SimulationExercise

AlekseyVosk

03 June 2018

Overview

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. We will show the sample mean and compare it to the theoretical mean of the distribution, how variable the sample is and compare it to the theoretical variance of the distribution, and the shape of the distribution.

Import packages and set seed for simulations.

```
library(ggplot2)
library(tibble)
set.seed(100)
```

Exponential Distribution

We create list of values, associated with our exponential distribution.

```
exp_distr = list(
  lambda = 0.2,
  mean = 1/0.2,
  sd = 1/0.2,
  var = (1/0.2)^2,
  se_expected = (1/0.2)/sqrt(40)     # expected standard error of the sample mean,
)                                    # estimated for sample size 40
print(exp_distr)
```

```
## $lambda
## [1] 0.2
##
## $mean
## [1] 5
##
## $sd
## [1] 5
##
## $var
## [1] 25
##
## $se_expected
## [1] 0.7905694
```

Simulations

Let's simulate 1000 random samples (n = 40) from exponential distribution, evaluate mean and variance of each sample and store results in tibble **sim_df**. Also let's evaluet standard error of means as sd of sample means.

```r
sim_matrix <- matrix (rexp(1000*40, 0.2), ncol = 40)
sim_means <- apply(sim_matrix, 1, mean)    # means for each of 1000 simulations
sim_var <- apply(sim_matrix, 1, var)       # variances for each of 1000 simulations
sim_df <- tibble(mean = sim_means, var = sim_var)

sim_se <- sd(sim_means)                    # sample standard error of the mean
sample_mean <- mean(sim_means)             # mean of the sampling mean distribution
sample_var <- mean(sim_var)                # mean of the sampling variance distribution
```

Additionally, for comparison, let's make 1000 random values from exponential distribution, and normal distribution with mean and variance corresponding to expected parameters of sampling distribution of mean.

```r
random_exp <- tibble(random1000 = rexp(1000))
normal_scale <- seq(2, 8, 0.05)
normal_value <- sapply(normal_scale, dnorm, exp_distr$mean, exp_distr$se_expected)
normal_df <- tibble(scale = normal_scale, value = normal_value)
```

Exploration

Comparison of sample mean and theoretical mean of distribution

First we evaluate relative difference between this parameters.

```
## [1] "theoretical mean" "5"
```

```
## [1] "sample mean"     "4.9997019268744"
```

```
## [1] "relative difference"  "5.96146251199414e-05"
```

We also can compare standart error of the mean of simulated sampling distribution and expected value of it.

```
## [1] "theoretical standart error" "0.790569415042095"
```

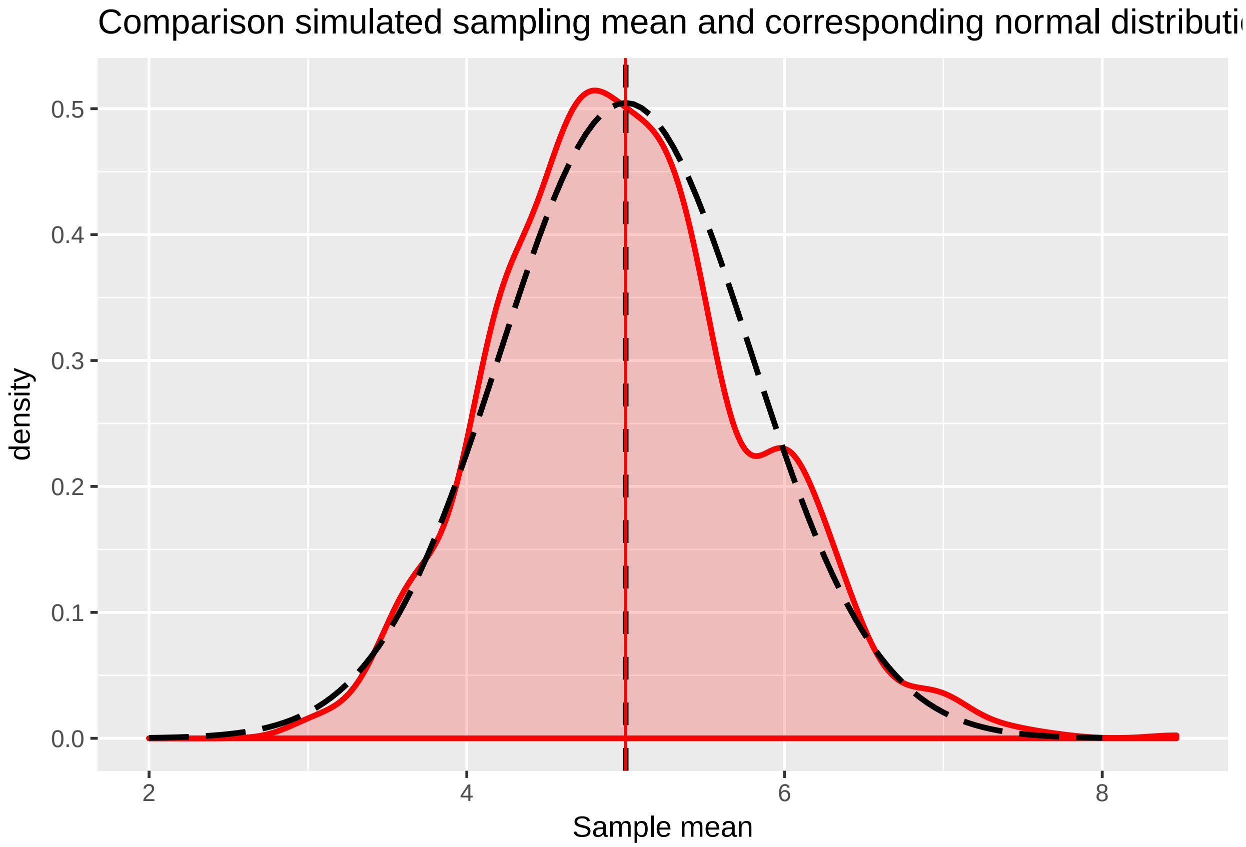
```
## [1] "sample standart error" "0.795946089646085"
```

```
## [1] "relative difference"  "-0.00680101519447738"
```

We can see that both sample mean and standart error are very close to there theoretical values.

To illustrate samplin distribution, let's plot it (usind smoothed version of the histogram) and compare it to the corresponding normal distribution.

```r
g1 <- ggplot() +
  geom_density(data = sim_df, mapping = aes(x = mean),
               colour = "red", fill = "red", alpha = 0.2, size = 1) +
  geom_line(data = normal_df, mapping = aes(x = scale, y = value),
            colour = "black", linetype = 5, size = 1) +
  geom_vline(aes(xintercept = exp_distr$mean), linetype = 2, size = 1) +
  geom_vline(aes(xintercept = sample_mean), linetype = 1, size = 0.5,colour = "red") +
  labs(x = "Sample mean",
       title = "Comparison simulated sampling mean and corresponding normal distributions")
print(g1)
```



Soild red line represents samplin distribution of mean, whereas black dashed line represents normal distribution. This lines are very close even for sample size 40, and we may assume, that normal distribution is good approximation for sampling distribution.

Comparison of sample variance and theoretical variance of distribution

Let's evaluaete relative difference between sample variance and theoretical variance.

```
## [1] "theoretical variance" "25"
```

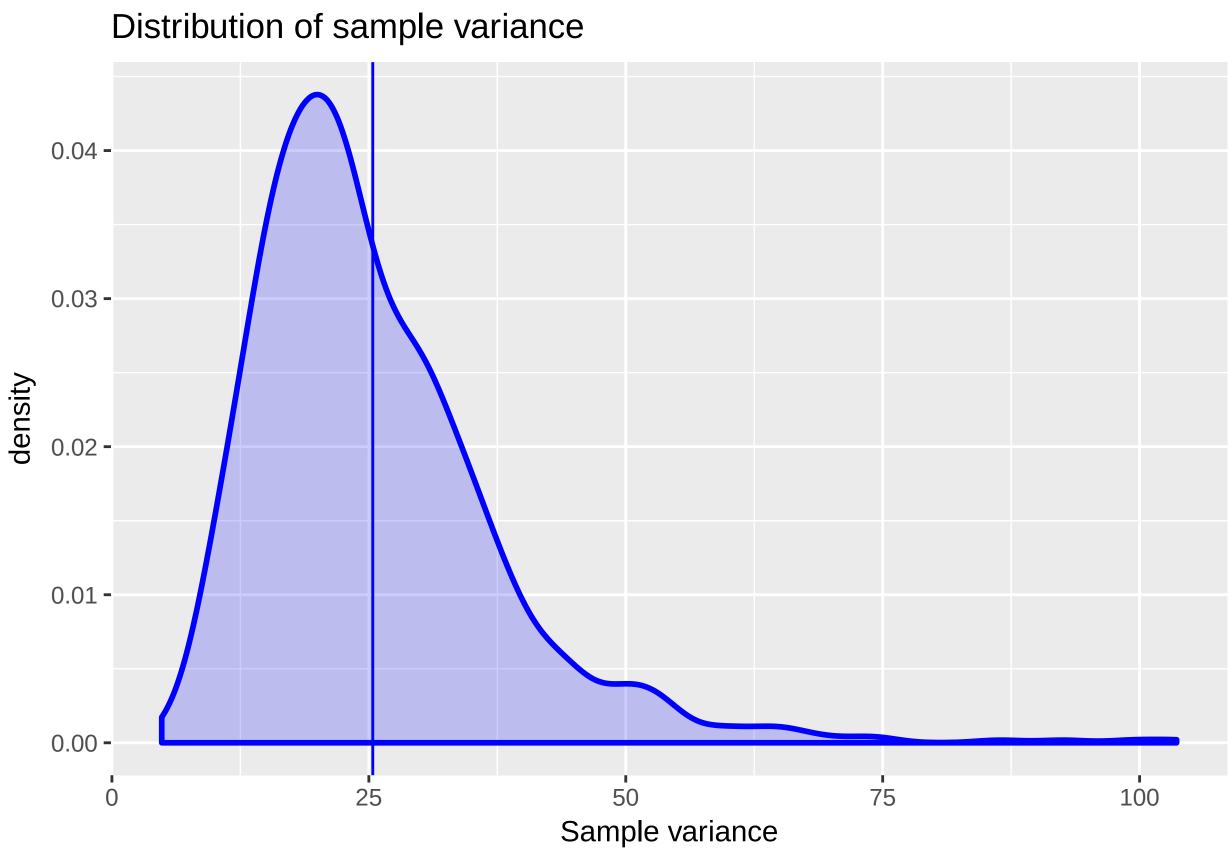
```
## [1] "sample variance"  "25.3828234274576"
```

```
## [1] "relative difference" "-0.0153129370983031"
```

There is less then 2% difference between sample and theoretical variance, so they are close.

Now let's plot sample variance distribution.

```
g2 <- ggplot() +
  geom_density(data = sim_df, mapping = aes(x = var),
               colour = "blue", fill = "blue", alpha = 0.2, size = 1) +
  geom_vline(aes(xintercept = sample_var),
             linetype = 1, size = 0.5,colour = "blue") +
  labs(x = "Sample variance",
       title = "Distribution of sample variance")
print(g2)
```

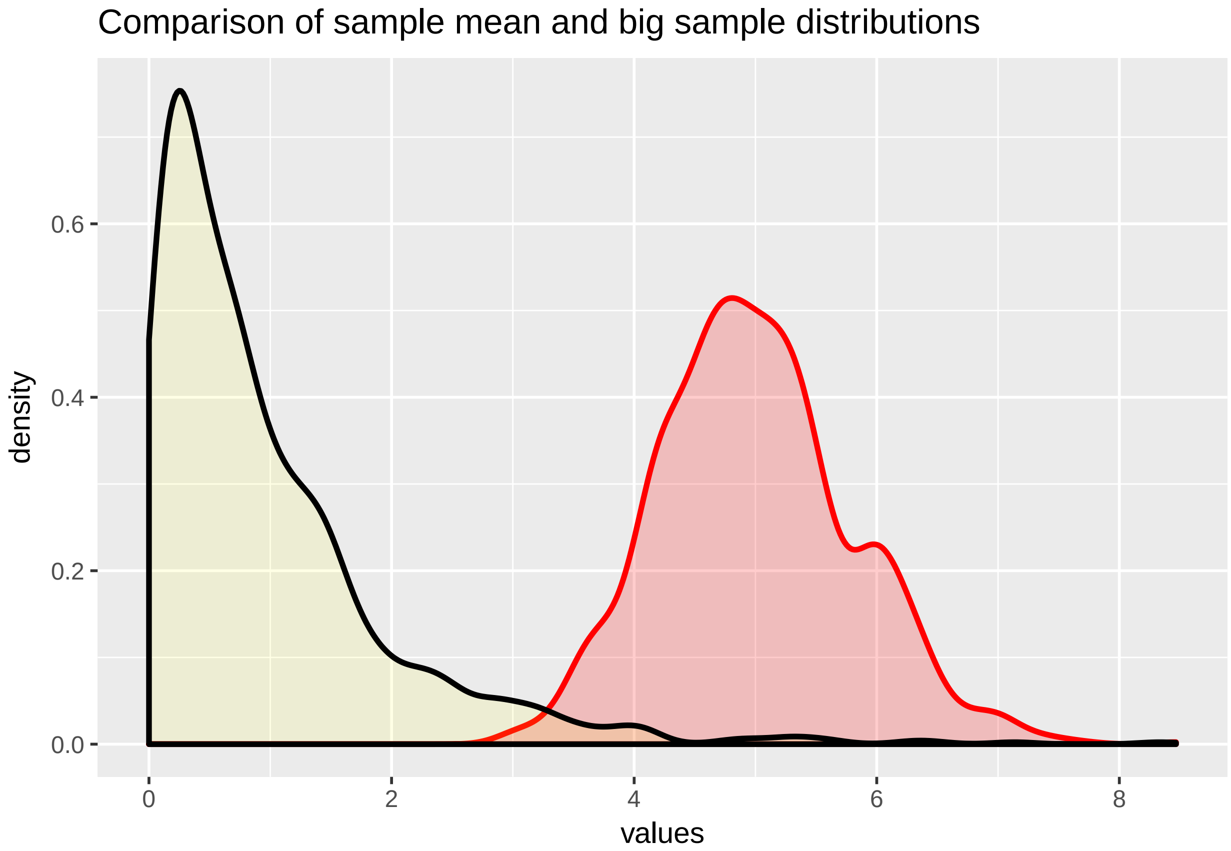


Sample variance distribution still looks bell-shaped, however it's right-skewed.

Comparison of sample mean and big sample distributions

Now let's compare our sample mean distribution and distribution of 1000 randomly generated values from exponential distribution.

```
g3 <- ggplot() +
  geom_density(data = sim_df, mapping = aes(x = mean),
               colour = "red", fill = "red", alpha = 0.2, size = 1) +
  geom_density(data = random_exp, mapping = aes(x = random1000),
               fill = "yellow", alpha = 0.1, size = 1) +
  labs(x = "values",
       title = "Comparison of sample mean and big sample distributions")
print(g3)
```



As we can see, while our sampling distribution (red on the plot) looks nearly normal, distribution of big sample (n=1000) of random exponentials (yellow on the plot) is highly skewed and isn't approximated by normal distribution.

Conclusions

Distributions of 1000 sample means and variances from exponential distribution of size 40 look nearly normal, and there means are close to the theoretical exponential distribution mean and variance.