Investigation of the distribution of the mean of samples (n=40) of exponential distributions

Michael Hunt

Thursday, June 04, 2015

Overview

This short report investigates the properties of the distribution of means of samples of size $n = 40$ exponential distributions. We show that as the number of samples increases the average value of this mean approaches the population mean μ, that the variance of the means approaches the value $\frac{\mu}{n}$ and that the distribution of the means approaches the normal distribution. These results illustrate the Law of Large Numbers and the Central Limit Theorem respectively.

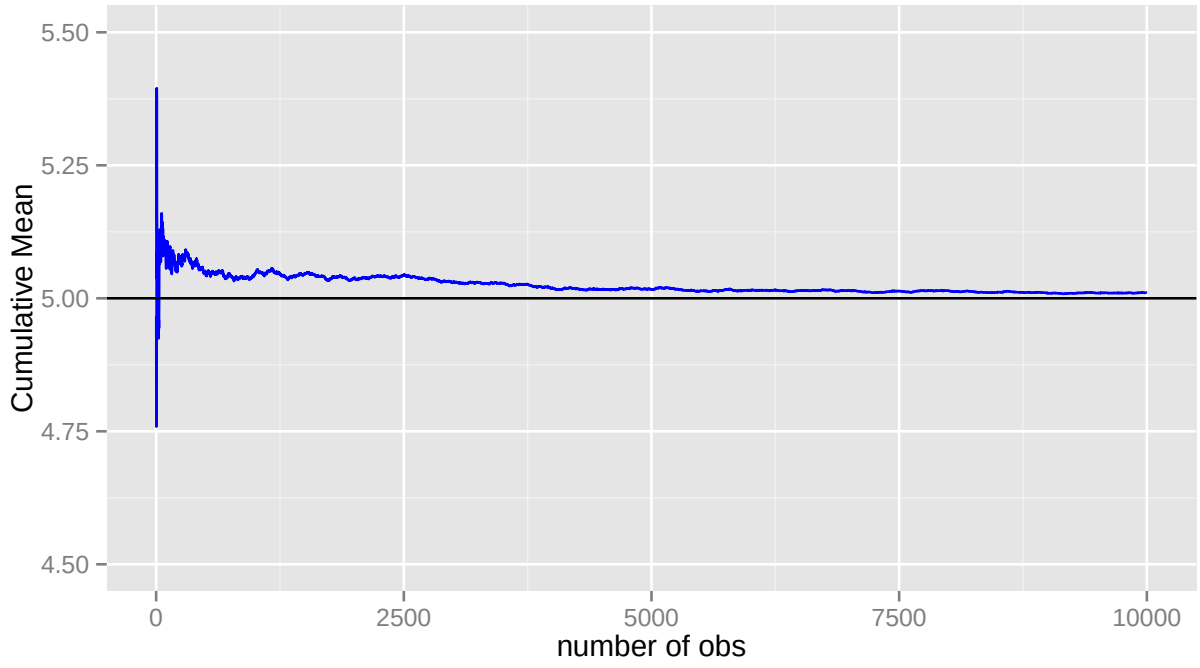
Simulations

Sample Mean versus Theoretical Mean

The theoretical expected value, or mean of the exponential distribution $f(x) = \lambda e^{(-\lambda/x)}$ is given by

$\mu = \int_0^\infty x f(x) dx = \int_0^\infty x \lambda e^{(-\lambda/x)} dx = \frac{1}{\lambda}$ Hence for $\lambda = 0.2, \mu = 5$. We shall use this value of λ throughout this report.

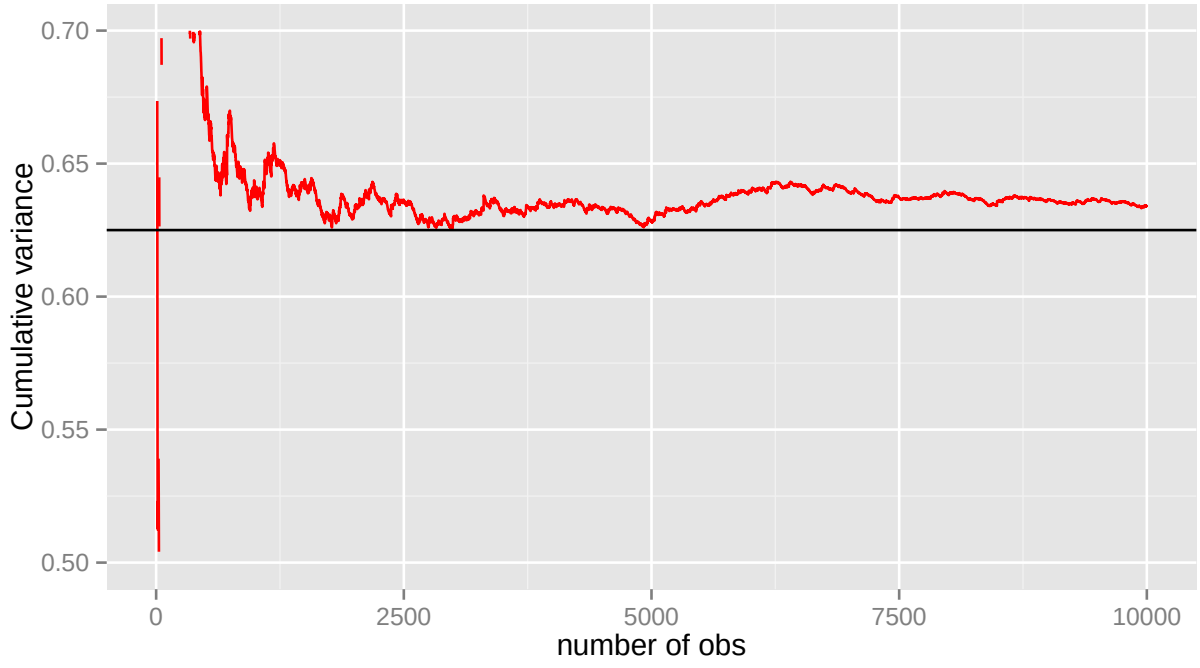
The mean of a collection of means of samples of exponential distributions of size n will tend towards the theoretical population mean as the number of samples tends towards infinity since the sample mean is a **consistent** estimator of the population mean. This is the Law of Large Numbers. We see this in the figure below, win which we plot the mean of the first 1,10....10000 values of 10000 means of samples of 40 exponential distributions, each with $\lambda = 0.2$. The asymptote is clearly equal to the population mean of 5.



Sample Variance versus Theoretical Variance

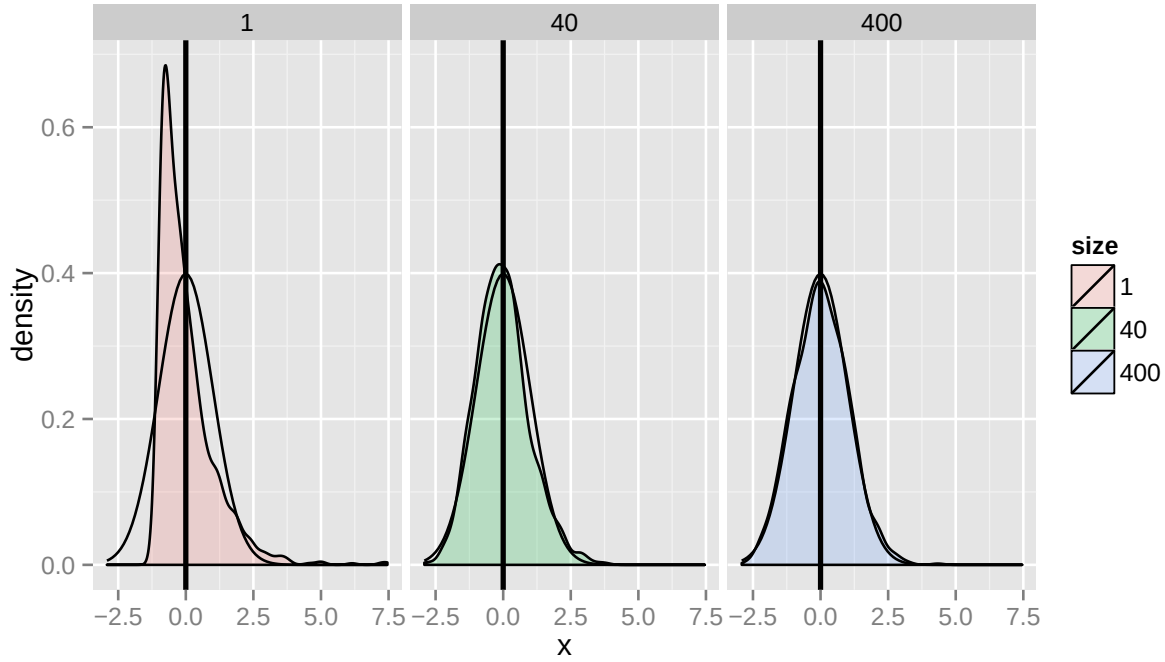
The variance of the exponential distribution is given by $var(f(x, \sigma = \lambda)) = \frac{1}{\lambda^2}$, and hence for $\lambda = 0.2$, the variance of a sample should tend towards $\frac{1}{0.2^2} = 25$.

As the number of samples increases, the variance of the distribution of averages of samples of size n should tend towards $var(f(x, \sigma = \lambda)) = \frac{\sigma^2}{n} = \frac{1}{n\lambda^2}$, since the sample variance is also a good estimator of the population variance. Hence for $\lambda = 0.2$ and $n_s = 40$ we should expect that the limiting value of the variance of the mean is $\frac{1}{n\lambda^2} = \frac{1}{40 \times 0.2^2} = 0.625$ This is illustrated in the figure below in which we plot the mean of the variance of the first 1..10..10000 of 10000 samples of 40 exponential distributions, each with $\lambda = 0.2$. The asymptote is plausibly 0.625.



Distribution

As the number of samples in each distribution increases, so the distribution of the means of those samples approaches the normal distribution. In the figure below we show the distribution of the normalised means of 1000 samples of size 1, 40 and 400 respectively, with a standard normal distribution ($\mu = 0, \sigma = 1$) superposed on top. By "normalised we mean that the population mean ($\frac{1}{\lambda} = \frac{1}{0.2} = 5$) has been subtracted from each of the sample means, and the resulting difference divided by the sample standard error $= \frac{\sigma}{\sqrt{(n)}} = \frac{5}{n}$ where n is the sample size.



The figure clearly shows that the normalised distribution of sample means approaches the standard normal distribution as the sample size increases (it is already a very good approximation for $n = 40$), despite that the sample distribution itself, in this case the exponential distribution, is not normal. The left -hand distribution in the figure, effectively the distribution of a sample of 1000 exponential distribution means, is clearly not normal, Collectively, this figure illustrates the Central Limit Theorem.

Appendix - Code used to generate the distributions and figures

Means

```
nr<-10000
ns<-40
lambda<-0.2
means<-apply(matrix(rexp(ns*nr,lambda),nr, ns), 1, mean)
cummeans<-cumsum(means)/(1:nr)
```

```
library(ggplot2)
library(dplyr)
dfm<-data.frame(1:nr,cummeans);
names(dfm)<-c("n","means")
gm<-ggplot(data=dfm,aes(x=n,y=means))+geom_line(colour = "blue")+
        ylim(4.5,5.5)+
        geom_abline(intercept = 1/lambda, slope = 0)+
        labs(x = "number of obs",y = "Cumulative Mean")
gm
```

Variances

```
vars<-apply(matrix(rexp(ns*nr,lambda),nr, ns), 1, var)
cumvars<-cumsum((means-(1/lambda))^2)/(1:nr)

dfv<-data.frame(1:nr,cumvars);
names(dfv)<-c("n","vars")
```

```
gv<-ggplot(data=dfv,aes(x=n,y=vars))+geom_line(colour = "red")+
        geom_abline(intercept = (1/lambda)^2/ns, slope = 0)+
        ylim(0.5,0.7)+
        labs(x = "number of obs",y = "Cumulative variance")
gv
```

Distributions

```
nosim<-1000
dat <- data.frame(
  x = c(apply(matrix((rexp(1*nosim,lambda)-1/lambda)/((1/lambda) / sqrt(1)),nosim, 1), 1, mean),
        apply(matrix((rexp(40*nosim,lambda)-1/lambda)/((1/lambda) / sqrt(40)),nosim, 40), 1, mean),
        apply(matrix((rexp(400*nosim,lambda)-1/lambda)/((1/lambda) / sqrt(400)),nosim, 400), 1, mean)
        ),
  size = factor(rep(c(1, 40, 400), rep(nosim, 3))))
```

```
g <- ggplot(dat, aes(x = x, fill = size)) + geom_density(alpha = .20,colour = "black")+
        stat_function(fun = dnorm, args = list(mean = 0, sd = 1),colour="blue")

g <- g + geom_vline(xintercept = 0, size = 1)
g+facet_grid(.~ size)
```