Final assignment of the Statistical Inference course. Part 1

Ivan Tiunov

In this part of the final assignmens we'll investigate the exponential distribution and find out if the Central Limit Theorem works for it. We'll draw 1000 samples of size 40 from the distribution and compair theit mean and variance distributions to the theoretical values.

Part 1. Central limit theorem

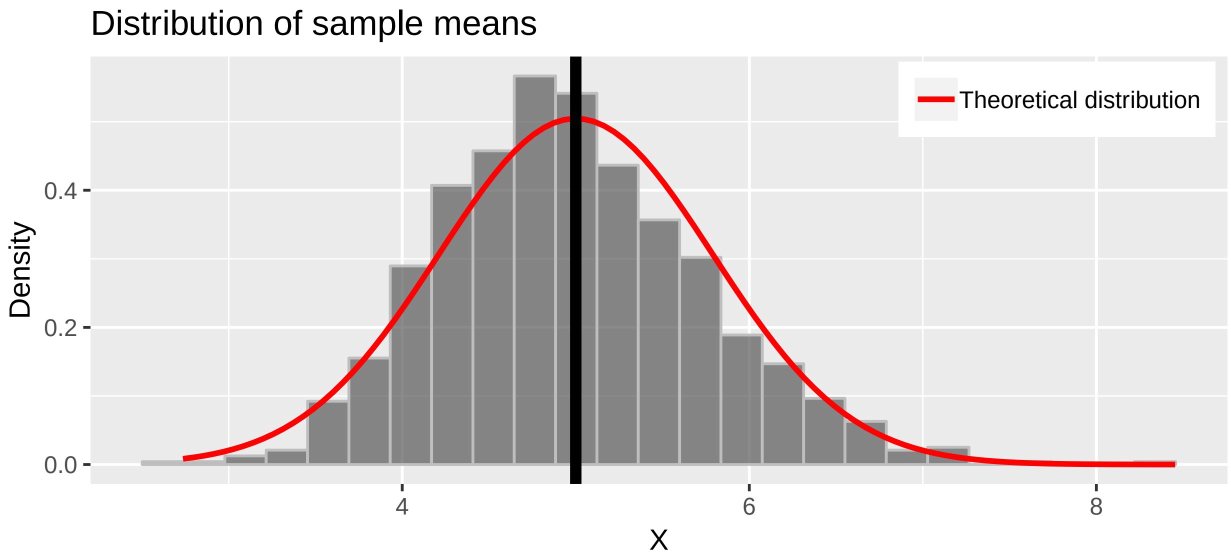
PDF of an exponential distribution is $f(x;\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$, mean and standard deviation are both $\frac{1}{\lambda}$. In our case $\lambda = 0.2$. Let's draw 1000 samples of size 40 from the distribution and calculate the mean for each sample.

```
set.seed(1024)
rate = 0.2  # Lambda
nosim <- 1000  # Number of simulations
n <- 40  # Sample size
samples <- matrix(rexp(nosim*n, rate = rate), nosim)
sample_means <- apply(samples, 1, mean)
```

According to the Central Limit Theorem, the distribution of the sample means is approximated by the normal distribution with parameters $N(1/\lambda, \frac{1}{\lambda\cdot\sqrt{n}})$, where $\lambda = 0.2$, $n = 40$. Let's calculate values of PDF for this distribution:

```
x <- seq(min(sample_means), max(sample_means), length.out = 100)
y = dnorm(x, mean = 1/rate, sd = 1/(rate*sqrt(n)))
```

Now let's plot the histogram of the sample means and compare it to the theoretical normal distribution of the means (see Supplementary section for the code):



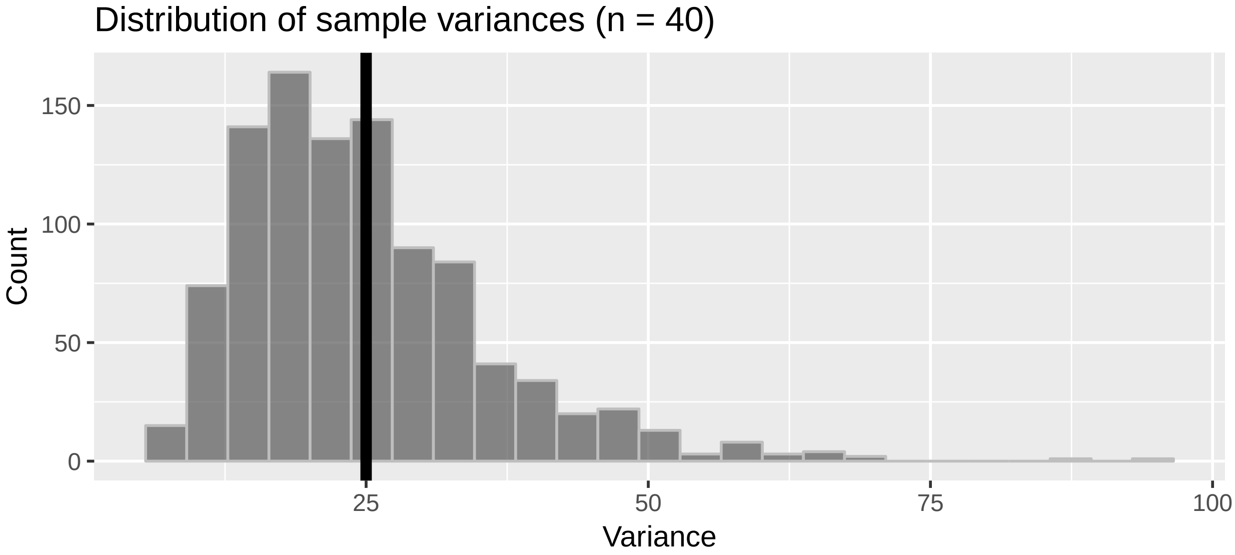
Black vertical line shows the theoretical mean. As we can see, the distribution of sample means is estimated by the theoretical normal distribution very well. The numerical values are the following (see Supplementary section for the code):

Distribution of means	Mean	SE of the mean
Sample	4.972	0.757
Theoretical	5.000	0.791

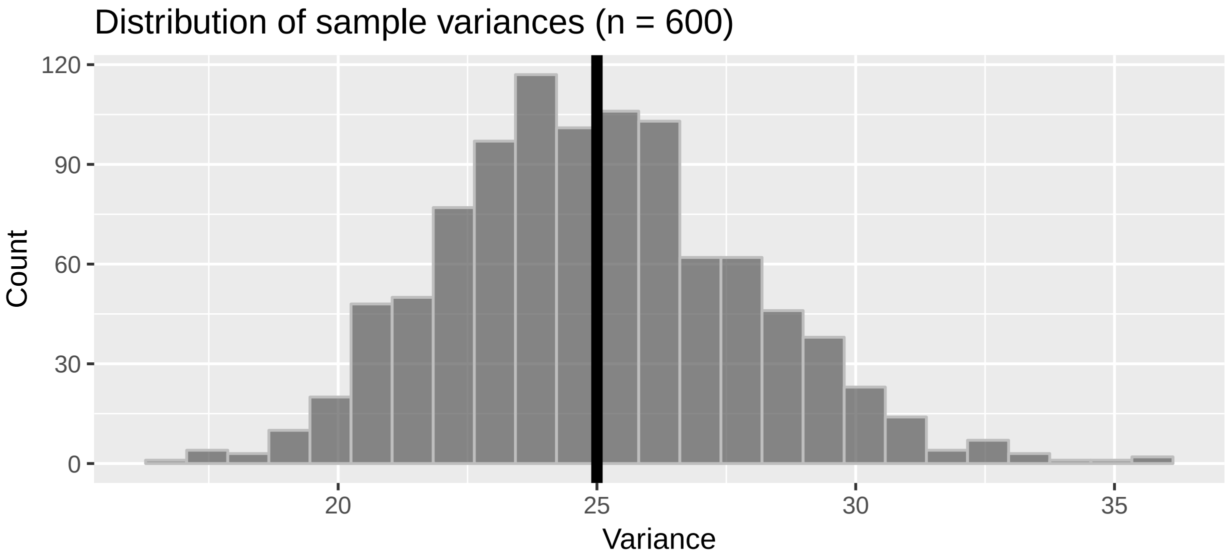
Let's also calculate variences of the samples from the exponential distribution:

```r
sample_vars <- apply(samples, 1, var)
```

And compare their distribution to the theoretical variance $(\frac{1}{\lambda})^2 = 25$ (see Supplementary section for the code):



The black vertical line shows the theoretical variance of the destribution. As we can see, the distribution of the variances is not normal. That is to be expected, because the sample size is relatively small for such an assymetric distribution as exponential, and the Central Limit Theorem only applies to the distribution of sample means. Apart from that, we can see that the sample variance slightly underestimates the theoretical variance. Again, this is due to the small sample size. As the sample size increases the sample distribution becomes more centered around the true value. For example, here is the the same picture but for the sample size of 600:



As expected, now the distribution looks more normal and correctly estimates the true variance.

Supplementary section

1. Plotting distribition of sample means

```
g <- ggplot() +
geom_histogram(aes(x=sample_means,
                   y = ..density..),
               bins = 25,
               color = "grey",
               alpha = .7) +
geom_line(aes(x = x,
              y = y,
              lty = "Theoretical distribution"),
          color = "red",
          size = 1) +
geom_vline(aes(xintercept = 1/rate),
           size = 2) +
labs(x = "X",
     y = "Density",
     title = "Distribution of sample means") +
theme(legend.title = element_blank(),
      legend.position = c(.85, .9))
```

2. Creating comparison table

```
dt <- data.frame(D = c("Sample", "Theoretical"),
                 M = c(round(mean(sample_means), 3), round(1/rate, 3)),
                 SEoM = c(round(sd(sample_means), 3), round(1/rate/sqrt(n), 3)))

kable(dt, "latex",
      booktabs = T,
      align = "c",
      col.names = c("Distribution of means", "Mean", "SE of the mean")) %>%
    kable_styling(position = "center")
```

3. Plotting distribition of variances (n = 40)

```
g <- ggplot() +
    geom_histogram(aes(x = sample_vars),
                   bins = 25,
                   color = "grey",
                   alpha = .7) +
    geom_vline(xintercept = (1/rate)^2,
               size = 2) +
    labs(x = "Variance",
         y = "Count",
         title = "Distribution of sample variances (n = 40)")
```