An Exploration of the Central Limit Theorem Relating to Sample Mean and Sample Variance of Random Variables from R Language's Pseudorandom Number Generator and base R's Implementation of the Exponential Distribution

Randall

Theoretical and Sample Mean of the Exponential Distribution

The exponential distribution describes the elapsed time between events in a Poisson process. Poisson processes are characterized by "memorylessness", which means that while the probability of total events having occurred accrues over time, the probability of an event occurring within any given time period is not affected by occurrences within other periods of time.

Where x is the elapsed time in units, λ is the rate, in events per unit of time. The probability density function of the exponential distribution which describes the elapsed time between events for $x \geq 0$ is $F(x;\lambda) = \lambda e^{-\lambda x}$ and its CDF is $F(x;\lambda) = 1 - e^{-\lambda x}$.

By definition, $1/\lambda$ corresponds to the expected time between events. For example, if 3600 hits are expected to a website every hour, then it is also expected that the rate will *average* to one hit per second. During this investigation $\lambda = 0.2$.

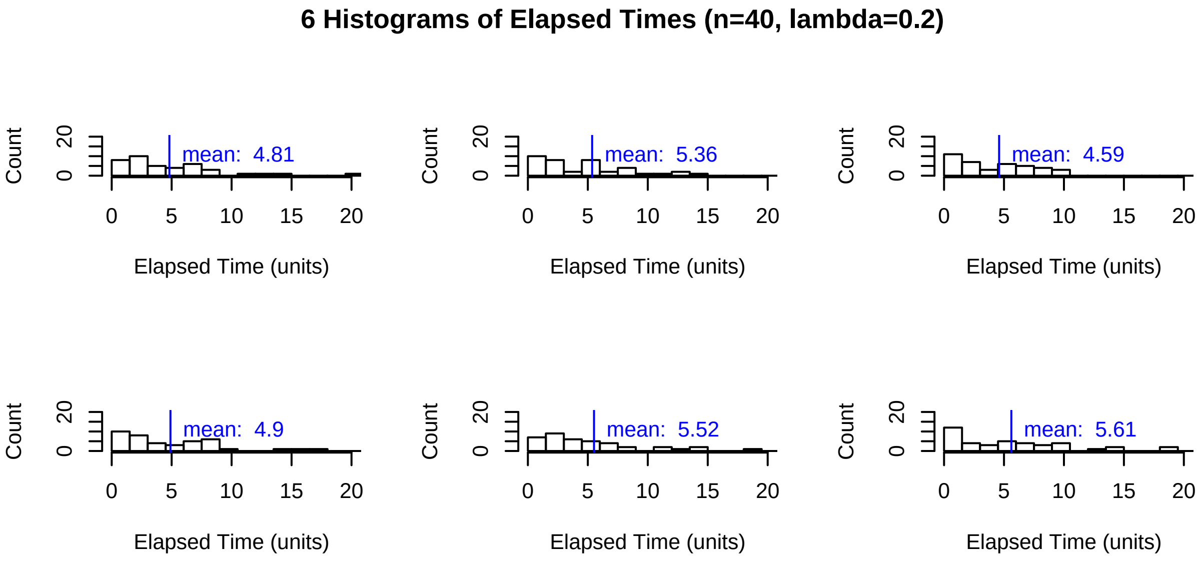
> Therefore:
> If $\lambda = 0.2 = 1/5$
> then $E[X] = 1/\lambda = 5$, the center of the distribution of elapsed times

We will explore 40 samples from the distribution, and in order to understand the Central Limit Theorem will see how *consistently* the mean begins to center around the center of this distribution.

```r
set.seed(123); lambda <- 0.2; num <- 40
exp.sample40 <- rexp(n = num, rate = lambda) # initialize accumulator
for (i in 1:999) { # accmulate additional experiments
    exp.sample40 <- cbind(exp.sample40, rexp(n = num, rate = lambda))
}
```

What do some of these look like:

```r
par(mfrow = c(2, 3), oma = c(0, 0, 1, 0))
for (i in 1:6) {
    hist(x = exp.sample40[, i], breaks = seq(from = 0, to = 80, by = 1.5), ylim = c(0, 20),
         xlim = c(0, 20), main = NULL, ylab = "Count", xlab = "Elapsed Time (units)")
    abline(v = mean(exp.sample40[, i]), col = "blue")
    text(x = mean(exp.sample40[ ,i]), y = 10, pos = 4, col = "blue", labels = paste(
        "mean: ", round(mean(exp.sample40[, i]), digits = 2)))
}
title(main = "6 Histograms of Elapsed Times (n=40, lambda=0.2)", outer = TRUE)
```



Sample and Theoretical Center of the Distribution

As stated the theoretical center of the distribution is **5**. Additionally, the 1,000 instances of 40 samples also converges to approximately 5.

```
mean(mean(exp.sample40)) #SAMPLE MEAN
```

```
## [1] 5.011911
```

Sample and Theoretical Spread of the Distribution

However, it is more interesting how consistently the data are spread out when there are only 40 samples.

Variance is the squared difference between outcomes and the mean. One way of expressing this is $\sum \frac{(\mu-x)^2}{n}$, however in practice the actual population mean is unknown, so the sample mean has to be used instead: $\sum \frac{(\overline{x}-x)^2}{n}$

```
vars <- var(exp.sample40[, 1])
for (i in 2:1000) {vars <- c(vars, var(exp.sample40[,i]))}
mean(vars) #SAMPLE VARIANCE
```

```
## [1] 24.84317
```