Statistical Inference Course Project Part 1 - Simulation Exercise

Dushan Yovetich

Overview

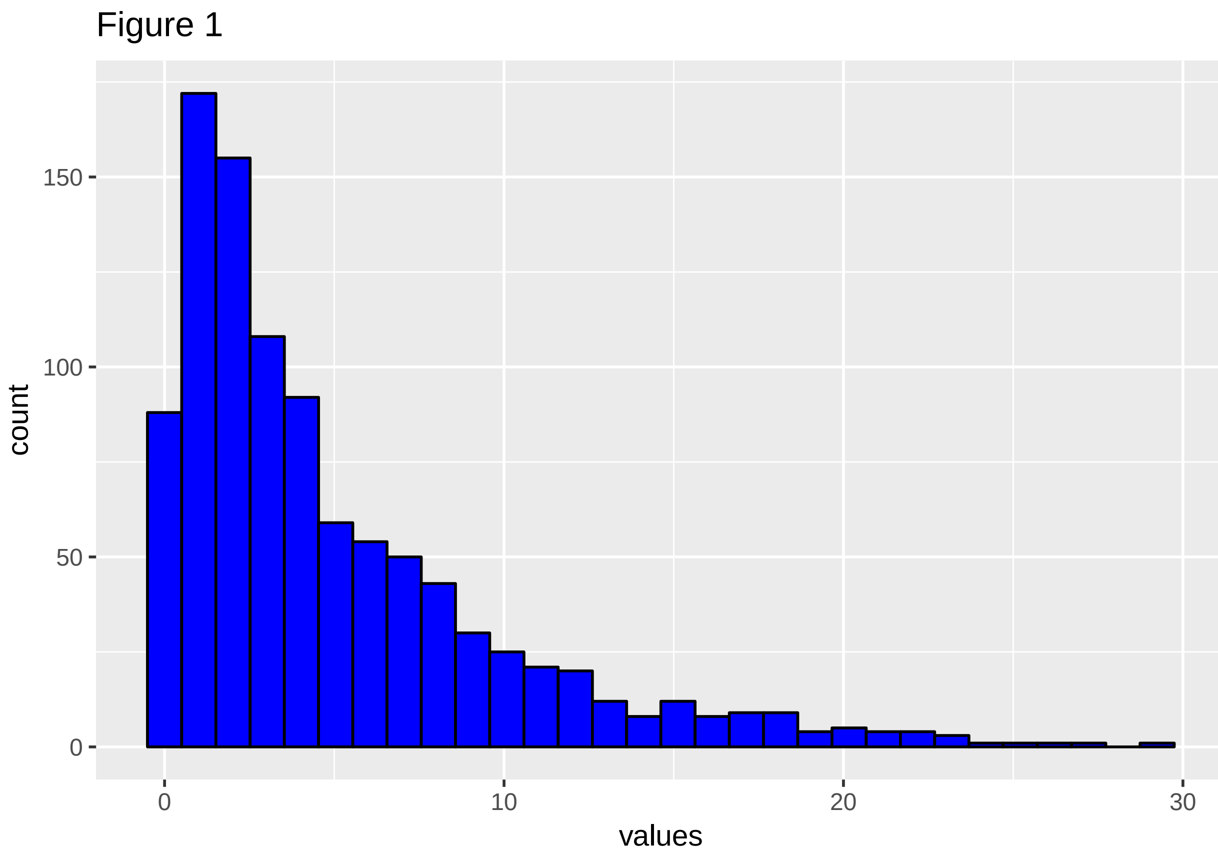
In this exercise, we investigate the exponential distribution and how the Central Limit Theorem applies. An exponential distribution is the probability distribution that describes the time between events in a process which occurs continuously and independently at a constant average rate (see https://en.wikipedia.org/wiki/Exponential_distribution). The Central Limit Theorem states if we take a large sample of averages of independent and indentically distributed variables from this distribution, the resulting distribution should be a normal one.

Simulations

An exponential distribution can be simulated using R and the function `rexp(n, lambda)`. In this function, the parameter **n** represents the number of observations we want, and **lambda** represents our constant average rate. Figure 1 is an example of a simulated exponential distribution. We generate a distribution with a 1000 variables with a constant average rate of 0.2 and plot it. As can be seen, an exponential distribution is skewed to the right.

```r
library(ggplot2)

set.seed(123)
qplot(rexp(1000,.2),geom = "histogram",fill = I("blue"),color = I("black"),
      xlab = "values",main = "Figure 1")
```

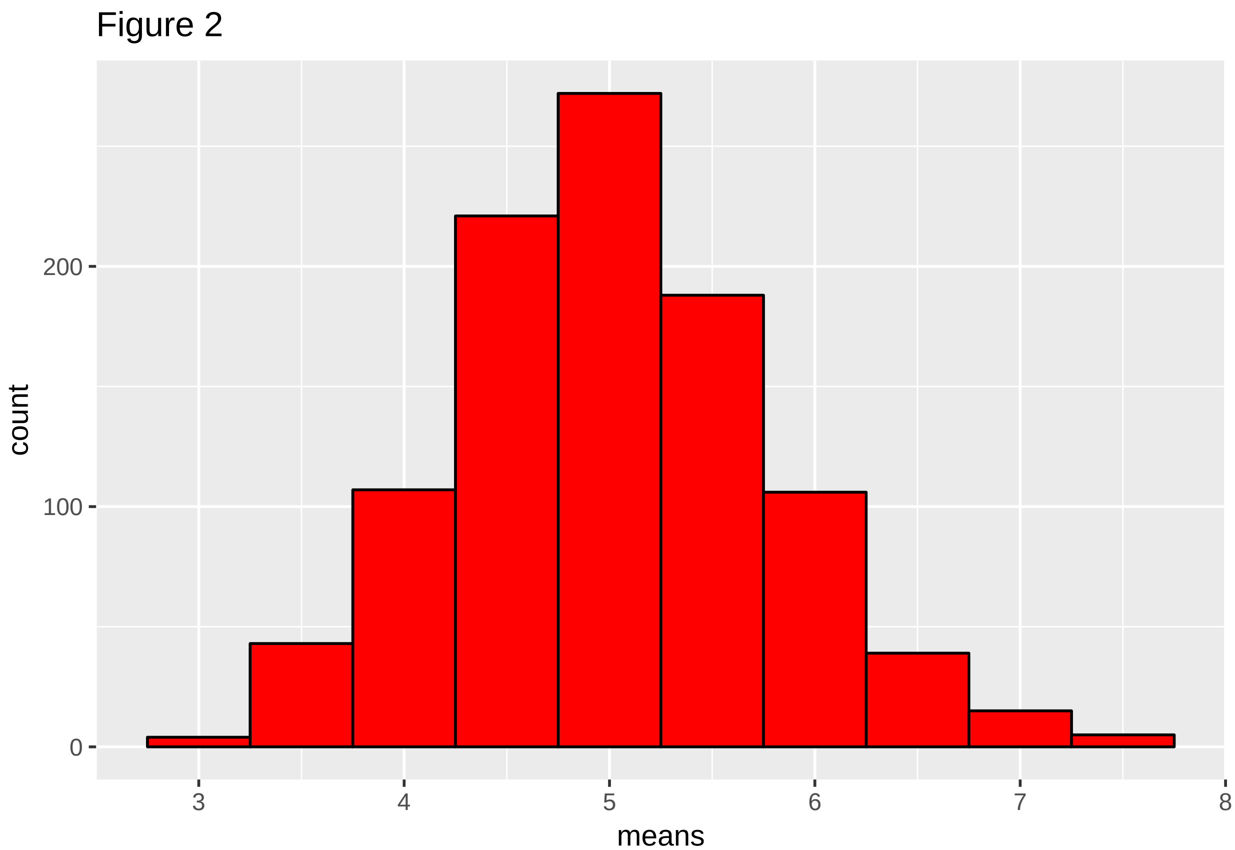


Let's see what happens when we take the average of a set of observations using a large sample. For the purpose of our analysis, we set n to 40 observations and lambda to 0.2. Additionally, because we want to capture a large set of averages, we take a sample of 1000 averages. The resulting distribution is illustrated in Figure 2.

```r
# defined parameters
n <- 40
lambda <- .2

# simulation
set.seed(123)
means <- NULL
for(i in 1:1000) means <- c(means,mean(rexp(n,lambda)))

qplot(means,geom = "histogram",binwidth = .5,fill = I("red"),
      color = I("black"),main = "Figure 2")
```



Sample Mean Versus Theoretical Mean

Next, we compare the mean of our simulated sample to the theoretical mean. Under Central Limit Theorem, the mean of a population should be the mean of the distribution. According to the exponential distribution, the theoretical mean should be equal to 1/lambda (again, lambda being our constant average rate).

```
# theoretical mean
1/lambda
```

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

For our sample mean, we simply take the mean of all the means in our sample

```
# our sample mean
mean(means)
```

```
## [1] 5.011911
```

The theoretical mean and the observed mean in our simulated sample are equitable.

Sample Variance Versus Theoretical Variance

Let's compare the variance of our simulated sample to the theoretical variance as well. Our theoretical deviation is noted to be 1/lambda. However, under Central Limit Theorem, the standard deviation should equal the standard error of mean for a normal distribution. Therefore, the variance should be the square of this standard error of the mean

```r
# theoretical variance
(1/lambda)^2/n
```

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

For our sample variance, we take the variance of all the means in our sample

```r
# our sample variance
var(means)
```

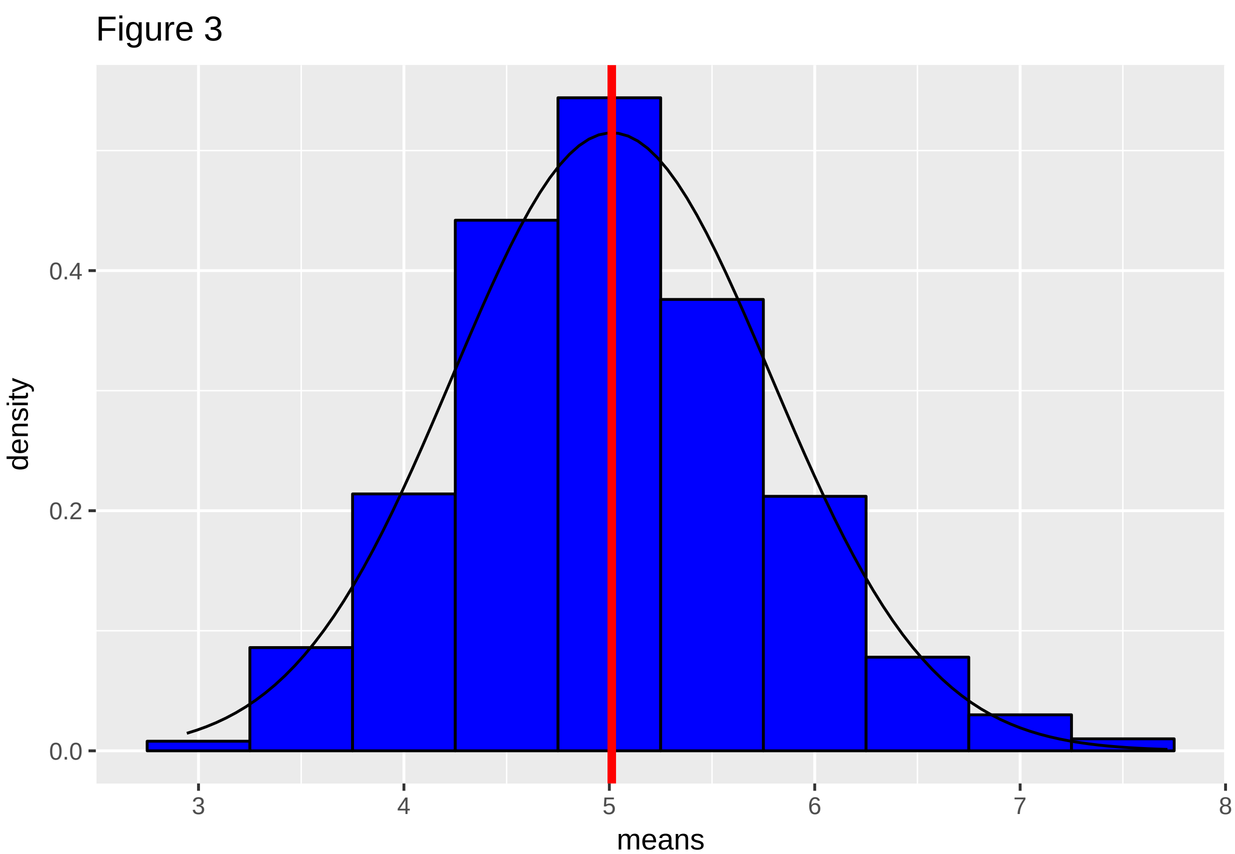
```
## [1] 0.6004928
```

The theoretical variance and the observed variance in our simulated sample appear to be similar.

Distribution

We take our population of 1000 averages simulated above (see Figure 2) and overlay the calculated sample mean and corresponding density curve. See Figure 3. As is illustrated, it is centered at the sample mean with the population equally distributed on either side. Further, the density curve is bell-shaped.

```r
qplot(means,geom = "blank", main = "Figure 3")+
  geom_histogram(aes(y =..density..),binwidth = .5, color = I("black"),fill = I("blue"))+
  stat_function(fun = dnorm,args = list(mean = mean(means),sd = sd(means)))+
  geom_vline(aes(xintercept = mean(means)),color = "red",size = 1.5)
```



In conclusion, the averages taken from a large sample of exponential distributions illustrates the Central Limit Theorem. As the sample size increases, the distribution becomes a normal distribution.