Data analysis - gamma distribution

Weronika Pyrka

TASK 1

Generate 3 samples from the Gamma distribution with parameters of your choice (different ones) for k and lambda. Compare the obtained means and variances in the samples with theoretical values. Plot histograms for each sample and compare them with the theoretical density. Also, draw theoretical cumulative distribution functions (preferably all 3 on one plot for comparison). *Create box plots for the samples (preferably all 3 on one plot).

```r
k_1 <- 4
k_2 <- 6
k_3 <- 7
lambda_1 <- 2
lambda_2 <- 3.3
lambda_3 <- 3.5

sample_1 <- rgamma(1000, k_1, lambda_1)
sample_2 <- rgamma(1000, k_2, lambda_2)
sample_3 <- rgamma(1000, k_3, lambda_3)

mean_1 <- k_1 / lambda_1
mean_2 <- k_2 / lambda_2
mean_3 <- k_3 / lambda_3

variance_1 <- k_1 / (lambda_1^2)
variance_2 <- k_2 / (lambda_2^2)
variance_3 <- k_3 / (lambda_3^2)

cat(paste0("Theoretical mean value 1 = ", mean_1, "\nSample mean 1 = ",
           mean(sample_1)), '\n')

## Theoretical mean value 1 = 2
## Sample mean 1 = 2.02277432439581

cat("\n")

cat(paste0("Theoretical mean value 2 = ", mean_2, "\nSample mean 2 = ",
           mean(sample_2)), '\n')

## Theoretical mean value 2 = 1.81818181818182
## Sample mean 2 = 1.81052679412771

cat("\n")
```

```r
cat(paste0("Theoretical mean value 3 = ", mean_3, "\nSample mean 3 = ",
           mean(sample_3)), '\n')
```

```
## Theoretical mean value 3 = 2
## Sample mean 3 = 1.97873492610495
```

```r
cat(paste0("Theoretical variance value 1 = ", variance_1, "\nSample variance
1 = ",
           var(sample_1)), '\n')
```

```
## Theoretical variance value 1 = 1
## Sample variance 1 = 1.00628266069105
```

```r
cat("\n")
```

```r
cat(paste0("Theoretical variance value 2 = ", variance_2, "\nSample variance
2 = ",
           var(sample_2)), '\n')
```

```
## Theoretical variance value 2 = 0.550964187327824
## Sample variance 2 = 0.553412283516398
```

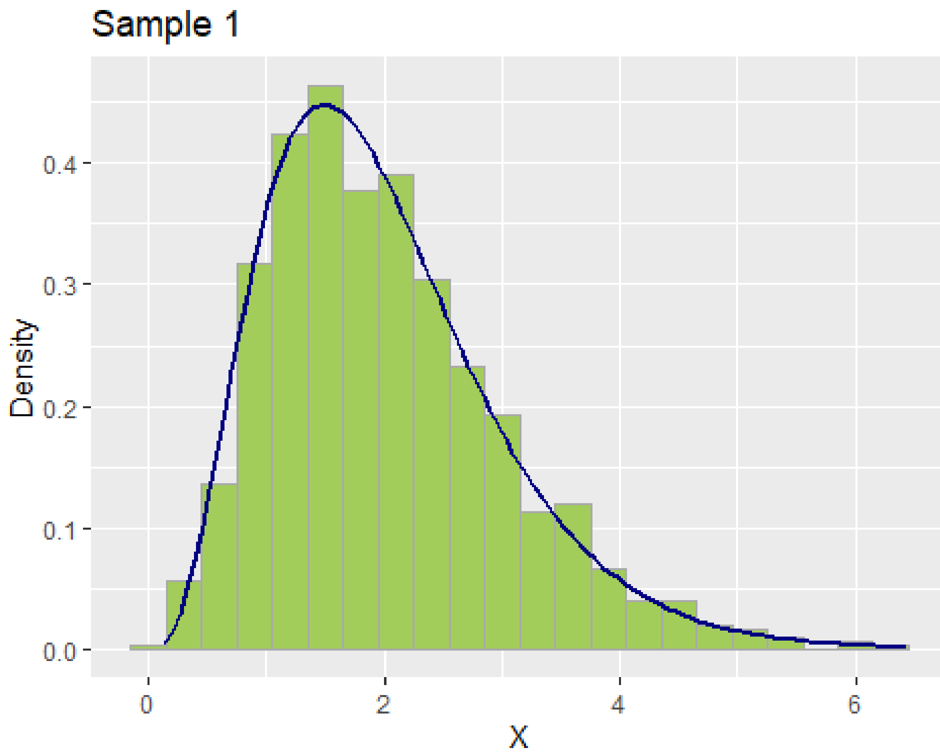
```r
cat("\n")
```

```r
cat(paste0("Theoretical variance value 3 = ", variance_3, "\nSample variance
3 = ",
           var(sample_3)), '\n')
```

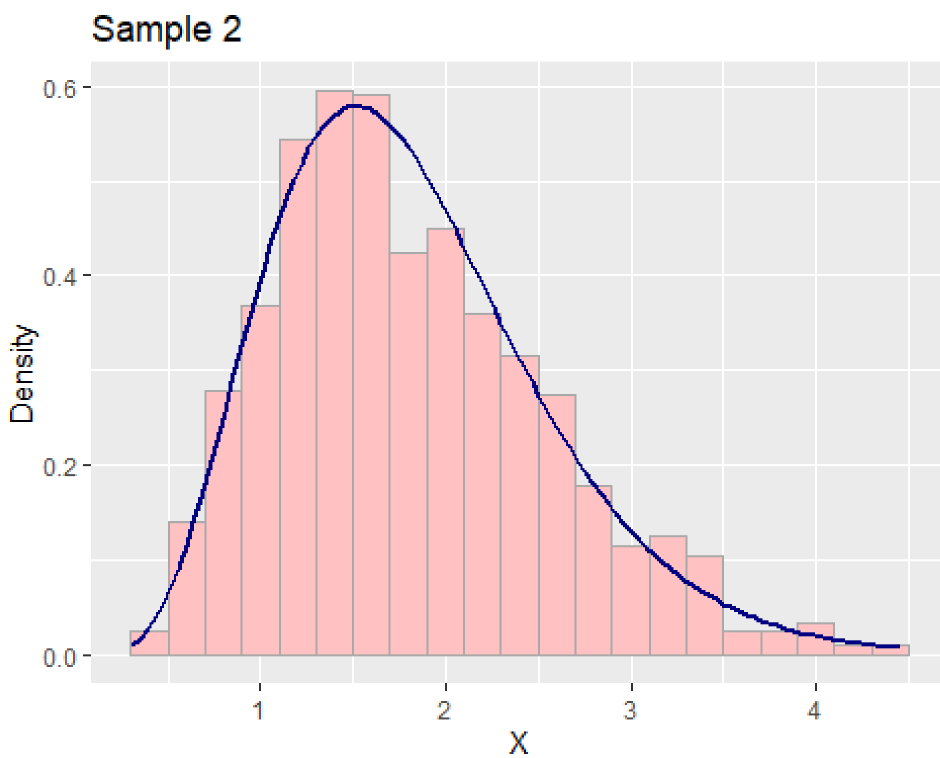
```
## Theoretical variance value 3 = 0.571428571428571
## Sample variance 3 = 0.541261450805318
```

As we can see, the obtained mean values are very close to their theoretical counterparts. The situation with variances is similar to that of means. While the obtained values are not identical to the theoretical ones, they are very close to each other. Therefore, we can infer that the mean is an appropriately chosen estimator of the expected value.

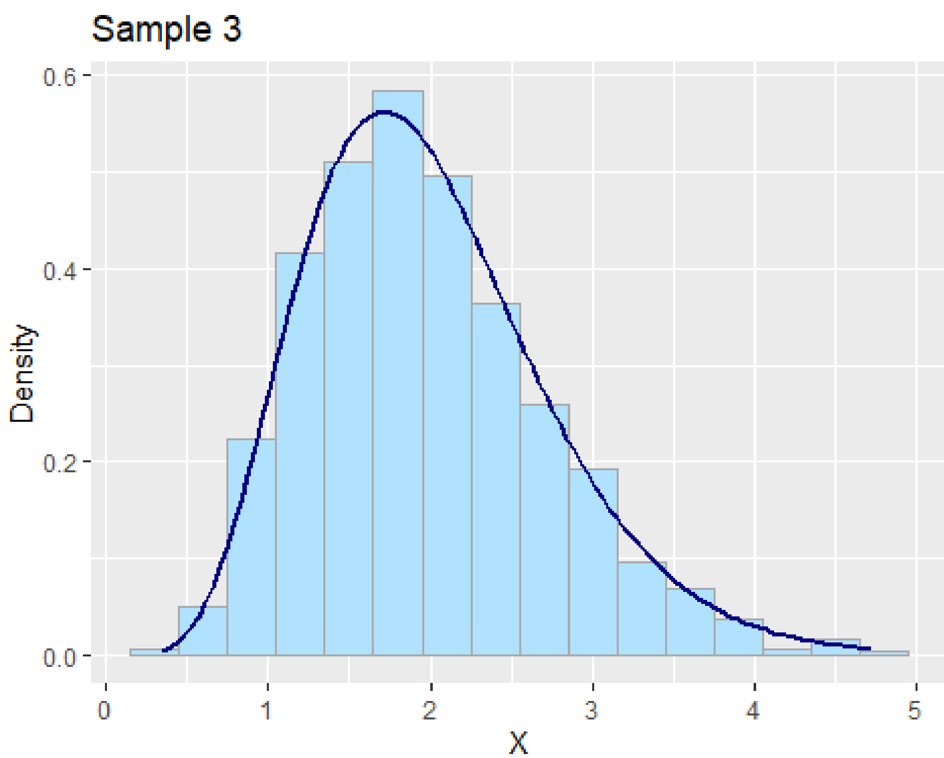
```r
frame_1 <- data.frame(x = sample_1)

plot_1 <- ggplot(frame_1, aes(x)) +
         geom_histogram(binwidth = 0.3, aes(y = after_stat(density)),
                        fill = "darkolivegreen3",color = "darkgrey") +
         stat_function(fun = dgamma, args = list(k_1, lambda_1),
                       color = "navy", lwd=1) +
         ggtitle("Sample 1") + xlab("X") + ylab("Density")

plot_1
```




```
frame_2 <- data.frame(x = sample_2)

plot_2 <- ggplot(frame_2, aes(x)) +
         geom_histogram(binwidth = 0.2, aes(y = after_stat(density)),
                        fill = "rosybrown1",color = "darkgrey") +
         stat_function(fun = dgamma, args = list(k_2, lambda_2),
                       color = "navy", lwd=1) +
         ggtitle("Sample 2") + xlab("X") + ylab("Density")
plot_2
```

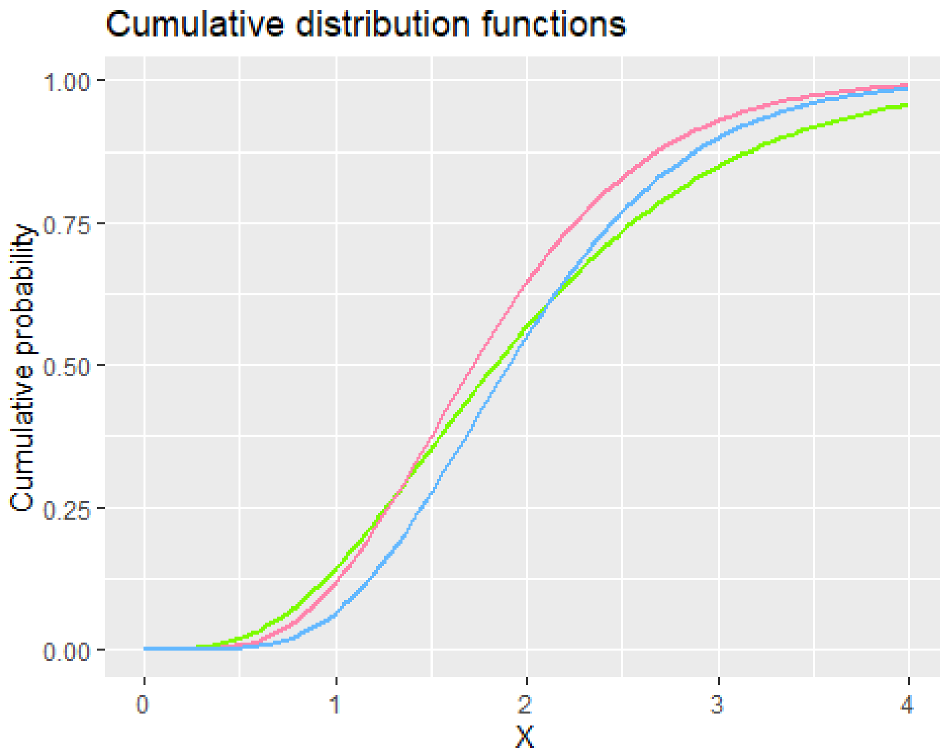



```
frame_3 <- data.frame(x = sample_3)

plot_3 <- ggplot(frame_3, aes(x)) +
         geom_histogram(binwidth = 0.3, aes(y = after_stat(density)),
                        fill = "lightskyblue1",color = "darkgrey") +
         stat_function(fun = dgamma, args = list(k_3, lambda_3),
                       color = "navy", lwd=1) +
         ggtitle("Sample 3") + xlab("X") + ylab("Density")

plot_3
```



On the plots, it is noticeable that when smaller parameters are provided, the histogram closely resembles the theoretical density (as seen in the plot for the first sample). As larger parameters are used, the plots begin to diverge more from each other (comparing, for instance, the plot of the first sample, where smaller parameters were used, to the plot of the last sample, where larger parameters were used).

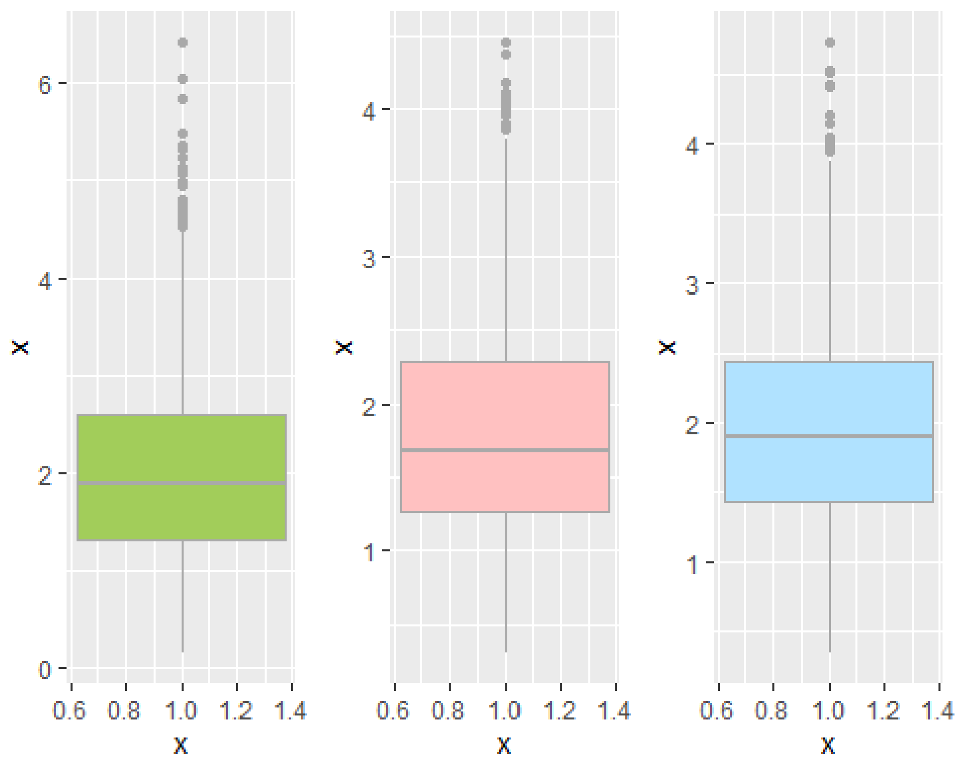
```
sequence = seq(0, 4, 1)
cumulative <- ggplot(data.frame(sequence), aes(x=sequence)) +
             stat_function(fun = pgamma, args = list(k_1, lambda_1),
                           color = "lawngreen", lwd=1) +
             stat_function(fun = pgamma, args = list(k_2, lambda_2),
                           color = "palevioletred1", lwd=1) +
             stat_function(fun = pgamma, args = list(k_3, lambda_3),
                           color = "steelblue1", lwd=1) +
             xlab("X") + ylab("Cumulative probability") +
  ggtitle("Cumulative distribution functions")

cumulative
```



From the cumulative distribution function plot, we can observe that the larger the obtained mean was, the slower the cumulative distribution function grows.

More precisely, the pink sample had the smallest mean, so the cumulative distribution function on the plot increases the fastest. On the other hand, the mean of the green sample was the largest, hence the cumulative distribution function increases the slowest.

```r
plot_box_1 <- ggplot(frame_1, aes(x = 1, y = x)) +
  geom_boxplot(fill = "darkolivegreen3", color = "darkgrey")

plot_box_2 <- ggplot(frame_2, aes(x = 1, y = x)) +
  geom_boxplot(fill = "rosybrown1", color = "darkgrey")

plot_box_3 <- ggplot(frame_3, aes(x = 1, y = x)) +
  geom_boxplot(fill = "lightskyblue1", color = "darkgrey")

grid.arrange(plot_box_1, plot_box_2, plot_box_3, ncol=3)
```

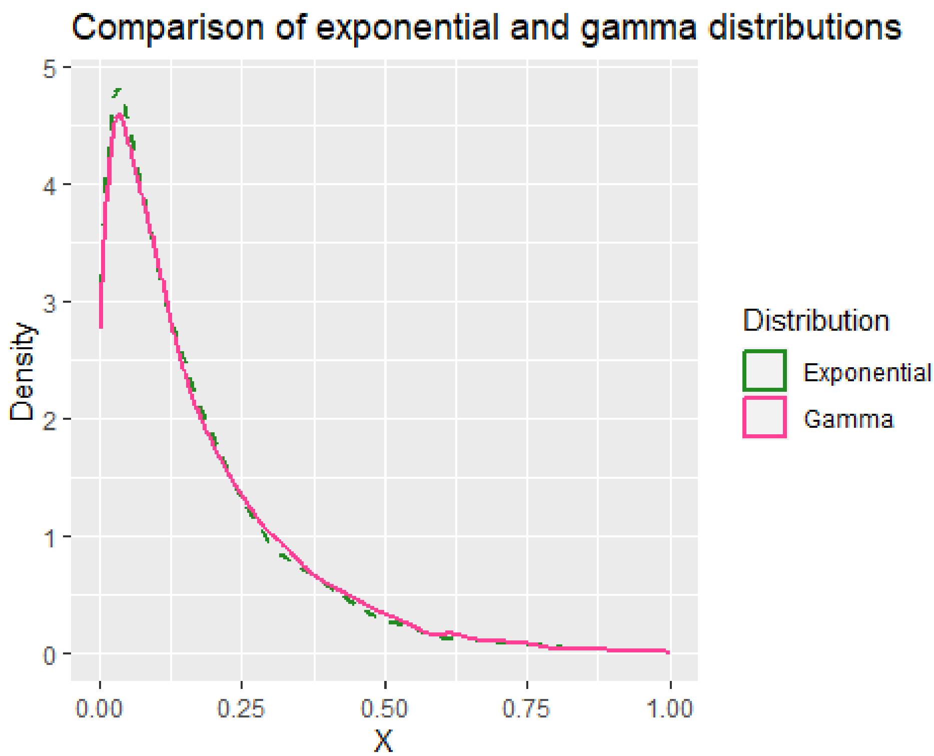


TASK 2

Empirically verify (e.g. using a density plot) that the exponential distribution with parameter lambda is a special case of the Gamma distribution with parameters lambda and k.

```
lambda <- 6
p <- 10000
k <- 1

exp_sample <- rexp(p, lambda)

gamma_sample <- rgamma(p, k, lambda)

exp_density <- data.frame(x = exp_sample, y = dexp(exp_sample, lambda))
gamma_density <- data.frame(x = gamma_sample, y = dgamma(gamma_sample, k,
lambda))

sequence = seq(0, 4, 1)
distributions <- ggplot(data.frame(sequence), aes(x=sequence)) +
         geom_density(data = exp_density, aes(x = x, color = "Exponential"),
                      alpha = 0.5, linewidth=1, linetype = "dashed") +
         geom_density(data = gamma_density, aes(x = x, color = "Gamma"),
                      alpha = 0.5, linewidth=1) +
         scale_color_manual(name="Distribution",
```

```
                            values = c("forestgreen", "violetred1")) +
         xlim(0, 1) + xlab("X") + ylab("Density") +
  ggtitle("Comparison of exponential and gamma distributions")

distributions

## Warning: Removed 23 rows containing non-finite values (`stat_density()`).

## Warning: Removed 19 rows containing non-finite values (`stat_density()`).
```



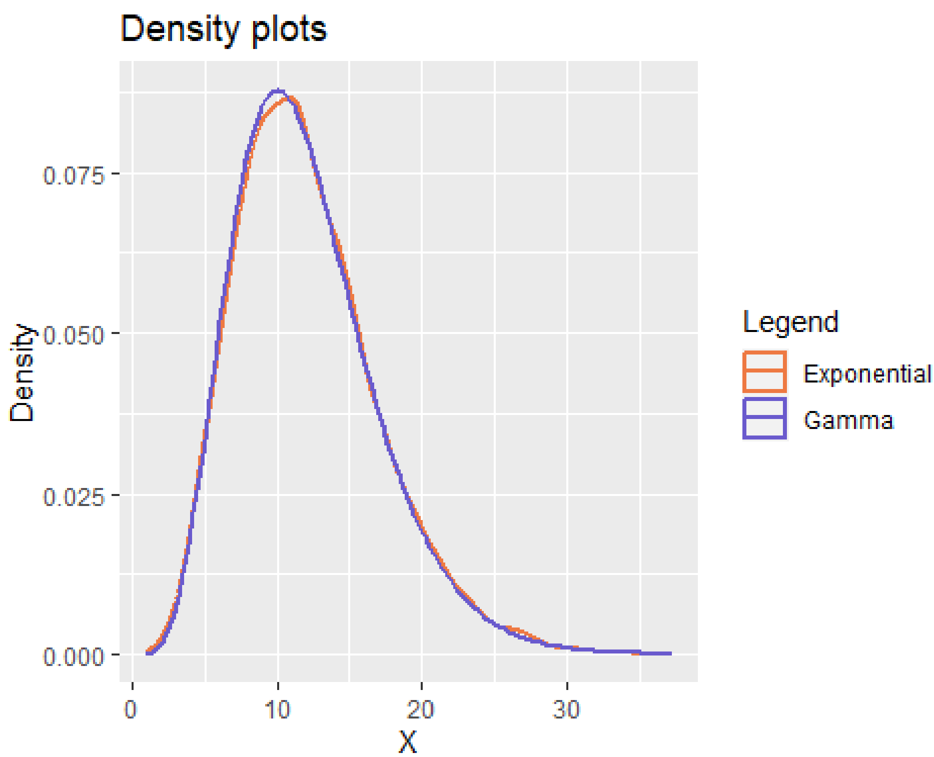
The plot clearly shows that the distributions have very similar densities, thus it can be concluded that the exponential distribution with parameter lambda is a special case of the Gamma distribution with parameters lambda and k.

TASK 3

Empirically verify (e.g. using density plots) that the sum of k independent variables with an exponential distribution with the same parameter lambda has a Gamma distribution with parameters lambda and k.

```
k <- 6
lambda <- 0.5

variable_1 <- rexp(10000, lambda)
variable_2 <- rexp(10000, lambda)
variable_3 <- rexp(10000, lambda)
```

```
variable_4 <- rexp(10000, lambda)
variable_5 <- rexp(10000, lambda)
variable_6 <- rexp(10000, lambda)

variables_sum <- variable_1 + variable_2 + variable_3 + variable_4 +
variable_5 + variable_6

plot_variables <- ggplot(data.frame(variables_sum),
                         aes(x=variables_sum)) +
           geom_density(aes(x = variables_sum,
                            color = "Exponential"), alpha = 0.8, lwd = 1) +
           stat_function(fun=dgamma, args=list(k,lambda),
                         aes(color = "Gamma"), lwd = 1)+
           scale_color_manual(name="Legend",
                              values = c("Exponential" = "sienna2", "Gamma" =
"slateblue")) +
           xlab("X") + ylab("Density") +
  ggtitle("Density plots")

plot_variables
```

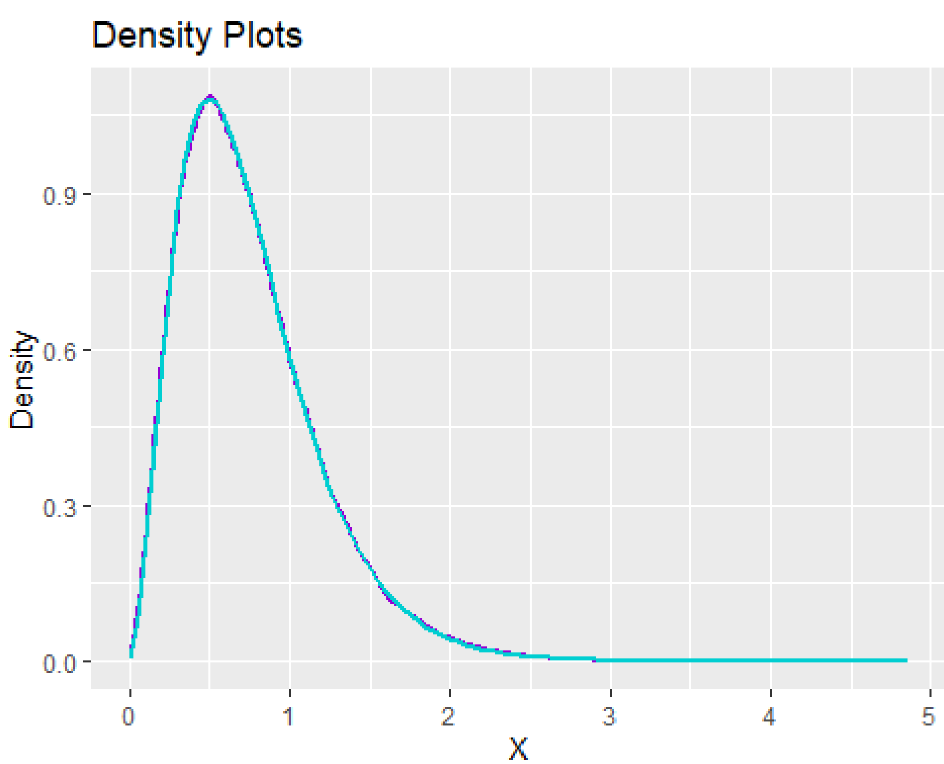


Since the density plots of the distributions have the same shape on the graph, it can be concluded that the sum of k independent variables with an exponential distribution with the same parameter lambda follows a Gamma distribution with parameters lambda and k.

TASK 4

Empirically verify (e.g. using density plots) the statement that if the variable X follows a Gamma distribution with parameters lambda and k, then the variable cX (for some c>0) follows a Gamma distribution with parameters lambda/c and k.

```
lambda <- 2
k <- 3

c <- 0.5

gamma_distribution = rgamma(100000, k, lambda)
cgamma_distribution = c * gamma_distribution

density_plot <- ggplot(data.frame(cgamma_distribution),
                       aes(x=cgamma_distribution)) +
          geom_density(color = "darkviolet", lwd=1) +
          stat_function(fun = dgamma, args = list(k, lambda/c),
                        color="darkturquoise", lwd=1) +
          xlab("X") + ylab("Density") + ggtitle("Density Plots")

density_plot
```

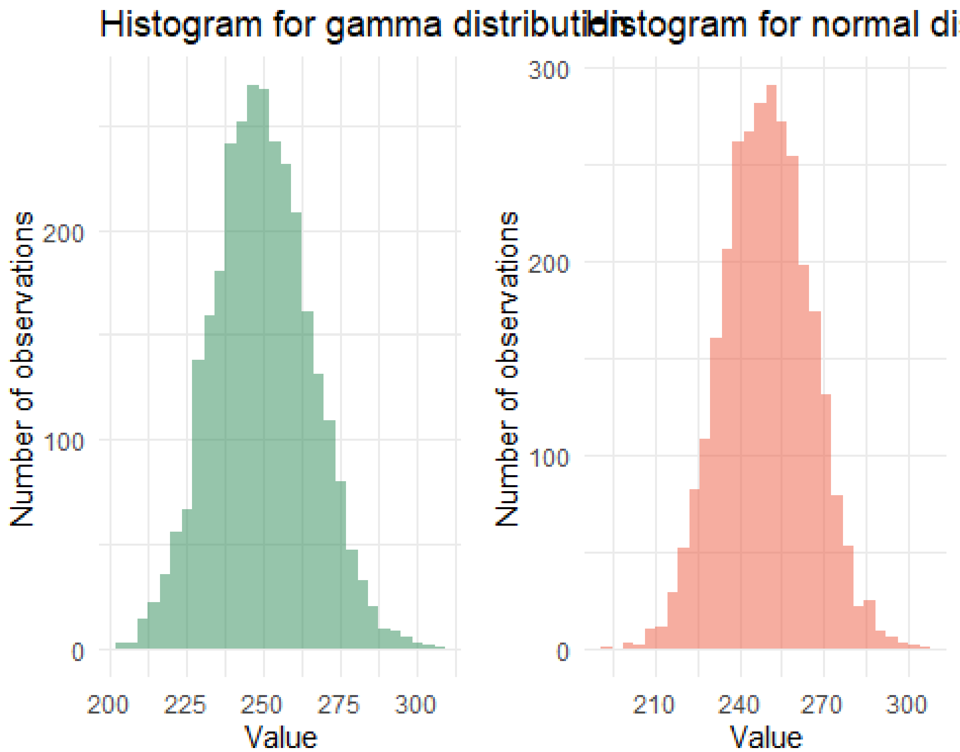


The plot shows that the Gamma distribution (purple color) and the cGamma distribution (turquoise color) have the same shape and overlap each other. Therefore, it can be inferred that the cGamma distribution is also a Gamma distribution.

TASK 5

Compare samples from a Gamma distribution with a fixed parameter lambda and large parameters k (suggested k > 50) and samples from a normal distribution with mean k/lambda and variance k/lambda^2.

```
lambda <- 1
k <- 250
n <- 3000

gamma_sample <- rgamma(n, k, lambda)
normal_sample <- rnorm(n, k/lambda, sqrt(k)/lambda)

hist_gamma <- ggplot() +
           geom_histogram(aes(gamma_sample), bins = 30,
                          alpha = 0.5, fill = "seagreen") +
           labs(x = "Value", y = "Number of observations",
                title = "Histogram for gamma distribution") +
  theme_minimal()

hist_normal <- ggplot() +
          geom_histogram(aes(normal_sample), bins = 30,
                         alpha = 0.5, fill = "tomato2") +
          labs(x = "Value", y = "Number of observations",
               title = "Histogram for normal distribution") +
  theme_minimal()

grid.arrange(hist_gamma, hist_normal, ncol = 2)
```



As the value of k increases, the gamma distribution becomes more similar to the normal distribution.