Statistical Inference Project : Part 1

Pradeepta Das

8th November 2020

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

Part 1: Simulation Exercise Instructions

Overview

The Central Limit Theorem (CLT) states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases. The important part is that it doesn't assume any distribution for the underlying iid variables. They could be drwan from any distribution. So here, for our simulation purposes, we will use exponential distribution.

Exponential Distribution

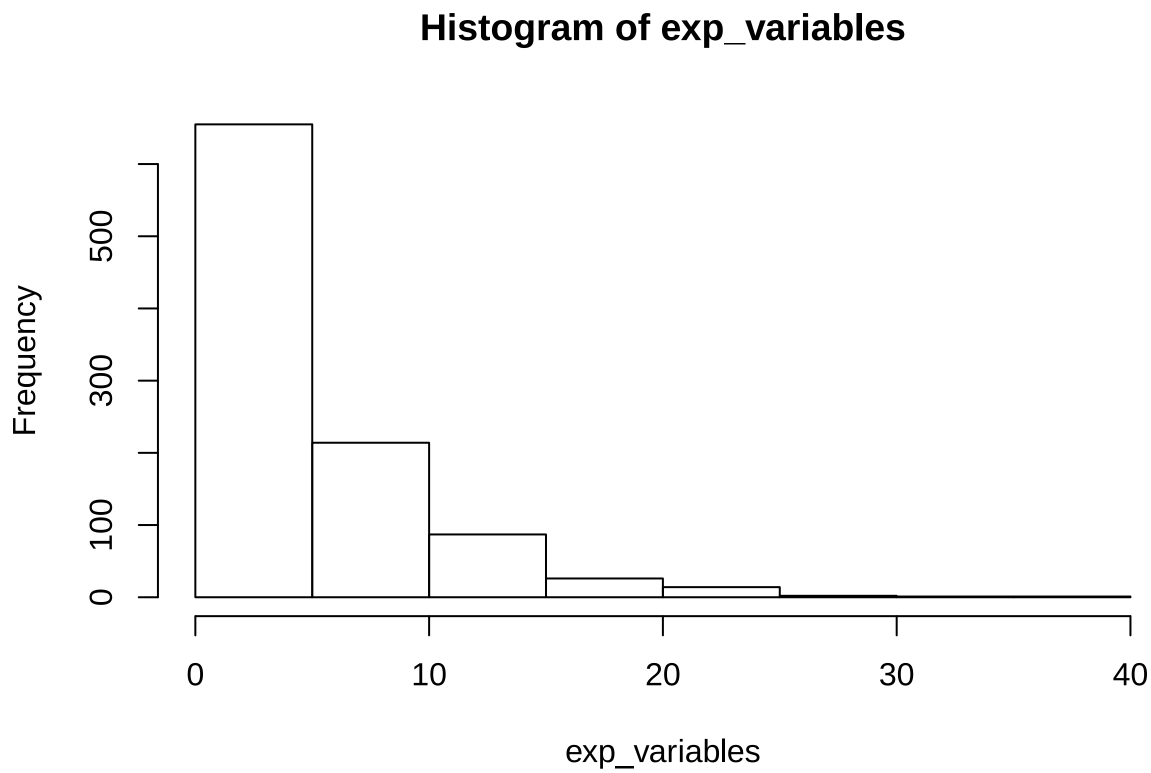
The exponential distribution is the probability distribution of the time between events in a Poisson point process. $PDF = \lambda e^{-\lambda x}$; where λ is the rate paramter.The mean of exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$.

Simulations

First, let us plot distribution of 1000 variables drawn from an exponential distribution with lambda = 0.2.

```
lambda <- 0.2
sample_size <- 40
no_sim <- 1000

exp_variables <- rexp(no_sim, lambda)
hist(exp_variables)
```



Population mean is:

```
mean(exp_variables)
```

```
## [1] 4.814912
```

We can see that it is close to the mean $1/0.5 = 5.0$

Sample Mean versus Theoretical Mean

Now, lets take 1000 samples of 40 sample size each and calculate mean for each sample.

```
sim <- data.frame(ncol=2,nrow=1000)
names(sim) <- c("Index","Mean")

for (i in 1 : no_sim){
  sim[i,1] <- i
  sim[i,2] <- mean(rexp(sample_size, lambda))

}
```

The sample mean is

```
sample_mean <- mean(sim$Mean)
sample_mean
```

```
## [1] 4.996567
```

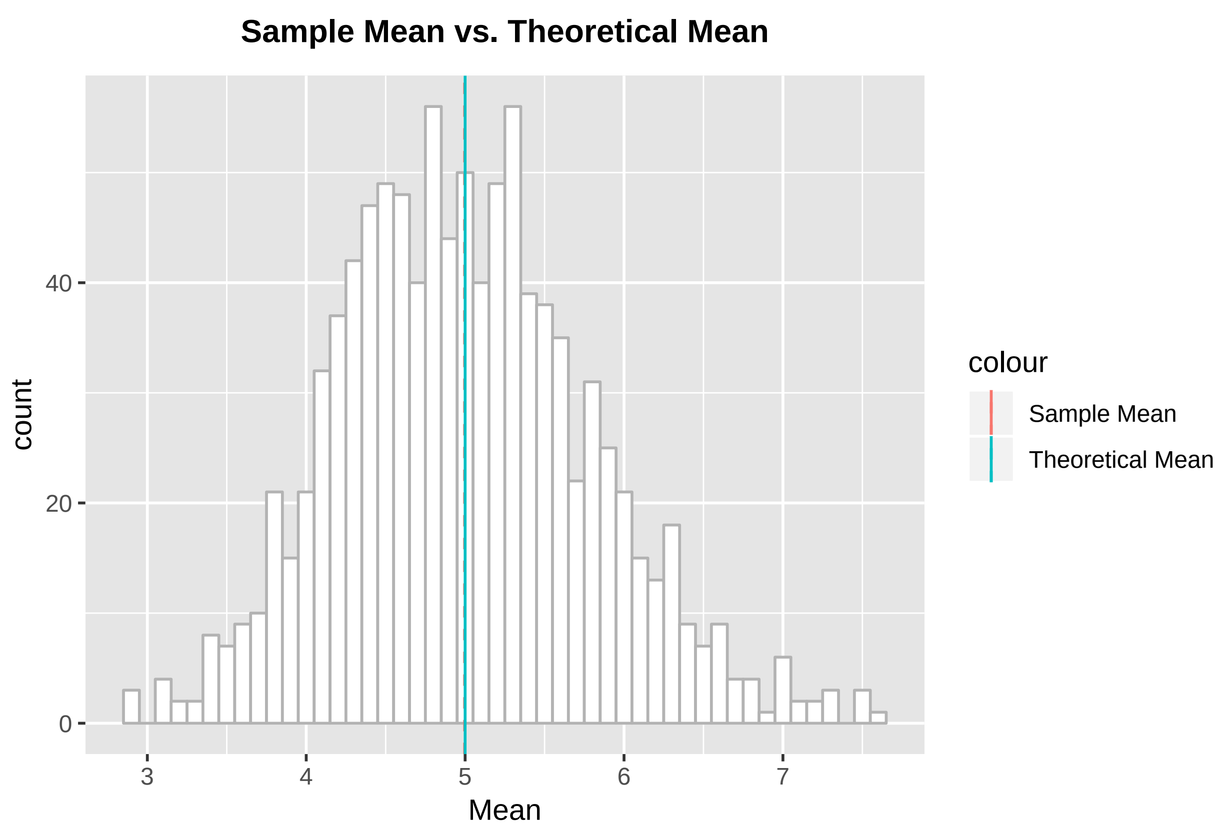
The theoretical mean is $1/\lambda =$

```r
theoretical_mean <- 1/lambda
theoretical_mean
```

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

The simulation mean of 4.9965674 is close to the theoretical value of 5.

```r
plt1 <- ggplot(data = sim, aes(x = Mean)) +
  geom_histogram(binwidth = .1, color = "grey70", fill = "white") + #plotting histogram
  geom_vline(aes(xintercept = sample_mean, color = "Sample Mean"), linetype = "dashed") +
  geom_vline(aes(xintercept = theoretical_mean, color = "Theoretical Mean")) +
  labs(title = "Sample Mean vs. Theoretical Mean") +
  theme(plot.title = element_text(size = 12, face = "bold",
      margin = margin(10, 0, 10, 0), hjust = 0.5),
      panel.background = element_rect(fill = "grey90")
  )
print(plt1)
```



Sample Variance versus Theoretical Variance

The sample variance is

```r
sample_var <- var(sim$Mean)
sample_var
```

```
## [1] 0.6335771
```

The theoretical variance is $1/\lambda =$

```
theoretical_var <- 1/(lambda*lambda)/sample_size
theoretical_var
```

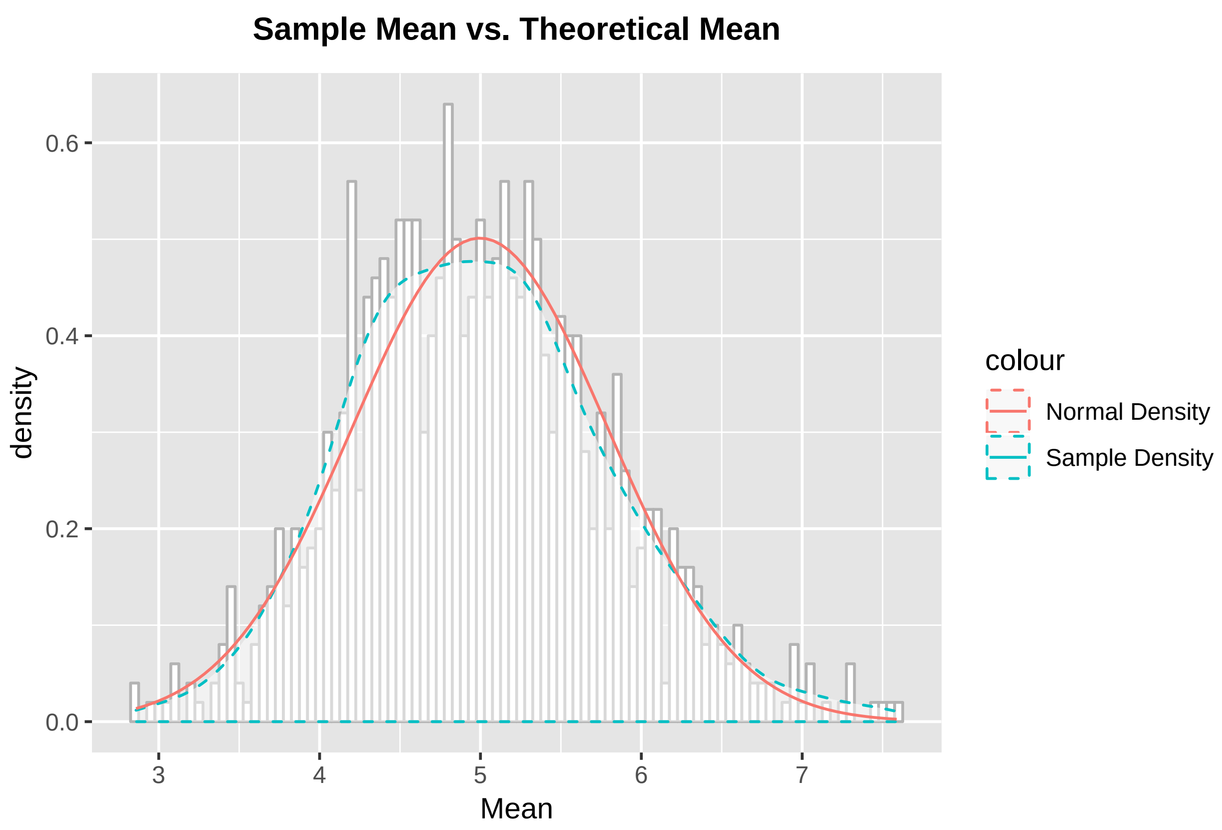
```
## [1] 0.625
```

The sample variance also matches with the theoretical variance.

Distribution

The following investigates whether the exponential distribution is approximately normal. Due to the Central Limit Theorem, the means of the sample simulations should follow a normal distribution.

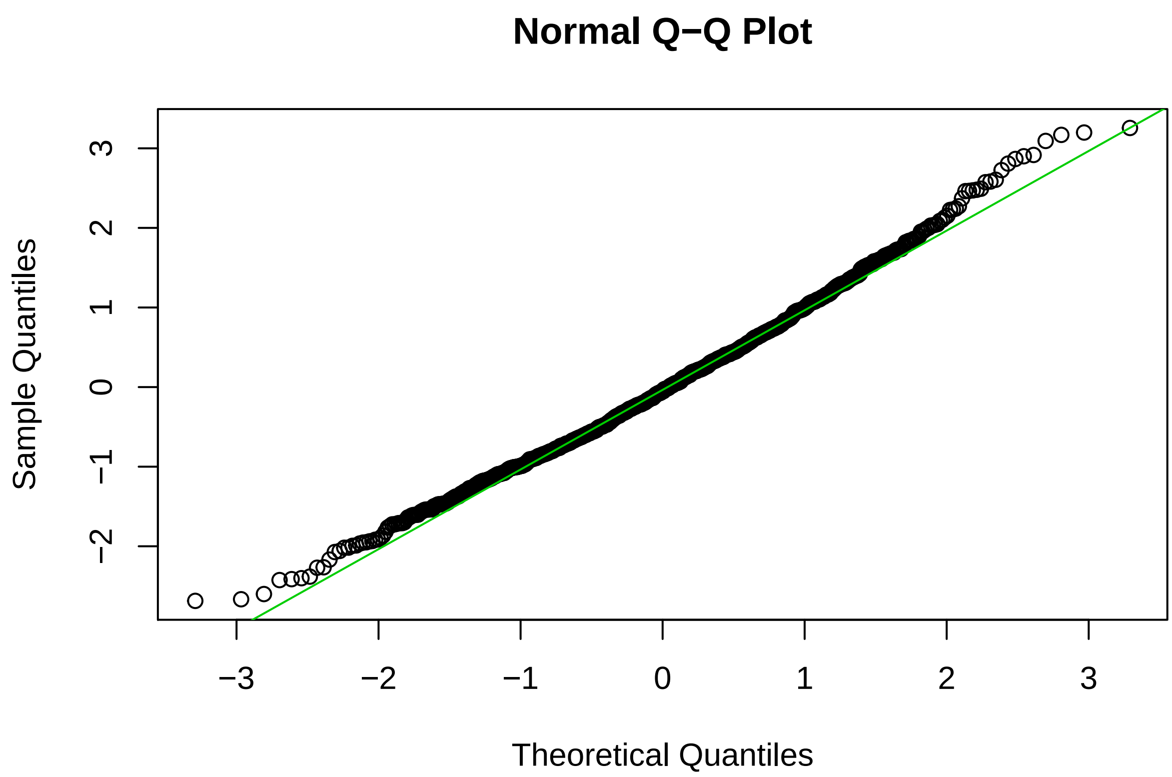
```
plt2 <- ggplot(data = sim, aes(x = Mean))+
  geom_histogram(binwidth = 0.05,aes(y=..density..), color = "grey70", fill = "white") +
  geom_density(alpha = .5, fill = "white", aes(color="Sample Density"), linetype="dashed") +
  labs(title = "Sample Mean vs. Theoretical Mean") +
  theme(plot.title= element_text(size = 12, face = "bold",
      margin = margin(10, 0, 10, 0), hjust = 0.5),
      panel.background = element_rect(fill = "grey90")) +
  stat_function(fun = dnorm, args = list(mean = sample_mean, sd = sqrt(sample_var)),
                aes(color="Normal Density"))
print(plt2)
```



From the QQ plot also, we can observe that the distribution of the mean (adjusted by its mean and variance) is close to the distribution of a standard normal variable.

```
means <- sim$Mean
means <- means - sample_mean
```

```
means <- means / sqrt(sample_var)
qqnorm(means, main ="Normal Q-Q Plot")
qqline(means, col = "3")
```



Conclusion

As shown above, the distribution of means of the simulated exponential distributions follows a normal distribution due to the Central Limit Theorem. If the number of samples increase (currently at 1000), the distribution should be even closer to the standard normal distribution (the solid line, above). The dotted line is the simulated curve.