Statistical Inference Course Project - Part 1

Fu-Ching Yang

Monday, October 03, 2016

Overview

In this project, I'll show how to do statistical inference using exponential distribution as an example. First, I'll illustrate the exponential distribution. Second, I'll show how to approximate the population mean and variance (standard deviation). Then, I compare the exponential distribution with the Central Limit Theorem. Finally, I show how to evaluate confidence interval and perform hypothesis test.

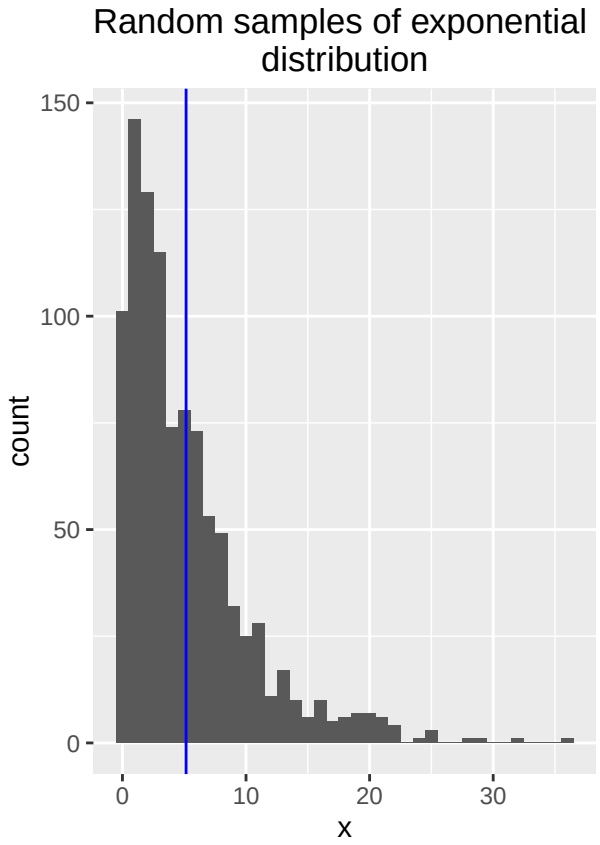
Simulations

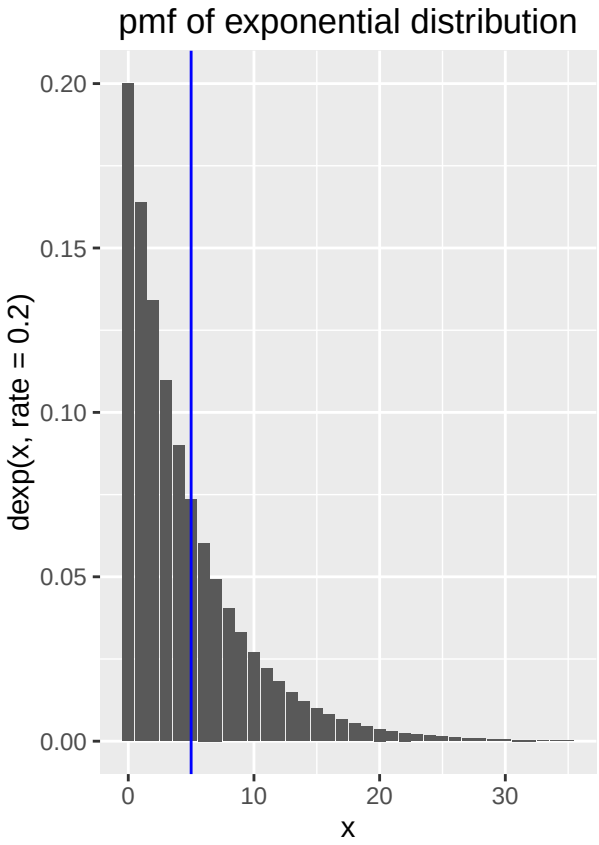
We assume the population is the exponential distribution with lambda = 0.2.

The mean and standard deviation of exponential distribution is 1/lambda, which is 5 in this case.

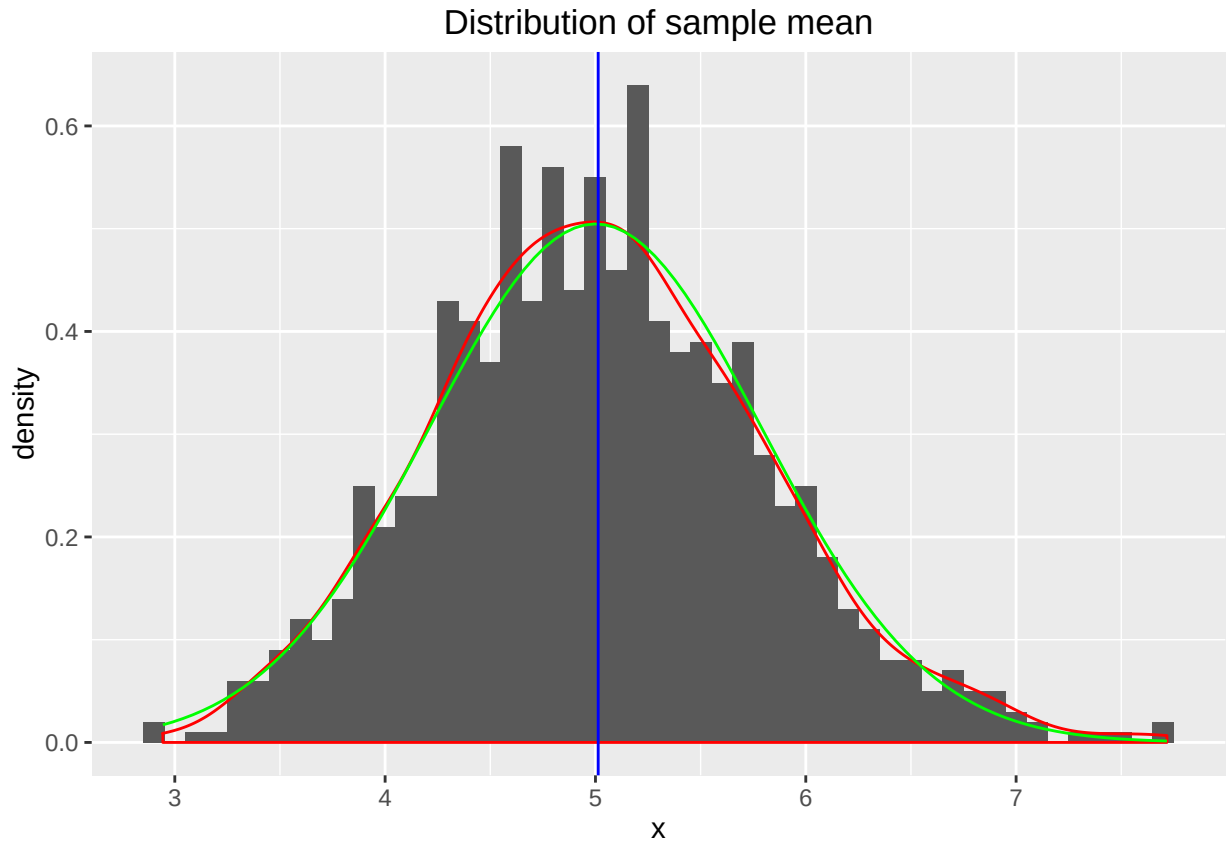
To visualize the exponential distribution, we plot the histogram of 1000 random number generated from exponential distribution and also the pmf (probability mass function).

The mean is at 5, indicated by the blue line.





Now, assume 40 samples are collected from such distribution, and the sample mean is calculated. We can do this 1000 times to observe the distribution of sample mean. Following is the plot of the sample mean distribution.



Sample Mean versus Theoretical Mean

By comparing the original exponential distribution and the sample mean distribution, you can see they are quite different. However, the sample mean approximates the theoretical mean.

From the sample mean distribution figure, although the sample mean varies, it shows up highest frequently at around 5 (blue line), which is very closed to the theoretical mean 5 (1/0.2).

```
mean(mns)
```

```
## [1] 5.012962
```

Sample Variance versus Theoretical Variance

The theoretical variance can also be approximated by the sample variance. So does the standard deviation.

To prove it, we collect 1000 sample standard deviation, each from a sample size 40. The average of these 1000 standard deviations is centered at around 5, which is very closed to the theoretical standard deviation (5=1/0.2).

```r
sds = NULL
for (i in 1 : 1000) sds = c(sds, sd(rexp(n=40, rate=0.2)))
mean(sds)
```

```
## [1] 4.870977
```

Distribution

According to CLT, the sample mean's distribution will be like Normal distribution. In the sample mean distribution that we have shown, we also plot the its density distribution in red curve. As we can see, it is approximately Normal. In fact, what it approximates is exactly (mean, sd/sqrt(n)), which is the green curve.

As you can see, they are very much alike.

Hypothesis testing

Given the sample mean and sample variance, we can perform hypothesis testing. Since we aleady knew the sample mean distribution approxmiates Normal(mean, sd/sqrt(n)), we can calculate the two-sided 95% confidence interval, with the estimated sample mean and sample variance. I use the R qnorm() function to do this for me. The 95% confidence interval is between 3.503459 and 6.522466

```r
conf_int <- qnorm(c(0.025, 0.975), mean=mean(mns), sd=mean(sds)/sqrt(40))
conf_int
```

```
## [1] 3.503459 6.522466
```

With such distribution, we can answer questions like: is the mean greater than 5, given alpha=5% in one-sided test.

The null hypothesis: mean = 5 The alternative hypothesis: mean > 5

Let's use cdf of the sampling distribution to calculate the probability greater than 5. This probability is exact p-Value.

The result shows p-Value is 49%, which is far bigger than the 5% alpha we specified. Therefore, we fail to reject the hypothesis that mean is equal to 5. In other words, the mean equals to 5, given 5% type-I error rate.

```r
pvalue <- 1-pnorm(mean(mns), mean=5, sd=mean(sds)/sqrt(40))
pvalue
```

```
## [1] 0.4932859
```