Mean and Variance of The Exponential Distribution

Barret Miller

Overview

In this project we will look at sample means and sample vairances from the Exponential Distribution. We'll compare these values found empirically to what theory tells us about the mean and variance for this distribuion. We know that the Exponential Distribution has mean 1/lambda and also standard deviation 1/lambda. For our study, we will use a lambda value of 0.2. Theoretically, the mean of the population should be 1/0.2, and the variance should be (1/0.2)^2.

```
sprintf("Mean:%2.2f, Variance:%2.2f", round(1/0.2,2), round((1/0.2)^2,2))
```

```
## [1] "Mean:5.00, Variance:25.00"
```

Simulations

I'll first generate a matrix, called expSamples, of 1000 rows by 40 columns. Each row will represent 1 set of 40 random samples from the exponential distribution. Each of the sets of 40 will be averaged and have their sample variance taken, using the apply function to apply the mean and var functions across the rows of the matrix. Then we will have 1000 means and 1000 variances of sets of size 40 from the exponential distribution. I'll store the mean results in a vector called meansOfFortySamps, and the variances in a vector called varsOfFortySamps. The mean and variance of these sets, themselves are different random variables, both having a mean themselves, and these will approach our theoretical mean and theoretical variance respectively.

```
expSamples = matrix(rexp(40000, rate=.2), 1000)
meansOfFortySamps = apply(expSamples, 1, FUN=mean)
varsOfFortySamps = apply(expSamples, 1, FUN=var)
```

Sample Mean vs. Theoretical Mean

Let's look at the mean and standard deviation of all of our samples, represented in the graph as random variable X. As you can see below, the mean is about 5, and so is the standard deviation, so this seems to be in line with what we know about the distribution. What you will notice is that the distribution of the samples is skewed right and doesn'tlook Gaussian or symmetrical around the mean at all. All samples are positive.

Now take a look at the mean of the means of 40 samples on the right, Stats-on-stats. Notice that the mean of the distribution of means of samples of size 40 is still about 5 (centered in the same place as the samples). However, the standard error is now much smaller, not even close to the mean. It's critical to remember here that this is the standard deviation of a new distribution, X'. This is the distribution of the sample means taken from samples of size 40. We know from theory that this distribution's mean and the sample mean should both be centered at the same place (around 5), and they clearly both are. We also know from theory that X' distribution's variance should be the population variance divided by n. The calculation below shows both. The graph gets more red and the height of the histogram is larger as more samples fall at the particular location on the x axis. The green dashed lines are just marking the sample standard deviation and the standard error.

```
suppressMessages(library(ggplot2)); suppressMessages(library(gridExtra))
m0 <- mean(as.vector(expSamples)); m1 <- mean(meansOfFortySamps)
s0 <- sd(as.vector(expSamples)); s1 <- sd(meansOfFortySamps)
sprintf("X mean:%2.2f | X' mean:%2.2f", m0, m1)
```

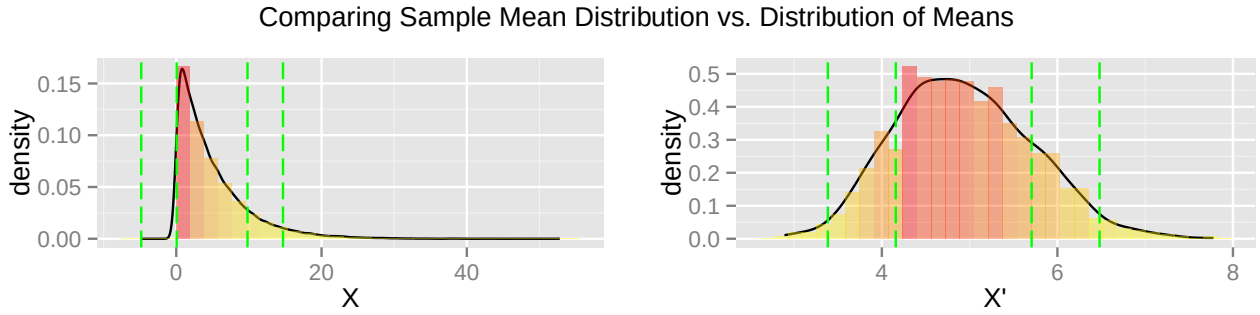
```
## [1] "X mean:4.93 | X' mean:4.93"
```

```
sprintf("Theoretical variance of X':%2.2f | Calculated variance of X':%2.2f",
        ((1/0.2)^2)/40, var(meansOfFortySamps))
```

```
## [1] "Theoretical variance of X':0.62 | Calculated variance of X':0.60"
```

```
p0 <- qplot(as.vector(expSamples), geom = "blank", xlab="X") +
   geom_line(aes(y = ..density..), stat = 'density') +
   geom_histogram(aes(y=..density.., fill = ..density..), alpha = 0.4) +
   scale_fill_gradient("Count", low = "yellow", high = "red", guide = F) +
   geom_vline(xintercept = m0+c(-1*s0, s0, -2*s0, 2*s0),
              colour="green", linetype = "longdash")
p1 <- qplot(meansOfFortySamps, geom = "blank", xlab="X'") +
   geom_line(aes(y = ..density..), stat = 'density') +
   geom_histogram(aes(y=..density.., fill = ..density..), alpha = 0.4) +
   scale_fill_gradient("Count", low = "yellow", high = "red", guide = F) +
   geom_vline(xintercept = m1+c(-1*s1, s1, -2*s1, 2*s1),
           colour="green", linetype = "longdash")

suppressMessages(grid.arrange(p0, p1, ncol=2,
             main="Comparing Sample Mean Distribution vs. Distribution of Means"))
```



Sample Variance vs. Theoretical Variance

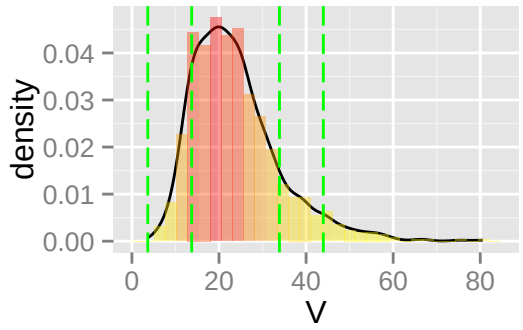
Now, lets take a look at the distribution of the sample variances for our sets of samples of size 40. For this, I'll graph the new random variable, displayed as V, that contains 1000 sample variances all calculated from sets of 40 samples taken from the Exponential distribution earlier. We know from theory that the distribution of sample variances should be centered at what it is estimating. You can see in the graph below, that it is in fact centered at 25 or (1/0.2)^2, where 0.2 was our lambda for this exponential. This also proves the sample variance taken across 40000 samples to be a good estimator of the true variance.

```
m2 <- mean(varsOfFortySamps)
v2 <- var(varsOfFortySamps)
s2 <- sd(varsOfFortySamps)
```

```
v0 <- var(as.vector(expSamples));
sprintf("Sample Variance: %2.2f | V mean:%2.2f", v0, m2)
```

```
## [1] "Sample Variance: 23.81 | V mean:23.80"
```

```
p2 <- qplot(varsOfFortySamps, geom = "blank", xlab="V") +
   geom_line(aes(y = ..density..), stat = 'density') +
   geom_histogram(aes(y=..density.., fill = ..density..), alpha = 0.4) +
   scale_fill_gradient("Count", low = "yellow", high = "red", guide = F) +
   geom_vline(xintercept = m2+c(-1*s2, s2, -2*s2, 2*s2),
              colour="green", linetype = "longdash")
suppressMessages(print(p2))
```



Distribution and Normality

One final thought is that the distribution of the sample means (X') and the distribution of the sample variances (V) are much more normal than the distribution of the samples themselves (X). You can just compare the symmetry and the Gaussian look of those two distributions (X' and V) to the distribution of the samples (X).