Statistical Inference Course Project

```r
nsamples <- 1000
lambdas <- rep(0.2, nsamples)
sample_size <- 40
```

1.1 Assignment

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of the exponential distribution is:

$\mu = 1/\lambda$

and the standard deviation is

$\sigma = 1/\lambda$.

Set $\lambda = 0.2$ for all of the simulations. We will investigate the distribution of averages of 40 exponentials. Note that we will do 1000 simulations.

We will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. We will

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

In point 3, we focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 1000 exponentials.

1.2 Overview

This task demonstrates the central limit theorem, which states the distribution of the sample means follows the central limit theorem, which states:

- mean of sample means = $\bar{X} = \sigma$
- sd of sample means = $s = \sigma/\sqrt{n}$
- where n is the sample size
- and sample means are normally (gaussian) distributed regardless of the population distribution, exponential in this case

We do this by obtaining 1000 samples of sample size 40 from a population exponentially distributed with lambda (rate) of 0.2.

If we define the exponential random variables as X, and the random variables representing the sample means as Y, According to the central limit theorem, the distribution of the means of the $n_Y = 1000$ samples should be normally distributed with mean $\mu_Y = 1/\lambda$ and standard deviation $\sigma_Y = (1/\lambda)/\sqrt{n_X}$.

With $\lambda = 0.2$, this translates to $\mu_Y = 5$ and $\sigma_Y = 0.7905694$.

1.3 Simulations

Instructions: Include English explanations of the simulations you ran, with the accompanying R code. Your explanations should make clear what the R code accomplishes.

The following code generates 1000 samples of $n_X = 40$ from an exponential distribution with $\lambda = 0.2$

```
samples <- lapply(lambda, function(x) rexp(sample_size, x))
```

The following code calculates the means and standard deviations of the sample means, $\overline{Y}$, and s_Y.

```
means <- sapply(samples, mean) # vector of length 1,000 (length(means))
sample_means = mean(means) # approx 1/lambda = 1/0.2 = 5
print(sample_means)

## [1] 4.976754

sample_sds = sd(means) # approx 5/sqrt(40) = 5/6.25 = 0.8
print(sample_sds)

## [1] 0.7823584
```

These are close to the theoretical mean and standard deviation, respectively, which are calculated in the next section.

1.4 Sample Mean versus Theoretical Mean

Instructions: Include figures with titles. In the figures, highlight the means you are comparing. Include text that explains the figures and what is shown on them, and provides appropriate numbers.

```
# Function factory. DRY.
# This function allows us to create histogram/distribution overlay plots
# for any distribution. Used here for the Normal and Exponential.
plotFunction <- function(fun, pl=seq(2,8)) {

        function(histx, argslist, plotlimits=pl, col="red") {
                df <- data.frame(vals=histx)
                ggplot() +
                        geom_histogram(data = df,
                                       mapping = aes(x=vals, y=..density..)) +
                        stat_function(data = data.frame(x=plotlimits),
                                      mapping = aes(x),
                                      fun = fun, # dnorm, dchisq, dexp
                                      args = argslist, # for different dist'ns
                                      color = col)
        }
}
```

The following code generates a histogram of the sample means overlaid with a normal distribution with the theoretical mean and sd of the sample means.

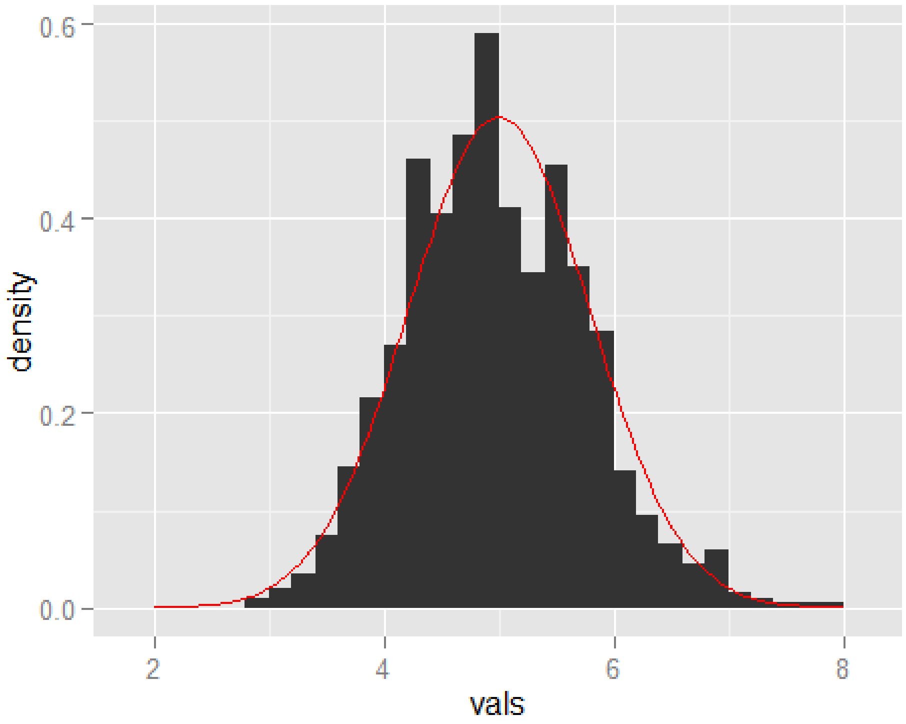
```
theoretical_mean_of_means = 1/lambdas[1]
print(theoretical_mean_of_means)
```

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

```
theoretical_sd_of_means = (1/lambdas[1])/sqrt(sample_size)
print(theoretical_sd_of_means)
```

```
## [1] 0.7905694
```

```
normPlot <- plotFunction(dnorm)
p <- normPlot(means, list(mean=theoretical_mean_of_means, sd=theoretical_sd_of_means)
)
print(p)
```



By inspection, the plot suggests the means are normally distributed with

$\mu = \overline{Y}$ and $\sigma = s_X/\sqrt{n_X}$.

The 95% confidence interval for μ_Y is

$\overline{Y} \pm z_{0.025} \times \sigma_X/\sqrt{n_X}$.

Which is:

```
mean(means) + c(-1,1) * qnorm(0.975) * sd(means)/sqrt(sample_size)
```

```
## [1] 4.734303 5.219205
```

Given this, we can see the theoretical mean $\mu_X = \mu_Y = \mu = 5$ is within the 95% confidence interval.

1.5 Sample Variance versus Theoretical Variance

Instructions: Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances.

The CLT states the mean of the means is estimated by $\overline{Y}$. It also states the sd of the means is estimated by $\sigma_Y = \sigma_X/\sqrt{n_X}$

Therefore, given $Y \sim N(\mu_Y, \sigma_Y^2)$,

$(n_Y - 1)s_Y^2/\sigma_Y^2 \sim \chi^2(n_Y - 1)$

A $(1 - \alpha)100\%$ confidence interval for σ_Y^2 is

$\frac{(n_Y-1)s_Y^2}{\chi^2(n_Y-1)_{\alpha/2}} < \sigma_Y^2 < \frac{(n_Y-1)s_Y^2}{\chi^2(n_Y-1)_{1-\alpha/2}}$.

In this case, $n_Y = 1000$. Taking the square root of both sides to convert from the variance to the standard deviation, we get the confidence interval for σ_Y:

```
alphas <- c(0.975,0.025)
sigma_ci <- sqrt((nsamples-1) * sd(means)**2/qchisq(alphas, nsamples-1))
print(sigma_ci)

## [1] 0.7495090 0.8182416
```

$$\sigma_Y = \sigma_X/\sqrt{n_X} =$$

```
1/lambdas[1]/sqrt(sample_size)

## [1] 0.7905694
```

is within this confidence interval.

Also, σ_Y^2 in this formula is equal to σ_X^2/n_X where $n_X = 40$.

Solving for σ_X by multiplying both sides by $\sqrt{40}$, the 95% confidence interval for σ_X is:

```
sqrt(40) * sigma_ci

## [1] 4.740311 5.175014
```

$\sigma_X = 5$, which is within this confidence interval.

1.6 Distribution

Instructions: Via figures and text, explain how one can tell the distribution is approximately normal.

There are statistical tests one can run to test for normality, but between the CLT and the visual appearance of the figure above where we overlaid the histogram of the data with a normal plot of $\mu = \overline{Y}$ and $\sigma = s_Y$, we can tell the distribution is approximately normal.

We can show that the quantiles are similar:

```
quantile(rnorm(n=1000, mean=5, sd=1/lambdas[1]/sqrt(sample_size)))
```

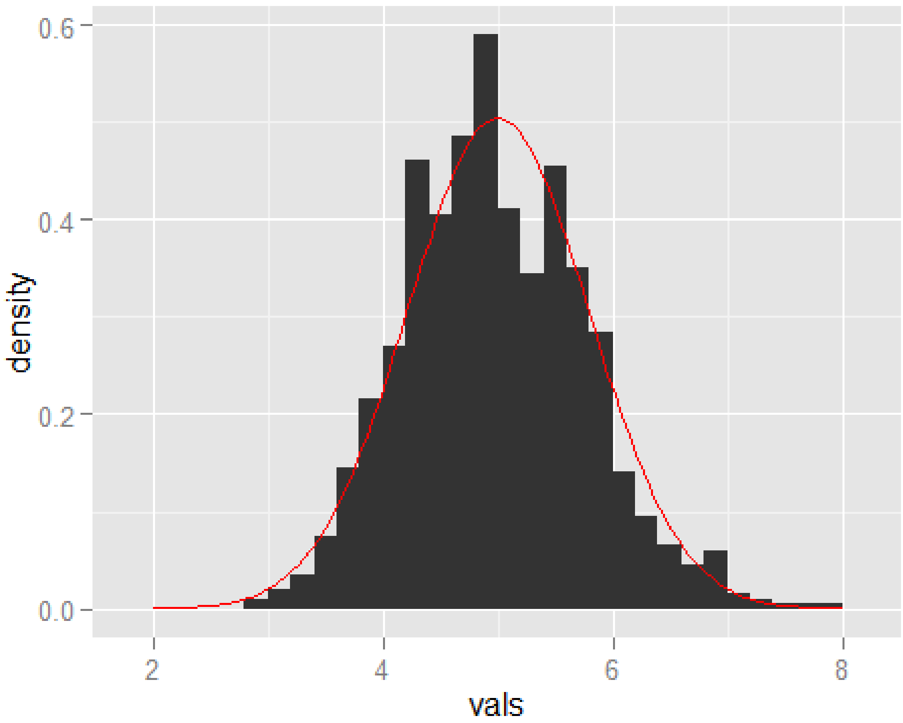
```
##        0%       25%       50%       75%      100%
## 2.404706 4.472100 4.969785 5.540411 7.546106
```

```
quantile(means)
```

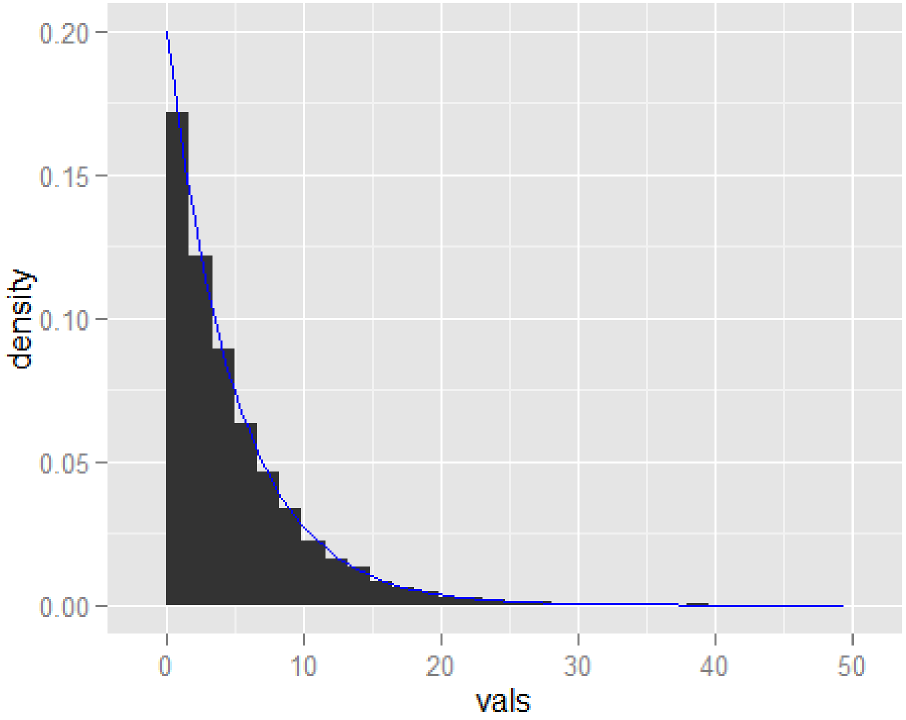
```
##        0%       25%       50%       75%      100%
## 2.924724 4.406117 4.927258 5.513350 7.843717
```

The histogram of the means overlaid with the normal plot is shown again below. This is compared with a histogram of all 40000 negative exponentials as a single sample, overlaid with an exponential distribution with $\lambda = 0.2$. This is clearly exponentially distributed.

```
print(p)
```



```
expPlot <- plotFunction(dexp, pl=seq(0,30))
s <- unlist(samples)
ep <- expPlot(s, list(rate=0.2), col="blue")
print(ep)
```


0.20
0.15
0.10
0.05
0.00
density
0
10
20
30
40
50
vals