Statistical Inference Course Project Part 1

Author: YZ

Part 1: Simulation Exercise

First part of the report is focused on the exponential distribution and its comparison to standard normal using Central Limit Theorem. Inputs provided in the instructions:

- Lambda = 0.2 (Mean = 1/lambda, Standard Deviation = 1/lambda)
- Number of randoms: 40; Number of simulations: 1,000

Step 1. Generate data

To perform the analysis we will generate a dataset with 1000 sample of 40 exponentials, take a mean of each sample and calculate cumulative mean

```
dat <- data.frame(matrix(replicate(1000, rexp(40, 0.2)), 1000))
dat$mean <- apply(dat, 1, mean)
dat$cum_exp_mean <- cumsum(dat$mean)/(1:1000)
```

Step 2. Show the sample mean and compare it to the theoretical mean of the distribution.

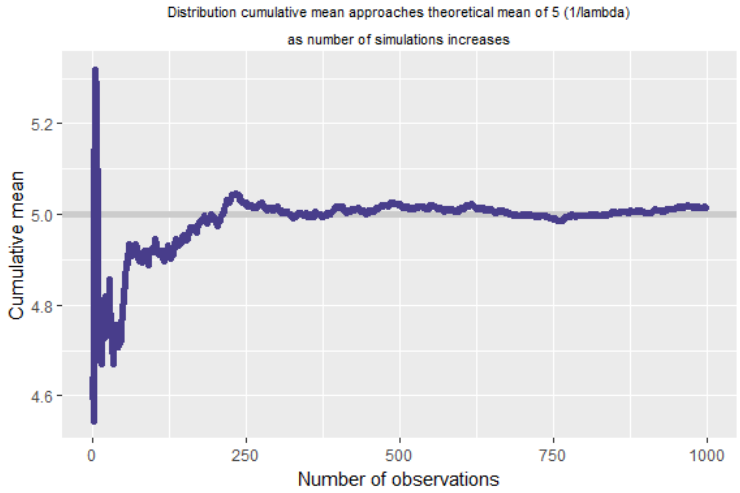
To answer this question, we will check what happens to sample mean as the number of simulations increases. In theory, sample mean should approach 5 (1/lambda) as the number of samples increases. The chart below demonstrates that well.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
g <- ggplot(data.frame(x = 1 : 1000, y = dat$cum_exp_mean), aes(x = x, y = y)) +
    geom_hline(yintercept = 5, size = 2, col = "grey80") +
    geom_line(size = 2, col = "darkslateblue") +
    labs(x = "Number of observations",
        y = "Cumulative mean",
        title = "Distribution cumulative mean approaches theoretical mean of 5 (1/lambda)",
        subtitle = "as number of simulations increases") +
        theme(plot.title = element_text(size = rel(0.75), hjust = 0.5), plot.subtitle = element_text(size = rel(0.7
5), hjust = 0.5))
g
```

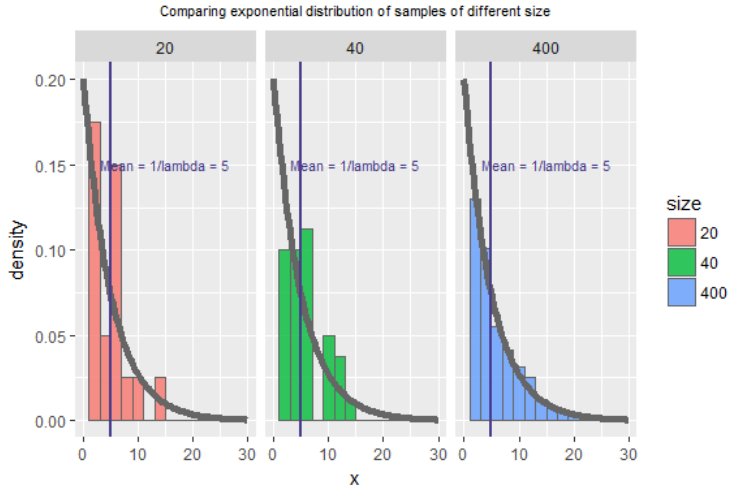


Step 3. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

To answer this question it would be helpful to compare the variance of the sample with 40 randoms to smaller (20 randoms) and larger (400 randoms) samples. As we can see in the chart below, the larger the sample the closer its variance to the theoretical distribution. 40 randoms is still small sample to clearly observe exponential distribution curve.

```
dat_compare <- data.frame(
    x = c(rexp(20, 0.2),
          rexp(40, 0.2),
          rexp(400, 0.2)),
    size = factor(rep(c(20, 40, 400), rep(c(20, 40, 400)))))

c <- ggplot(dat_compare, aes(x = x, fill = size)) +
        geom_histogram(alpha = .8, aes(y = ..density..), binwidth=2, col = "grey40")  +
        stat_function(fun = dexp, size = 2, args = (mean=0.2), col = "grey40") +
        ylim(0,0.2) +
        geom_vline(xintercept = 5, size = 1, col = "darkslateblue") +
        facet_grid(. ~ size) +
        annotate("text", x = 15, y = 0.15, size = 3, label = "Mean = 1/lambda = 5", col = "darkslateblue") +
        labs(title = "Comparing exponential distribution of samples of different size") +
        scale_x_continuous(limits = c(0,30)) +
        theme(plot.title = element_text(size = rel(0.75), hjust = 0.5))
c
```

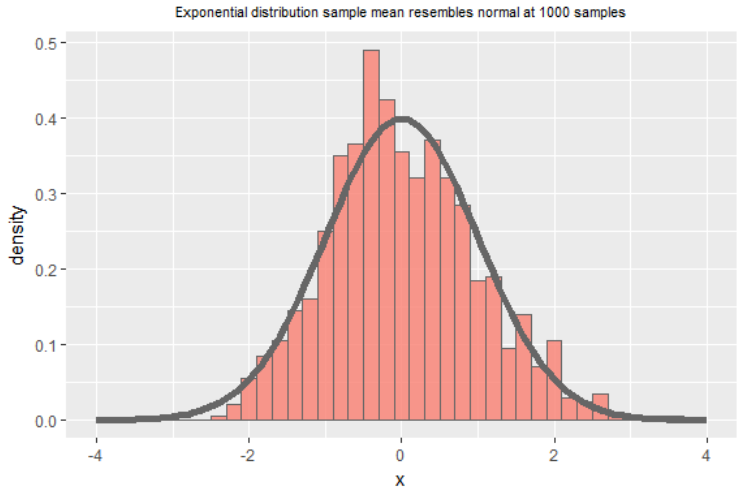


Step 4. Show that the distribution is approximately normal.

The first step is to standardize the mean of the exponential distribution. That is the purpose of cfunc. Based on the histogram below, the distribution of the mean of 1000 samples of 40 exponential randoms closely resembles standard normal distribution.

```
cfunc <- function(x, n) sqrt(n) * (mean(x) - 1/0.2) / (1/0.2)
dat2 <- data.frame(x = apply(dat, 1, cfunc, 40))

z <- ggplot(dat2, aes(x = x)) +
    geom_histogram(alpha = .8, binwidth=.2, fill = 'salmon', col = "grey40", aes(y = ..density..)) +
    stat_function(fun = dnorm, size = 2, col = "grey40") +
    xlim(-4, 4) +
    labs(title = "Exponential distribution sample mean resembles normal at 1000 samples") +
    theme(plot.title = element_text(size = rel(0.75), hjust = 0.5))
z
```



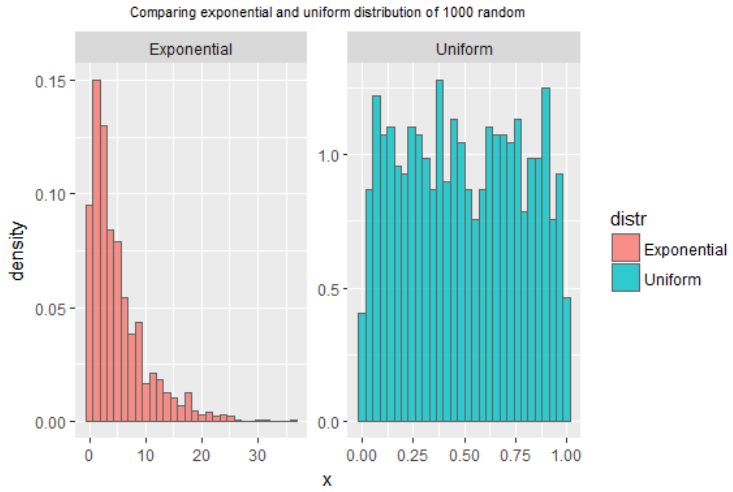
Step 5. Compare exponential to the distribution of 1000 random uniforms

For this exercise, we will use uniform distribution with a mean of 0.5 and variance of 1/12. As shown in the figure below, exponential and uniform distributions of 1000 randoms look very different from normal.

```
dat_unif <- data.frame(
    x = c(matrix(rexp(1000, 0.2), 1000),
          matrix(runif(1000), 1000)),
    distr = factor(rep(c("Exponential", "Uniform"), rep(1000, 2))))

u <- ggplot(dat_unif, aes(x = x, fill = distr)) +
    geom_histogram(alpha = .8, aes(y = ..density..), col = "grey40") +
    facet_wrap(~distr, scale = "free") +
    labs(title = "Comparing exponential and uniform distribution of 1000 random") +
    theme(plot.title = element_text(size = rel(0.75), hjust = 0.5))
u
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



However, the means of 1000 samples confirm Central Limit Theorem in both cases.

```
cfunc <- function(x, n, m, sd) sqrt(n) * (mean(x) - m) / sd
dat_unif2 <- data.frame(x = c(apply(matrix(replicate(1000, rexp(40, 0.2)), 1000), 1, cfunc, 40, 5, 5),
                              apply(matrix(replicate(1000, runif(40)), 1000), 1, cfunc, 40, 0.5, sqrt(1/12))),
                        distr = factor(rep(c("Exponential", "Uniform"), rep(1000, 2))))

z <- ggplot(dat_unif2, aes(x = x, fill = distr)) +
    geom_histogram(alpha = .8, binwidth=.2, col = "grey40", aes(y = ..density..)) +
    stat_function(fun = dnorm, size = 2, col = "grey40") +
    facet_wrap(~distr) +
    labs(title = "Central Limit Theorem in action", subtitle = "Both means distributions resemble standard normal a
t 1000 samples") +
    theme(plot.title = element_text(size = rel(0.75), hjust = 0.5), plot.subtitle = element_text(size = rel(0.75),
hjust = 0.5))
z
```