Simulation of Exponential Variables and Means

Benjamin Phillips

Course project for the course Inferential Statistics, part of the Coursera John Hopkins data-science specialisation. Project and code information can be found on my github account

Load packages

```
library(ggplot2)
library(dplyr)
library(gridExtra)
```

Set seed, so the same variables can be reproduced if necessary

```
set.seed(103)
```

Create our exponential random variables with mean and standard deviation 5 (or lambda=0.2)

```
num_per_group <- 40
num_groups <- 1000
my_exponentials <- rexp(num_per_group*num_groups,0.2)
```

Get the means of 1000 groups of 40 exponential variables

```
row_means <- my_exponentials %>%
   matrix(nrow = num_groups) %>%
   rowMeans() %>%
   data_frame() %>%
   `names<-`("means")
```

Create a graph showing the distribution of the means

```
g1 <- ggplot(data = row_means, aes(means)) +
   geom_histogram(binwidth = 1/10) +
   ggtitle("Distribution of the Means") +
   geom_vline(xintercept = 5, color = "red")
```

Find the mean and standard deviation of the distrution of means

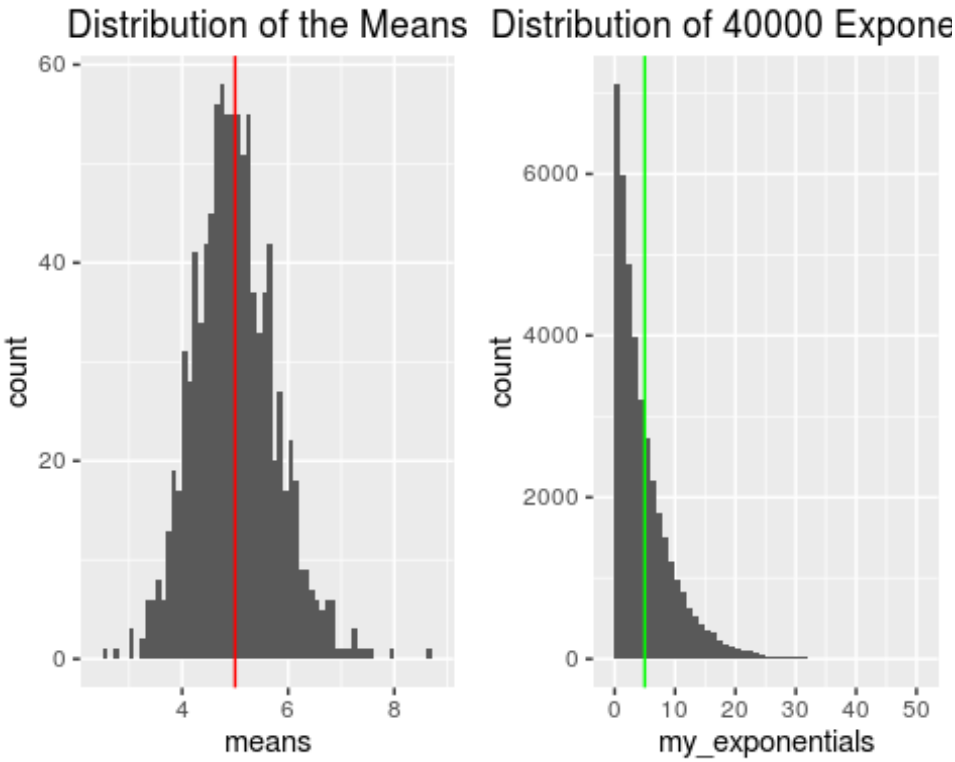
```
mean(row_means$means)

## [1] 4.992963

sd(row_means$means)

## [1] 0.7683478
```

Create a graph showing the distribution of the exponential variables
The exponential variables used previously are used here

```
my_exponentials <- data_frame(my_exponentials)
g2 <- ggplot(data = my_exponentials, aes(my_exponentials)) +
    geom_histogram(binwidth = 1) +
    ggtitle("Distribution of 40000 Exponentials") +
    geom_vline(xintercept = mean(my_exponentials$my_exponentials), color
= "green")
```

Print the graphs

```
grid.arrange(g1, g2, nrow = 1)
```



The graph on the left shows that the distribution of the means in normally distrubed, and from our calculations, has a mean of 4.99 and standard deviation of 0.77. This result is predicted be the Central Limit Theorem that states that for a (not necessarily normal) population of variables, the mean will be normally distributed. The red line on the left graph indicated the theoretical mean, which can be seen to be the centre of the distribution. The green line on the right is the mean of the 40000 variables.

Producing the distrubution of 1000 groups of 40 variables is useful because it allows us to find the variation in the mean.

```
t.test(row_means$means, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  row_means$means
## t = 205.49, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.945283 5.040642
## sample estimates:
## mean of x
##  4.992963
```

The t-test used on the row means shows that the 95 percent confidence interval of the mean is (4.945283, 5.040642). That is, we can be 95% certain that the mean lies between these values.