JHU Statistical Inference Simulation Exercise

Overview

The central limit theorem states that the sampling distribution of the mean, distribution of sample means from many samples, is nearly normal centred at the population mean, with standard error equal to the population standard deviation divided by the square root of the sample size.

This demonstration shows how taking sufficient samples from the exponential distribution will produce means whose distribution approximate the normal distribution.

Simulations

The exponential distribution has mean and standard deviation of $1/\lambda$.

We will simulate 1000 samples, each of size 40, with an exponential distribution of $\lambda = 0.2$.

```
knitr::opts_chunk$set(cache=TRUE, echo = TRUE)
library(ggplot2); library(ggthemes); library(kableExtra); library(moments)

count_sims <- 1000; lambda <-  0.2; n <- 40
```

Using `rexp`, we create the samples and then compute the average of each one.

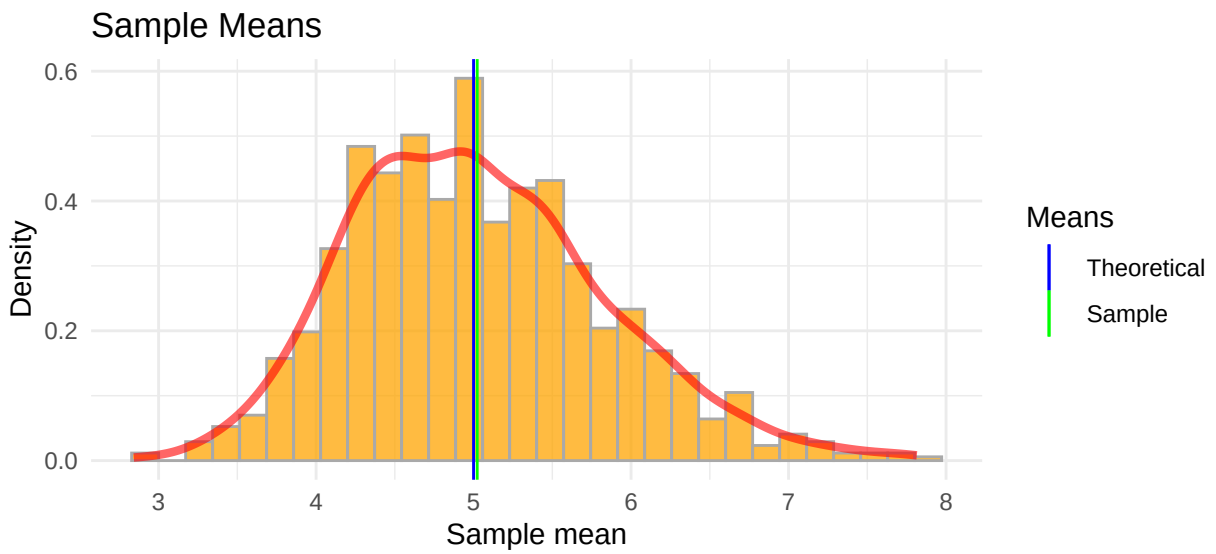
```
set.seed(123456)
simulation <- matrix(rexp(n = count_sims * n, rate = lambda), count_sims, n)
mean_sim_rows <- rowMeans(simulation)
```

Sample Mean versus Theoretical Mean

The theoretical mean of the exponential distribution is $\mu = 1/\lambda$. In this case, with $\lambda = 0.2$, the mean is $\mu = 1/0.2 = 5$.

The following code plots the distribution of sample means and calculates the overall mean of those values.

```
df_plot <- as.data.frame(mean_sim_rows)
g <- ggplot(df_plot, aes(mean_sim_rows)) + theme_minimal() +
    geom_histogram(alpha = 0.75, bins = 30, aes(y = stat(density)),
        position="identity", fill="orange", col="darkgrey") +
    geom_line(aes(y = ..density..), colour = 'red', stat = 'density', size = 1.5, alpha = 0.6) +
    geom_vline(aes(xintercept = 1/ lambda, colour="Theoretical")) +
    geom_vline(aes(xintercept = mean(mean_sim_rows), colour="Sample")) +
    ggtitle ("Sample Means ") + xlab("Sample mean") + ylab("Density") +
    scale_color_manual(name = "Means", values = c(Theoretical = "blue", Sample = "green")) +
    theme(plot.margin = margin(0, 0, 0.5, 0.5, "cm"))
df_tbl <- data.frame("Mean"=c(mean(mean_sim_rows), 1/ lambda))
rownames(df_tbl) = c("Simulation","Theoretical")
g; df_tbl
```




```
##                  Mean
## Simulation  5.022915
## Theoretical 5.000000
```

The sampled mean of 5.022915 is very close to the theoretical value of 5 while the distribution is approaching a normal Gaussian bell curve.

Sample Variance versus Theoretical Variance

Theoretical variance of an exponential distribution is given by $\sigma^2 = \frac{(1/\lambda)^2}{n} = \frac{(1/0.2)^2}{1000} = 0.625$

The sample variance can be found using the `var` function with sample means:

```
var(mean_sim_rows)
```

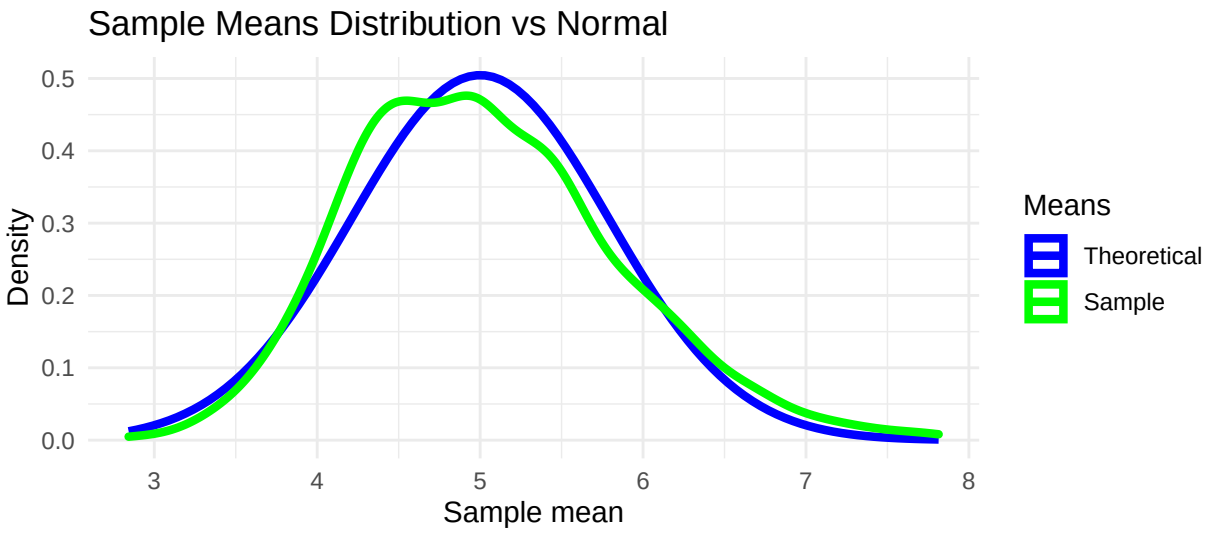
```
## [1] 0.6557463
```

Variance also shows a good approximation of the normal distribution. The small difference would decrease with greater sample size or number of simulations.

Distribution

First we can compare the density function of the sampled means with the theoretical normal distribution:

```
ggplot(df_plot, aes(mean_sim_rows)) +
    theme_minimal() +
    stat_function(fun = dnorm, aes(colour = "Theoretical"), size = 1.5,
        args = list(mean = (1/ lambda), sd = sqrt((1/ lambda)^2 / n )))+
    geom_density(aes(colour="Sample"), size = 1.5)+
    ggtitle ("Sample Means Distribution vs Normal") + xlab("Sample mean") + ylab("Density") +
    scale_color_manual(name = "Means", values = c(Theoretical = "blue", Sample = "green")) +
    theme(plot.margin = margin(0, 0, 0.5, 0.5, "cm"))
```



Overall, the distribution shows a good approximation to the normal with the following notes:

- the density function reaches maximum density to the left of the sample mean
- the sample mean distribution shows a small right skew, meaning it will be biased towards higher values, such that the mean of the distribution will exceed the median of the distribution.

We can confirm the latter point by examining the median and the skewness:

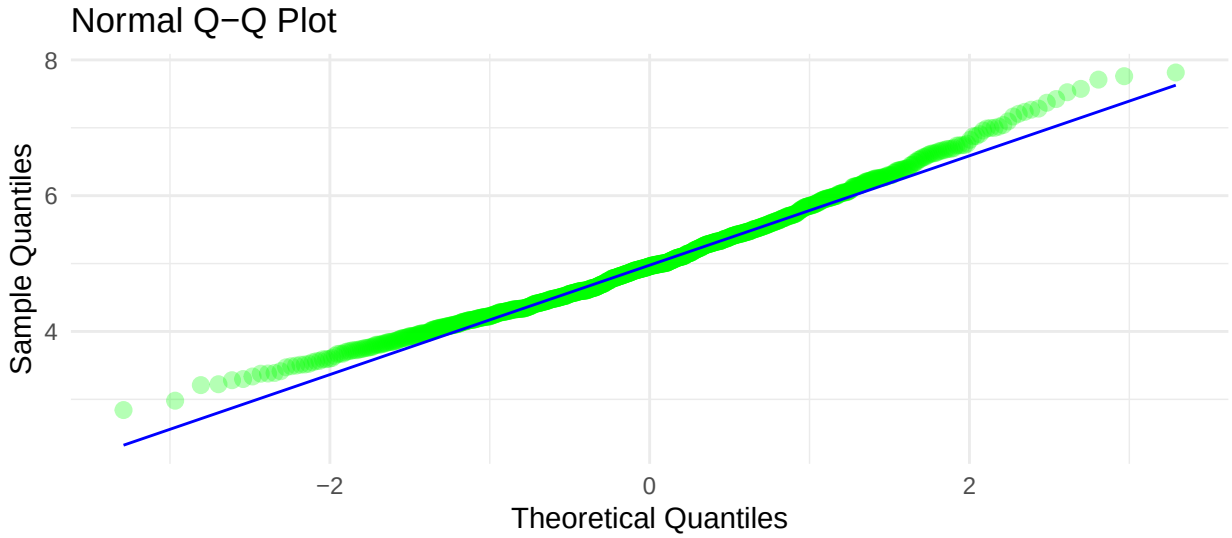
```
median(mean_sim_rows); skewness(mean_sim_rows)
```

```
## [1] 4.961684
```

```
## [1] 0.4634158
```

We can also use a QQ plot to compare the sample means with the normal distribution. Normal distribution plots will line up along the diagonal.

```
ggplot(df_plot, aes(sample = mean_sim_rows)) +
    theme_minimal() +
    ggtitle ("Normal Q-Q Plot") + xlab("Theoretical Quantiles") + ylab("Sample Quantiles") +
    stat_qq(colour = "green", alpha = 0.3, size = 2.5) + stat_qq_line(colour = "blue")
```



For the bulk of the data, the distribution lines up well with the normal distribution with the tails showing the small right skew previously mentioned.