Statistical Inference Course Project - Part 1: Simulations

Alejandro Osorio

March 28, 2018

General Parameters

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
set.seed(1)
nosim <- 1000
n <- 40
lambda <- 0.2
theor_values <- 1/lambda
```

Problem 1

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

First, we generate the sample of the 1.000 averages of 40 exponentials, as a 1.000 x 2 data frame with variables "mean" (averages of 40 exponentials) and "type" (in this case all with value 1, which will later be converted into a factor):

```
sample <- matrix(rexp(nosim * n, rate = lambda), nosim) %>%
        apply(1, mean) %>%
        as_data_frame() %>%
        mutate(type = 1)
names(sample) <- c("mean", "type")
```

Applying the head function, the obtained data frame looks like this:

```
head(sample)
```

```
## # A tibble: 6 x 2
##    mean  type
##   <dbl> <dbl>
## 1  4.90  1.00
## 2  5.23  1.00
## 3  6.40  1.00
## 4  4.74  1.00
## 5  5.18  1.00
## 6  5.17  1.00
```

The sample mean is obtained as follows:

Sample mean:

```
sample_mean <- mean(sample$mean)
```

Finally, a comparison between the sample mean and the theoretical mean of the function:

Sample Mean:

```
## [1] 4.990025
```

Theroretical mean of the function:

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

As we can see, the sample mean is quite close to the theoretical mean of the function, as expected.

Problem 2

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

Sample variance and standard deviation, obtained with R's functions:

```
sample_var_R <- var(sample$mean)
sample_sd_R <- sqrt(sample_var_R)
```

Sample theoretical variance and standard deviation were obtained as follows:

```
sample_var_theor <- theor_values^2/n
sample_sd_theor <- sqrt(sample_var_theor)
```

Finally, a comparison between the sample variance and the theoretical variance of the function:

Sample Variance:

```
## [1] 0.6177072
```

Theroretical variance of the function:

```
## [1] 0.625
```

Hence, the sample variance is quite close to the theoretical variance of the function, as expected.

Problem 3

Show that the distribution is approximately normal.

In order to compare the distribution of the samples of 40 averages, with the exponential distribution (as suggested), the following second sample was generated, with 1.000 exponential values:

```
sample2 <- c(rexp(nosim, rate = lambda)) %>%
        as_data_frame() %>%
        mutate(type = 2)
names(sample2) <- c("mean", "type")
```

Finally, both samples were joined into one unified dataset, with the "type" variable converted into a factor in order to differentiate both samples, as follows:

```
samples <- bind_rows(sample, sample2)
samples$type <- factor(samples$type, labels = c("40 means", "exps"))
```

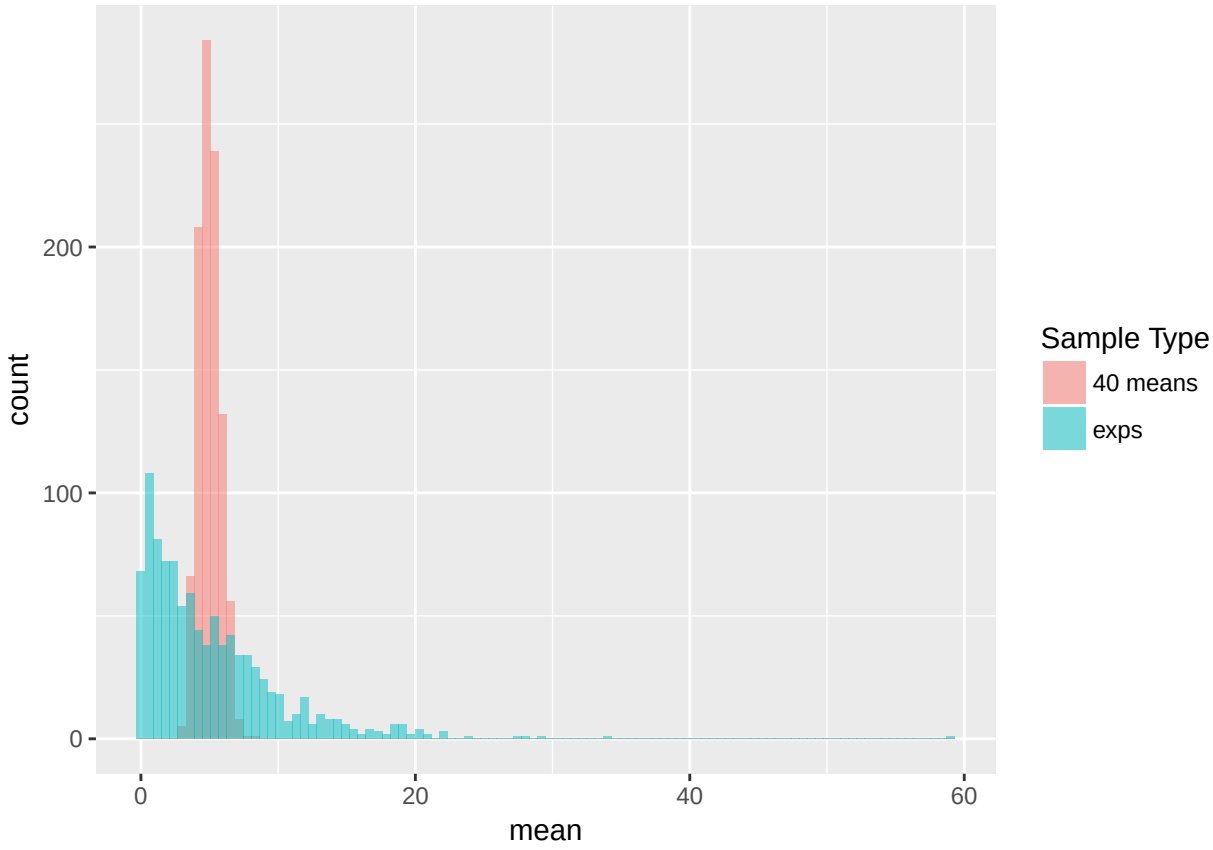
The resulting unified dataset looks like this:

```
head(samples)
```

```
## # A tibble: 6 x 2
##    mean type
##   <dbl> <fct>
## 1  4.90 40 means
## 2  5.23 40 means
## 3  6.40 40 means
## 4  4.74 40 means
## 5  5.18 40 means
## 6  5.17 40 means
```

If, as suggested, we now compare the distribution of a large collection of random exponentials (sample2) with the distribution of a large collection of averages of 40 exponentials (sample), the corresponding plot looks as follows:

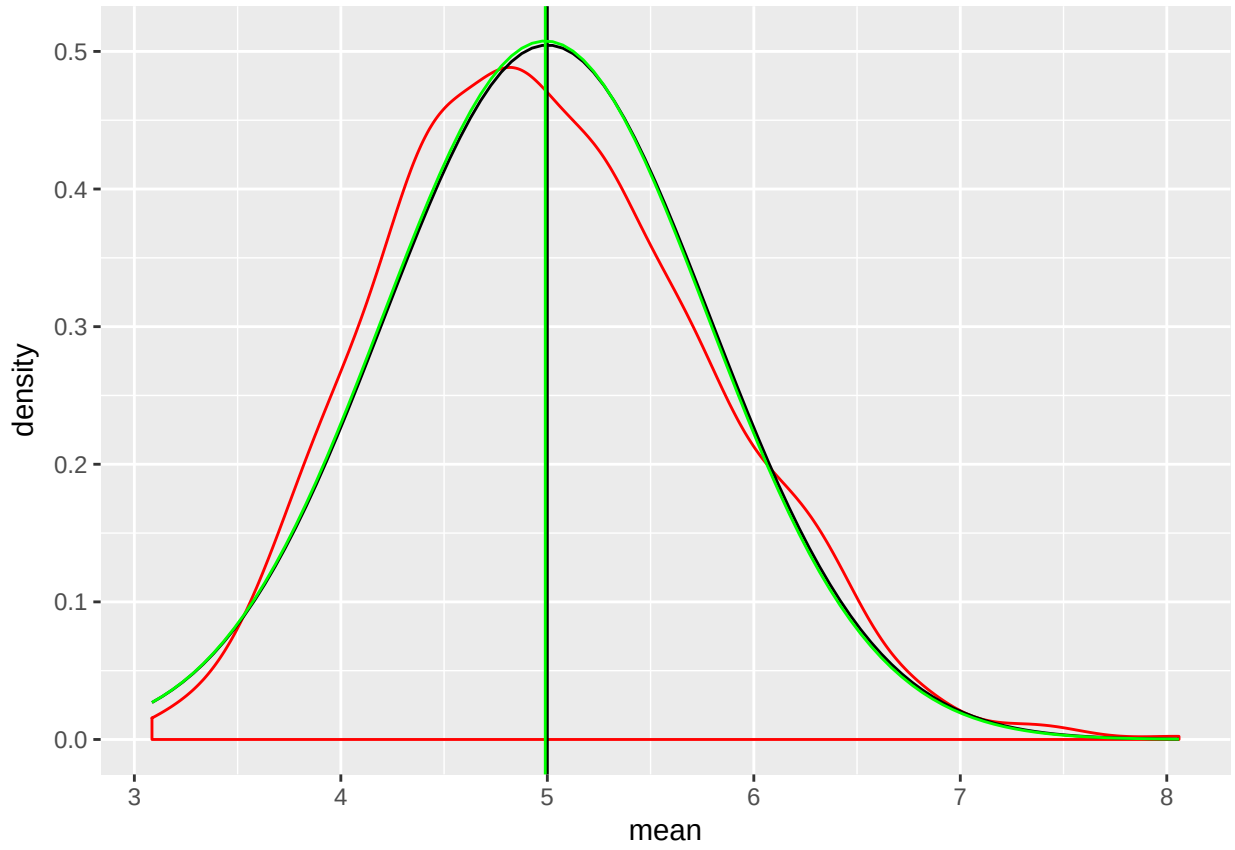
```
ggplot(samples, aes(x = mean)) +
        geom_histogram(bins = 100, aes(fill = samples$type), alpha = .5, position = "identity") +
        scale_fill_discrete(name = "Sample Type")
```



Note that, as expected, the distribution of the sample of averages of 40 exponential values (in red), looks quite normal in comparison with the distribution of exponential values, which looks, of course, quite exponential.

Finally, in order to conclude how close to the expected normal the first sample's distribution is, its density function (obtained with ggplot2's geom_density function) is compared to the functions obtained through the theoretical and sample's values (obtained previously) on the following plot:

```
ggplot(sample, aes(x = mean)) +
        geom_density(kernel = "gaussian", color = "red") +
        stat_function(fun = dnorm, args = list(mean = theor_values, sd = sample_sd_theor)) +
        geom_vline(xintercept = theor_values) +
        stat_function(fun = dnorm, args = list(mean = sample_mean, sd = sample_sd_R), color = "green") +
        geom_vline(xintercept = sample_mean, color = "green")
```



Where:

1) The red plot was obtained using a "gaussian" geom_density function, fed with the sample data.
2) The green normal was obtained with the sample mean and standard error.
3) The black normal was obtained with the distribution theoretical values.