Statistical Inference: Course Project

Part 1. Simulation exercise

Synopsis

In this simulation, we will investigate the distribution of averages of exponential deviates and illustrate via simulation and associated explanatory text the properties of the distribution of the mean of exponential deviates. We will

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

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

C. Show that the distribution is approximately normal

Distribution of Sample Mean

Let's generate 40,000 random deviates from the exponential distribution with rate lambda = 0.2 and simulate 1000 sample draws of size 40 from these deviates.

```
set.seed(1)
nosim <- 1000   # number of simulations
n <- 40         # sample size
lambda <- 0.2   # rate parameter
sample.mean <- apply(matrix(rexp(nosim * n, lambda), nosim), 1, mean)
```

sample.mean is a variable that stores the mean of a sample. sample.mean was calculated from random data; therefore, it is a random variable itself and has a distirbution with the following properties:

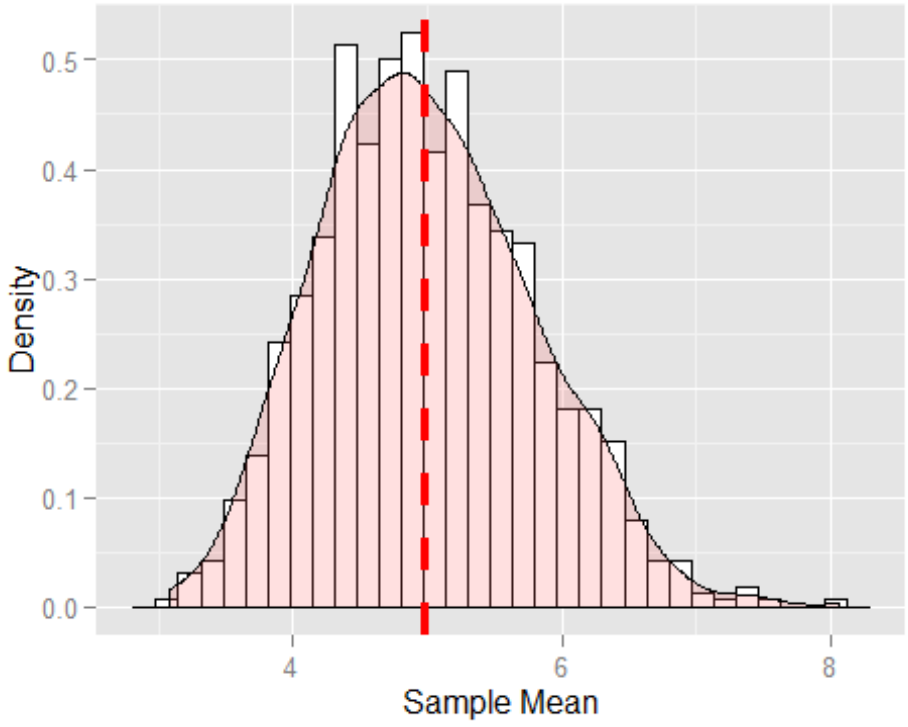
```
sample.range <- round(range(sample.mean), 2); sample.range      # range

## [1] 3.09 8.06

sample.average <- round(mean(sample.mean), 2); sample.average   # expected
value

## [1] 4.99

sample.sd <- round(sd(sample.mean), 2); sample.sd               # standard
deviation

## [1] 0.79

sample.var <- round(sample.sd ^ 2, 2); sample.var               # variance

## [1] 0.62
```

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

The mean of a distribution is a characterization of its center. Therefore, the distribution of sample.mean is centered at sample.average, 4.99, which estimates the theorectical mean of our exponential distribution, 1/lambda = 5.

Let's plot the histogram of the sample means along with the density curve and a vertical line indicating the average of sample means.

```
library(ggplot2)
df <- data.frame(x = 1:nosim, y = sample.mean)
ggplot(df, aes(x = sample.mean)) +
        geom_histogram(aes(y = ..density..), binwidth = abs(sample.range[1] -
sample.range[2])/30, color = "black", fill = "white") +
        geom_density(alpha = 0.2, fill = "#FF6666") +
        geom_vline(aes(xintercept = sample.average), color = "red", linetype
= "dashed", size = 1.5) +
        labs(y = "Density", x = "Sample Mean")
```



We see the dashed red, the average of sample.mean line at the center of the distribution.

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

Our exponential distribution has the following standard deviation and variance.

```
exp.sd = 1 / lambda; exp.sd
```

```
## [1] 5
```

```
exp.var = exp.sd ^ 2; exp.var
```

```
## [1] 25
```

The variance of sample.mean estimates the population variance. The theoretical relatioship (S^2 = sigma^2 / n) holds true where S^2 is the sample variance and sigma^2 is the population variance.

```
sample.var      # sample variance
```

```
## [1] 0.62
```

```
exp.var / n     # population variance
```

```
## [1] 0.625
```

C. Show that the distribution is approximately normal

The Central Limit Theorem (CLT) states that the distribution of means of iid samples becomes that of a standard normal as the sample size increases. From the plot, the distribution resembles a normal distribution with few outliers near the upper tail, symmetric at the average of sample.mean.

```
quantile(sample.mean)
```

```
##    0%   25%   50%   75%  100%
## 3.085 4.417 4.924 5.530 8.059
```

50th quantile of sample.mean is the same as sample.average (4.99) - a behaviour consistent with that of a normal distribution. We can also run Shapiro-Wilk test.

```
shapiro.test(sample.mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample.mean
## W = 0.9916, p-value = 1.759e-05
```

p-value is well below the 0.05; therefore, the test does verify the normality.