A Demonstration of the Central Limit Theorem

Simulation of the sample average of a set of exponentials

AKN

Overview

In this report, we simulate taking the average, $\bar{X}_n$, of a set of n *iid* observations from an exponentially distributed variable X. We will use the Central Limit Theorem(CLT) to calcuate the expected mean μ and other parameters (variance, sd, 95% CI) of the distribution of $\bar{X}_n$, and then compare to the matching observed parameters in the sample average distribution. This exercise will demonstrate the CLT by showing that the distribution of the sample average is normally distributed, given a large number of iterations.

Simulations

First, we simulate $\bar{X}_n$, the average of a sample of size n taken from a collection of *iid*. Our *iid* will come from a random, exponentially distributed variable X with rate parameter λ. The exponential distribution has a mean $\mu = \frac{1}{\lambda}$, and standard deviation $s = \frac{1}{\lambda}$. We set $\lambda = 0.2$, sample size $n = 40$, and perform 1000 iterations. The average of each of the 1000 samples are calculated and saved to a data frame along with a set of 1000 random exponentials.

```
n = 40  #sample size
reps = 1000 #number of simulations
lambda = 0.2
set.seed(936907)
#Simulations
samples <- replicate(reps, rexp(n,lambda))
a <- rexp(reps, lambda); b <- apply(samples, 2, mean)
data <- data.frame(exp = a, means = b)
```

Using CLT to calculate mean and variance

The exponential distribution has a theoretical mean equal to the inverse of rate parameter, $\frac{1}{\lambda} = 5$. We can compare this to the mean of the sample averages, $\bar{X}_n$. To calculate other parameters of the distribution of $\bar{X}_n$, we can use the CLT which states that $\bar{X}$ is approximately $N(\mu, \sigma^2/n)$, and that:

$$\text{sample variance} = \frac{\sigma^2}{n} = (\frac{1}{\lambda})^2 * 1/n = \frac{25}{40} = 0.625 \text{ and sample sd} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{40}} = 0.791$$

Also, the quantity $\bar{X}_n \pm z_{1-\alpha/2}\sigma/\sqrt{n}$ is the 95% interval of μ. How close are the theorized variables to the actual values in the simulated distribution of the sample averages? Below, we measure all the parameters of the sanple average and the variance, sd and 95% interval for the sample average distribution. Then we compare them directly to the theoretical counterparts, based on the formulas above. By showing how the observed values of the simulations match with calculations from theory, we demostrate the utility of the CLT.

```r
#Calculations
mu = 1/lambda; s = (1/lambda)/sqrt(n); sigma = s^2 #theoretical mu, sigma and s
obmean <- mean(data$means) #observed mu
sd_sample_mean <- sd(data$means); var_sample_mean <- var(data$means) #obs. s and sigma
s95CI <- obmean + c(-1,1)*2*(sd_sample_mean/sqrt(n))
t95CI <- mu + c(-1,1)*1.96*(s/sqrt(n))
#Table 1
t1 <- data.frame(Theoretical=c(mu, s, sigma, t95CI[1], t95CI[2]),
                 Simulated=c(obmean, sd_sample_mean, var_sample_mean, s95CI[1], s95CI[2]))
rownames(t1) <- c("Mean", "SD", "Variance", "95% CI", " ")
print(xtable(t1, digits=4, caption = "Sample Parameters vs. Theorized Parameters"), comment = FALSE)
```

	Theoretical	Simulated
Mean	5.0000	5.0154
SD	0.7906	0.7916
Variance	0.6250	0.6266
95% CI	4.7550	4.7651
	5.2450	5.2658

Table 1: Sample Parameters vs. Theorized Parameters

Distributions

We will use our simulated data to test the central limit theorem. By the CLT, the simulated sample mean will have a normal distribution, despite the shape of the underlying distribution. If one takes enough samples, the mean of those samples will be **normally distributed**. *Figure 2* shows density plots, with an overlay of the normal distribution centered at μ (showed by the red curve and red line for the mean). In this plot, we can see that **while the distribution of a large set of exponentials is not normal, the distribution of the sample average of exponentials is approximately normally distributed**. The right panel provides another way to confirm the normality of the distribution of means: the quantile-quantile plot, or Q-Q plot.

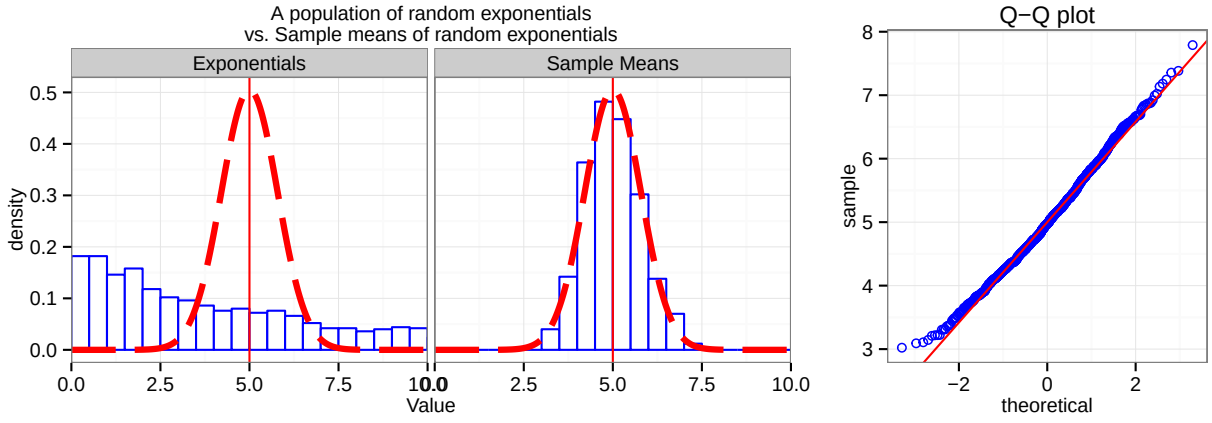

Figure 1: Demonstrating that the Distribution of Sample Means is Normal

Conclusions

1. We simulated sampling the mean of an exponential distribution. The simulated mean was 5.015, close to the theoretical mean, 5.

2. The simulated variance, 0.625, was cloe to the variance of the sampled means, 0.627.

3. The 95% CI of the distribution of sample means, 4.765, 5.266, contains μ, and lines up with the theoretical interval 4.755, 5.245.

4. We began with a population of exponentals that were not normally distributed. Then, we took the means of many samples of size n. The distribution of these sample means **are normally distributed**. This is the main prediction of the Central Limit Theorem.

Note: Code for graphs not included in this report due to space limits, but markdown file can be found here.