statinf

rohit jain

Sunday, September 20, 2015

Overview

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT).

The exponential distribution can be simulated in R with the `rexp(n, lambda)` function where lambda λ represents the rate parameter. The mean of an exponential distribution is $\mu = \frac{1}{\lambda}$ and the standard deviation is $\sigma = \frac{1}{\lambda}$.

According the CLT if a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean, the sampling distribution, is well approximated by a normal model which expressed in mathematical notation is: $\bar{x}_n \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

In general, a sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population. It is useful to think of a particular point estimate as being drawn from such a distribution.

In this analysis we will show that the sampling distribution of the mean of an exponential distribution with $n = 40$ observations and $\lambda = 0.2$ is indeed approximately $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$ distributed.

Comparison of the sample mean and the theoretical mean of the distribution

In the following we will draw 1000 samples of size 40 from an $Exp(\frac{1}{0.2}, \frac{1}{0.2})$ distribution. For each of the 1000 samples we will calculate the mean. Theoretically, this the same as drawing a single sample of size 1000 from the corresponding sampling distribution with $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$.

According to the CLT we would expect that each single mean of those 1000 means is already approximately $\frac{1}{\lambda} = \frac{1}{0.2} = 5$. Since we now caluculate the mean of 1000 sampled means we expect the output to be very close to 5.

We will check if this is the case.

```
set.seed(1234)

exp_sample_means <- NULL
for(i in 1:1000) {
  exp_sample_means <- c(exp_sample_means, mean(rexp(40, 0.2)))
}
mean(exp_sample_means)
```

```
## [1] 4.974239
```

$\bar{x}$ in our case is 4.97 which is very close to the mean of the theoretical distribution namely $\mu = \frac{1}{0.2} = 5$.

Comparison of the sample variance with the theoretical distribution

According to the CLT we would expect that the variance of the sample of the 1000 means is approximately $\frac{\frac{1}{0.2^2}}{40} = 0.625$.

```
var(exp_sample_means)
```

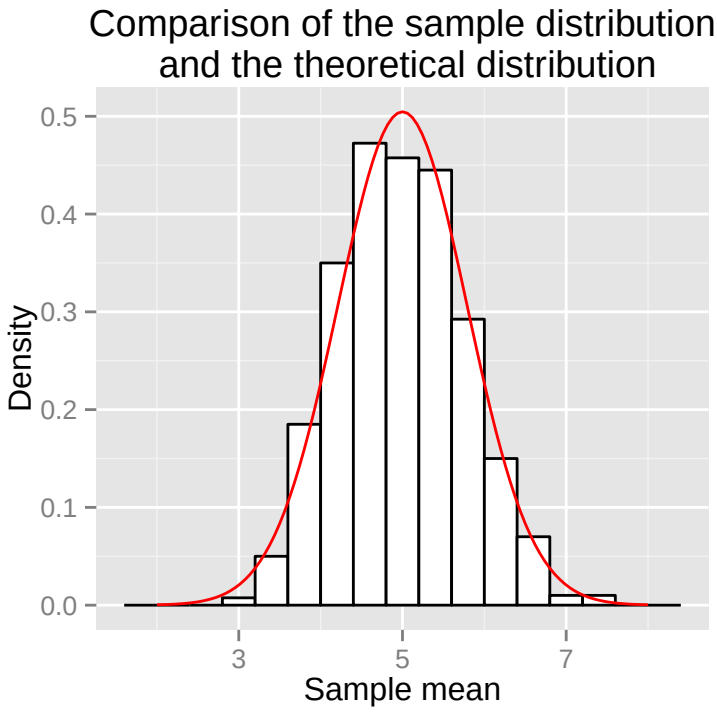
```
## [1] 0.5706551
```

s^2 in our case is 0.57 which is close to the variance of the theoretical distribution we mentioned above.

Showing that the sample distribution is approximately normal

In order to demonstrate that the sample distribution of the 1000 sampled means is approximately normal we will plot the correspondent histogram and overlay it with the density function from the theoretical sampling distribution which is $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$ distributed.

```
data <- as.data.frame(exp_sample_means)
ggplot(data, aes(x = exp_sample_means)) +
  geom_histogram(binwidth = 0.4, color = 'black', fill = 'white', aes(y = ..density..)) +
  stat_function(aes(x = c(2, 8)), fun = dnorm, color = 'red',
                args = list(mean = 5, sd = sqrt(0.625))) +
  xlab('Sample mean') +
  ylab('Density') +
  ggtitle('Comparison of the sample distribution\n and the theoretical distribution')
```



Conclusions

In this analysis we showed that the sampling distribution of the mean of an exponential distribution with $n = 40$ observations and $\lambda = 0.2$ is approximately $N(\frac{1}{0.2}, \frac{\frac{1}{0.2}}{\sqrt{40}})$ distributed.

The complete code of this report can be found on https://github.com/rohitjain2219/courserastatinf