Statistical Inference: Course Project

rtaph

November 23, 2014

Part I: The Exponential Distribution

To explore the exponential distribution, we begin by making a histogram of an experiment where we draw 40 random variables X from the distribution. For our purposes, we will use the exponential distribution with $\lambda = 0.2$, but any value of lambda greater than zero can be chosen. We can get a sense of what distribution looks like by plotting a histogram of this experiment:

```
set.seed(5678); n = 40; lambda = 0.2;
experiment1 = rexp(n, lambda)
hist(experiment1, xlab = "Value of Random Variable X")
abline(v=mean(experiment1), col = "red", lwd = 3)
legend("topright", legend = "Sample Mean", lwd=3, col="red")
```

In our experiment, the mean of the 40 draws is calculated to be 5.8191691 (shown in red above). Given the inherent variability in random data, we might expect this sample mean ($\bar{x}$) to differ slightly from the population mean from which it was drawn (μ).

To investigate this, we turn to asymptotics. By repeating the experiment many times through a monte carlo simulation, we may be able to gain stronger insight into the true nature of the population. We repeat the experiment a thousand times and take the mean of each experiment:

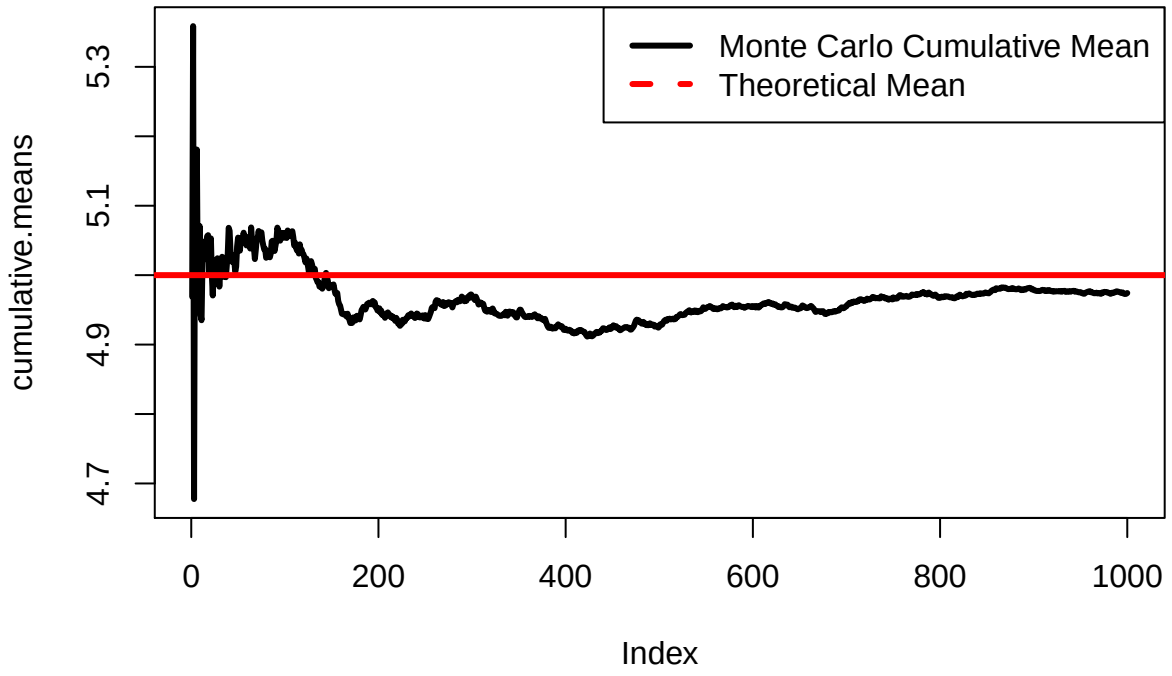
```
monte.carlo.means = NULL; m = 1000; set.seed(1234);
for (i in 1 : m) monte.carlo.means = c(monte.carlo.means, mean(rexp(n, lambda)))
```

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

The expected value of an exponentially distributed random variable X with rate parameter λ is given by $E[X] = \frac{1}{\lambda}$. For the distribution chosen in this paper, the theoretical mean evaluates to $1/0.2 = 5$.

By plotting the cumulative mean for each additional experiment (in black), we can see that the monte carlo mean asymptotes to the theoretical mean (in red).

```
theoretical.mean = (1/lambda)
cumulative.means <- cumsum(monte.carlo.means) / (1 : m)
plot(cumulative.means, type = "l", lwd = 3)
abline(h=theoretical.mean, col = "red", lwd = 3)
legend("topright", legend = c("Monte Carlo Cumulative Mean", "Theoretical Mean"),
       lty = 1:2, lwd=c(3,3), col=c("black","red"))
```



As we would expect, the cumulative mean from our repeated experiement evaluates to `cumulative.means[1000]` =4.9742388. This value is much closer to the theoretical mean of 5 than our single experiment mean of 5.8191691.

2. Show how variable it is and compare it to the theoretical variance of the distribution.

The theretical variance of the distribution is given by the expression $Var[X] = \frac{1}{\lambda^2}$. We compare both the theoretical and empirical results below:

```r
theoretical.var = 1/(lambda^2)
experiment.var = var(experiment1)
monte.carlo.var = NULL; set.seed(1234)  # same seed, same monte carlo simulation
for (i in 1 : m) monte.carlo.var = c(monte.carlo.var, var(rexp(n, lambda)))
cumulative.var <- cumsum(monte.carlo.var) / (1 : m)
c(Theoretical = theoretical.var, Experiment1 = experiment.var,
  MonteCarlo = cumulative.var[m])
```

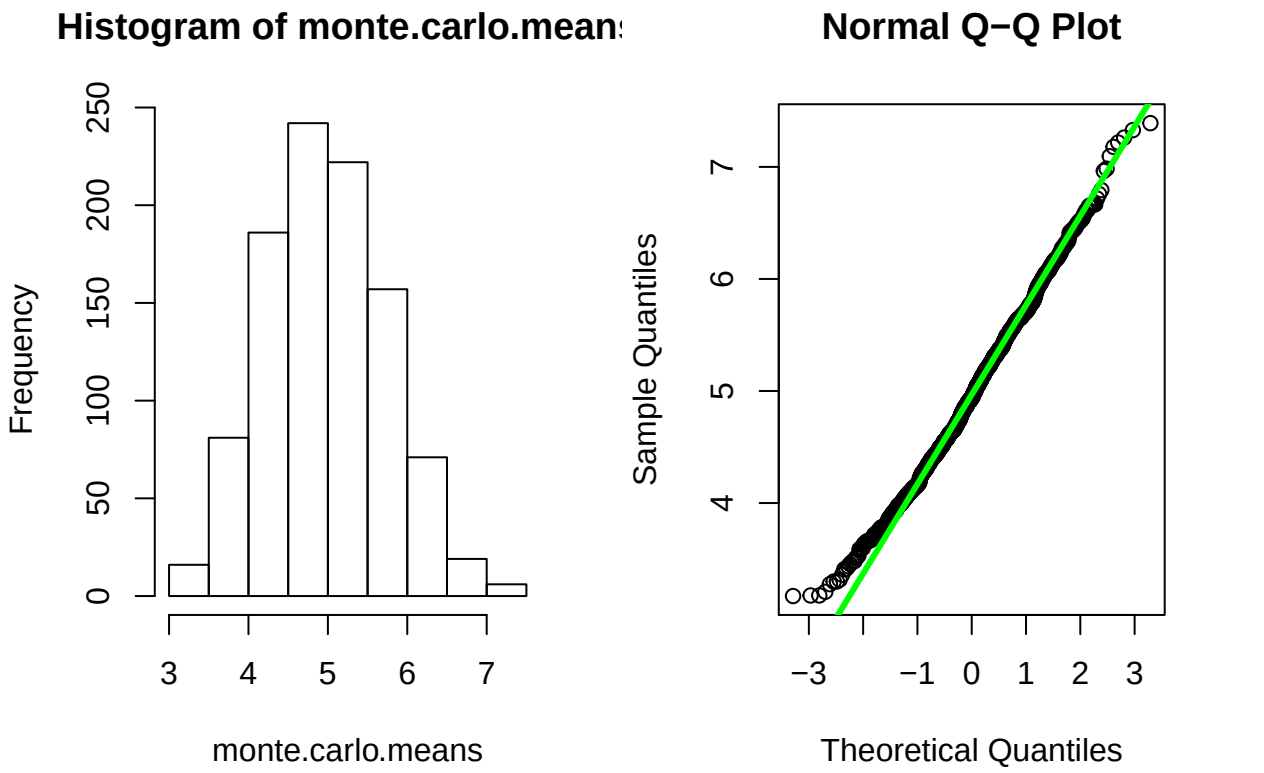
```
## Theoretical Experiment1  MonteCarlo
##    25.00000    21.41109    24.37801
```

The cumulative monte carlo variance is much closer to the theoretical value than the original experiment. We would expect this to be the case as a result of the Law of Large Numbers.

3. Show that the distribution is approximately normal.

The Central Limit Theorem (CLT) tells us that the resulting distribution of means from our 1000 experiments should tend towards normality. We can visually test this on the basis of our empirical data by plotting a histogram of the means and producing a QQ plot.

```r
par(mfrow=c(1,2))
hist(monte.carlo.means)
qqnorm(monte.carlo.means); qqline(monte.carlo.means, col="green", lwd=3)
```



Visual inspection of the histogram reveals that the distribution of experiment means resembles the normal distribution. The QQ plot supports this finding, with the green line representing the identity line for normal distribution. We attribute the small deviations from normality as monte carlo error (i.e. random noise from the simluation).