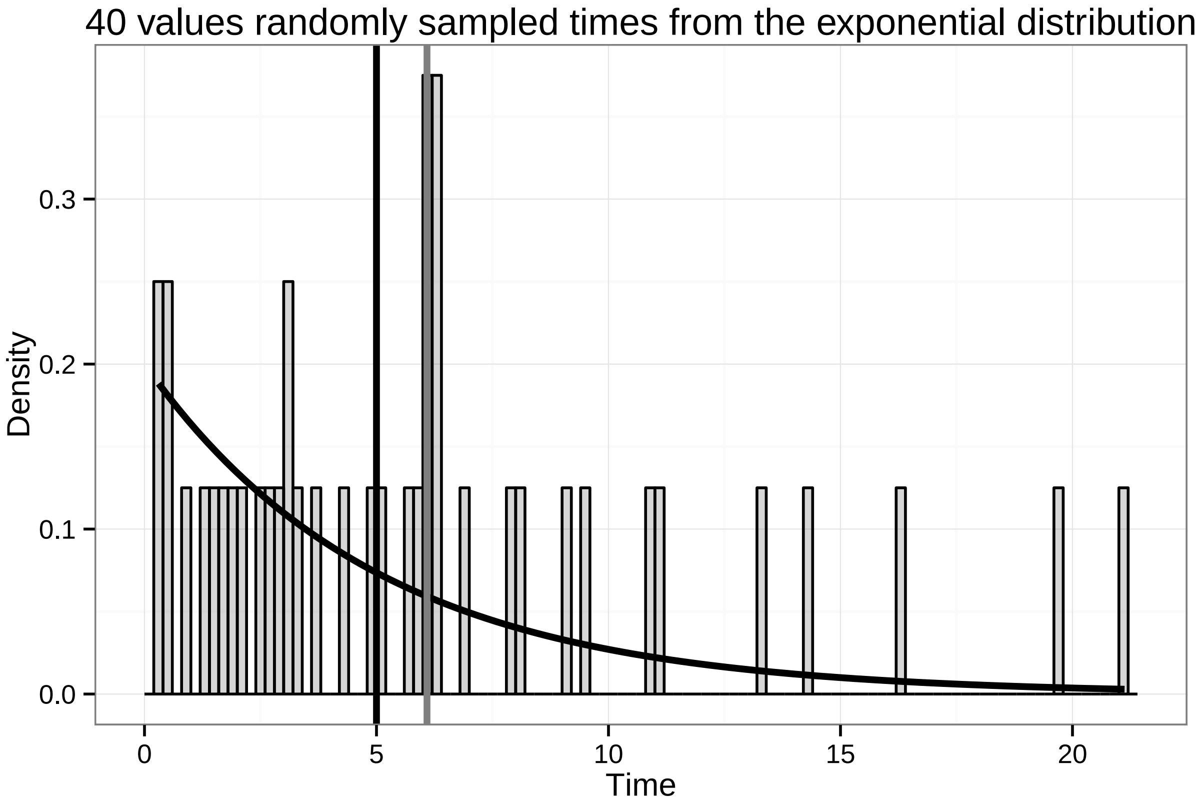
From the code above we know the average standard error from 1,000 samples is 0.7692069 and the true standard error for sample size 40 is 0.7905694. Our estimated variance is a little smaller than the true standard error. Both the estimaged standard error and true standard error appear on the graph as grey and black lines respectively.

Distribution

We can conclude from this analysis that that the distribution of sample means is roughly normally distributed as the sample size gets large. The sample means takes on the right shape for a normal distribution which is symmetric about the mean and bell or mound shaped. The standard error is not as close to normally distributed though it is centred about the true standard error in a mound. If the standard error is not know and the sample size is small a t distribution may be more appropriate for finding confidence intervals than a normal distribution. The t distribution has fatter tails and produces a wider confidence interval which take into account our greater level of uncertainty.

Appendix

Appendix 1: 40 iid values drawn from the exponential distribution

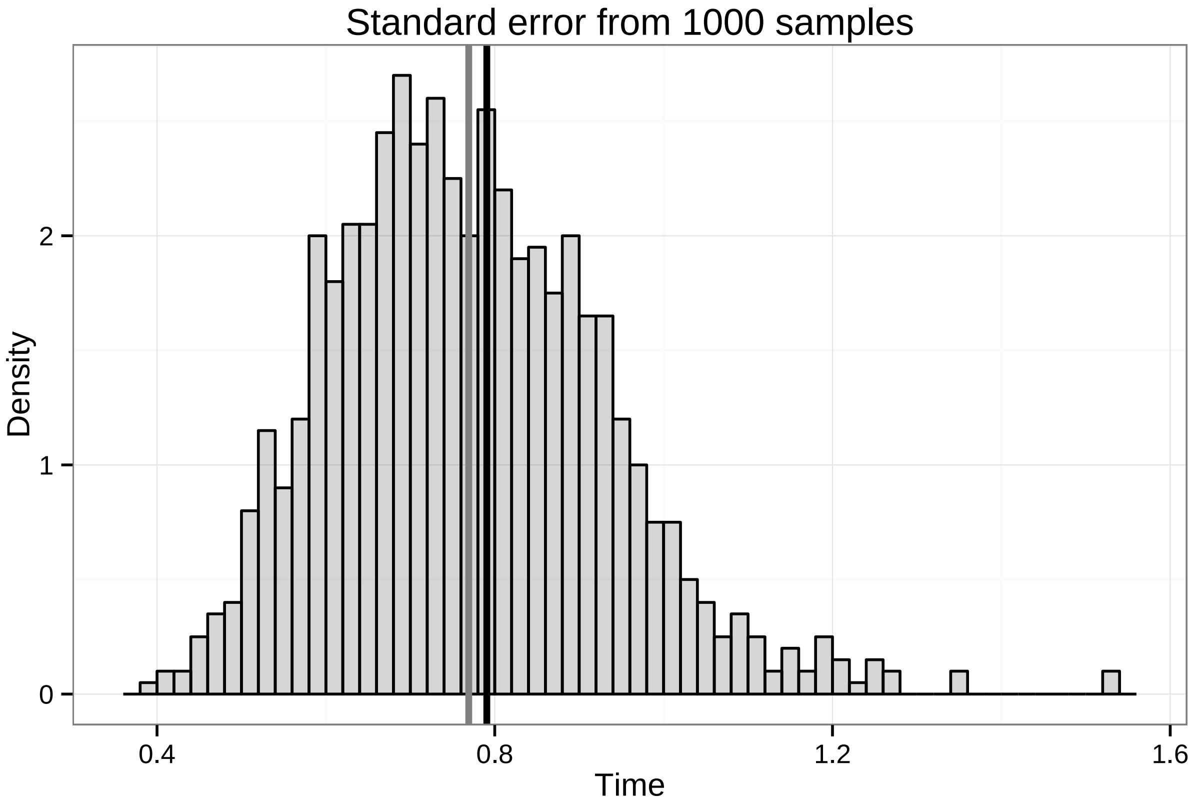


The sample mean, shown on the graph in grey, is an estimate of the distribution mean of 5 however we know if we took another 40 samples from the exponential distribution we would get a different sample mean.

Appendix 2: histogram of sample errors estimaged from 1,000 samples

```
stderror <- sqrt(vrs/40)
se <- 5/sqrt(40)
avese <- mean(stderror)
stderror <- data.frame(stderror)
g <- ggplot(data = stderror, aes(x = stderror)) + geom_histogram(alpha = .20, binwidth = .02, colour =
g <- g + labs(title = "Standard error from 1000 samples")
g <- g + geom_vline(aes(linetype = "Sample mean"), xintercept = avese, colour = "grey50", size = 1.2)
g <- g + geom_vline(aes(linetype = "Density mean"), xintercept = se, size = 1.2)
g <- g + labs(x = "Time", y = "Density")
g <- g + theme_bw()

g
```



The standard error, which was calculated using the variance from the samples of 40 iid draws from the exponential distribution, does not look as symetrical as the means of the samples. However when we take the average of all the standard error approximations we get 0.7692069 which is close to the theoretical standard error of 0.7905694. If we were to take more and samples the average of the estimated standard errors would get closer to the mean. Alternatively if we took larger samples we would also get a better estimate of the sample mean and hence the variance.