The Central Limit Theorem - simulation of random exponential samples

Synopsis

This project is about simulation of random variables and the central limit theorem. The means of 1000 samples of 40 iid random variables (drawn from an exponential distribution) are treated as a random sample, and compared to a normal distribution.

Exponential distribution

For random variables drawn from a (continuous) exponential distribution of rate λ, the propability distribution function is given by

$$P(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

This distribution has a population mean of $\mu = 1/\lambda$, a population variance of $1/\lambda^2$, and a standard deviation of $\sigma = 1/\lambda$.

Central Limit Theorem

The central limit theorem states that as the sample size increases the mean of a sample of iid random variables should itself behave like a random normal variable, with mean equal to the mean of the original population, and stamdard deviation equal to the standard error of the sample mean.

Simulation

Two datasets will be generated via simulation, the first is a sample of 1000 iid random variables drawn from an exponential distribution with rate $\lambda = 0.2$. The second dataset consists of 1000 sample means, where each sample consists of 40 iid random variables drawn from an exponential distribution with rate $\lambda = 0.2$. The R code used to compute all results in this report is listed in the appendix.

Sample Mean

The population mean for an exponential distribution is equal to $1/\lambda$, which in this case is $1/0.2 = 5$. For each sample, the mean should be close to this value. The mean of 1000 sample means should also be close to this value. For the samples generated for this report, the mean is equal to 5.0004, which is in good agreement with the expected value.

Sample Variance

As the sample under consideration is a set of means, the variance of this sample should be given by

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n},$$

which is the square of the standard error in the mean. For the exponential distribution, the variance (σ^2) is given by $1/\lambda^2$. The expected variance for our sample of means would then be

$$\mathrm{var}(\bar{X}) = \frac{1}{\lambda^2 n}$$

For samples of 40 variables, with $\lambda = 0.2$, we would expect the means to have variance 0.625. The value obtained from the simulated samples is equal to 0.594, which is within 5% of the expected value.

Another way to check that the distribution of the sample means is approximately normal is to evaluate the probability of a variable lying within a certain range of the mean. For a normal distribution, 68% of the values should lie within one standard deviation of the mean. For our sample of exponential means, this information is summarised in the table below

X	Area within $X\sigma$ of mean for a normal distribution	fraction within $X\sigma/\sqrt{40}$ for sample of means
1	0.683	0.706
2	0.954	0.957
3	0.997	0.998
4	1.000	1.000

There is good agreement between the proportions of sample means and the probabilities from the normal distribution. Histograms are plotted for the exponential sample and the means of 1000 exponential samples in the appendix. The sample means have a distribution that appears to be normal.

Conclusion

The means of simulated samples of random variables drawn from an exponential distibution are (at least approximately) normally distributed. This is consistent with the central limit theorem.

Appendix

load packages:

```
require(ggplot2)
require(gridExtra)
```

simulate random variable samples

```
set.seed(5074491)
samples<-1000
lambda<-0.2
samples1k<-rexp(samples,rate=lambda)
means1kx40<-replicate(samples,mean(rexp(40,rate=lambda)))
```

average of exponential sample means

```
themean<-mean(means1kx40)
themean
```

```
## [1] 5.000418
```

variance of exponential sample means

```
var(means1kx40)
```

```
## [1] 0.5938206
```

Distribution of exponential sample means. Count how many sample means lie with in multiples of the standard error of the mean. Use the pnorm function to get the probability for the relevant region of a normal distribution.

```
sigmas<-c(1,2,3,4)
fractions<-sapply(sigmas, function(x) { sum((means1kx40 < (themean + x/(lambda*sqrt(40)) ))
round(fractions,digits=3)
```

```
## [1] 0.706 0.957 0.998 1.000
```

```
# normal distribution
normfracs<-sapply(sigmas,function(x) { pnorm(x) - pnorm(-x)})
round(normfracs,digits=3)
```

```
## [1] 0.683 0.954 0.997 1.000
```

Plot stuff. On the left, a histogram is plotted from the simulated sample of random exponentials, with the exponential distribution function drawn in red (scaled based on the number of samples). On the right, a histogram is plotted from the means of each sample of 40 random exponentials. For the right hand plot, the curve in red shows a normal distribution function. In both plots, the theoretical curve in red agrees well with the histogram of the simulated quantities.

```r
# scale the 'd' functions so that they appear in proportion with the histograms
# scale factor is 1000 (samples) multiplied by binwidth (0.2)
dnormscale<-function(x,scale=1.0,...) { scale*dnorm(x,...)}
dexpscale<-function(x,scale=1.0,...) { scale*dexp(x,...)}
df<-data.frame(samples1k,means1kx40)
hist_exp<-ggplot(data=df,aes(x=samples1k)) +
geom_histogram(binwidth=0.2,aes(fill=..count..)) +
labs(title='distribution of 1000 random exponential variables',
    x='value',y='count') +
scale_fill_gradientn(colours=c('blue','purple')) +
stat_function(fun=dexpscale,colour='red',args=list(scale=samples*0.2,rate=lambda))+
 guides(fill=F)

hist_expMean<-ggplot(data=df,aes(x=means1kx40)) +
geom_histogram(binwidth=0.2,aes(fill=..count..)) +
labs(title='distribution of 1000 sample means of 40 random exponential variables',
    x='value',y='count') +
scale_fill_gradientn(colours=c('blue','purple')) +
guides(fill=F)+
stat_function(fun=dnormscale,colour='red',
    args=list(scale=0.2*samples,mean=1/lambda,sd=1/(lambda*sqrt(40))))

grid.arrange(hist_exp,hist_expMean,nrow=1)
```

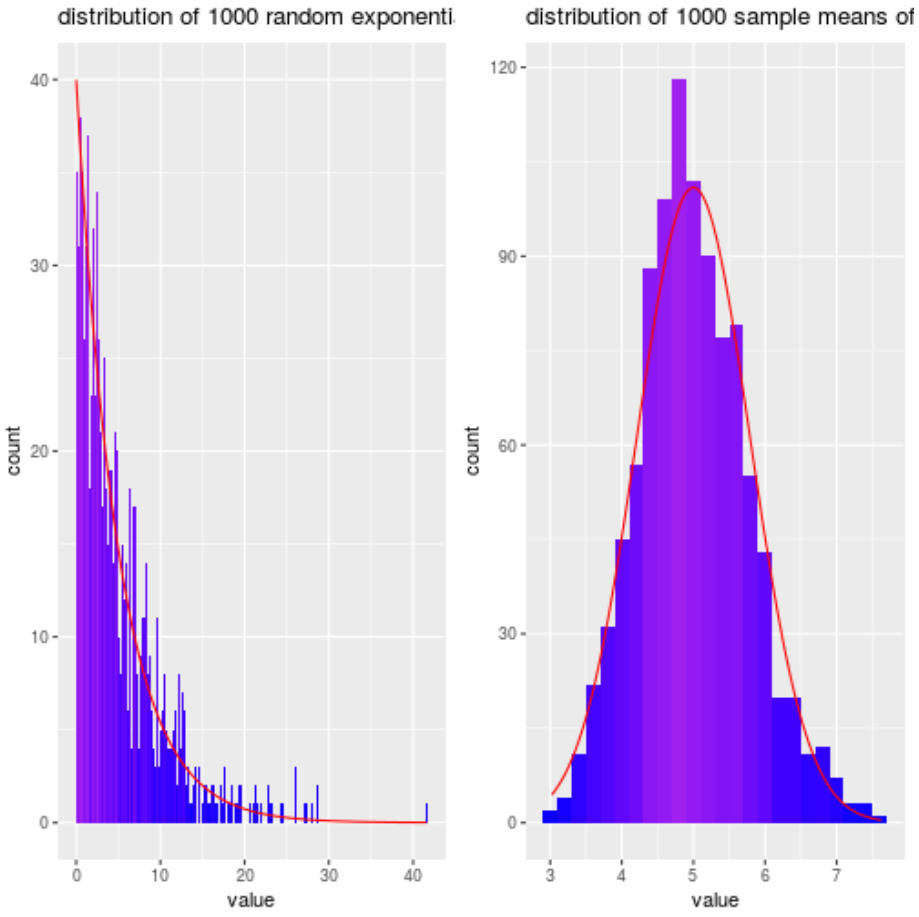

Figure 1: plot of chunk plot