Statistical Inferences and simulations

Shaun

February 19, 2017

Overview

For the purpose of this assignment, we'll be exploring the estimation of population parameters from samples by using the Central Limit Theorem. It will be shown that, given enough samples, the mean of will be approximately normal and act as a good estimate for the population.

Simulations

The following code generates a dataframe with a 1000 rows of 40 randomly generated numbers (40000 in total) from an exponential distribution. It then calculates the mean and standard deviation of each row.

```
# Load required libraries
library(ggplot2)
library(gridExtra)
set.seed(42)        # Set random number generator seed for reproducibility
n <- 40             # Number of exponentials to generate per simulation
lambda <- 0.2       # Rate parameter for rexp function
mu <- 1/lambda      # mean
sig <- 1/lambda     # standard deviation
vars <- sig^2       # calculate the variance
sim.count <- 1000   # Number of simulation to run
dis.matrix <- NULL  # Initialising data frame

# Run simulation to generate a sim.count by n (1000 x 40) data frame
for(i in 1:sim.count){
  dis.matrix <- rbind(dis.matrix,rexp(n, lambda))
}
dis.matrix.stacked <- as.data.frame(stack(as.data.frame(dis.matrix))$values)

# Calculate the row means
rms <- as.data.frame(rowMeans(dis.matrix))
colnames(rms) <- "mean"

# Calculate the row standard deviations
sds <- as.data.frame(apply(dis.matrix, 1, sd))
colnames(sds) <- "std"
```

Sample Mean versus Theoretical Mean

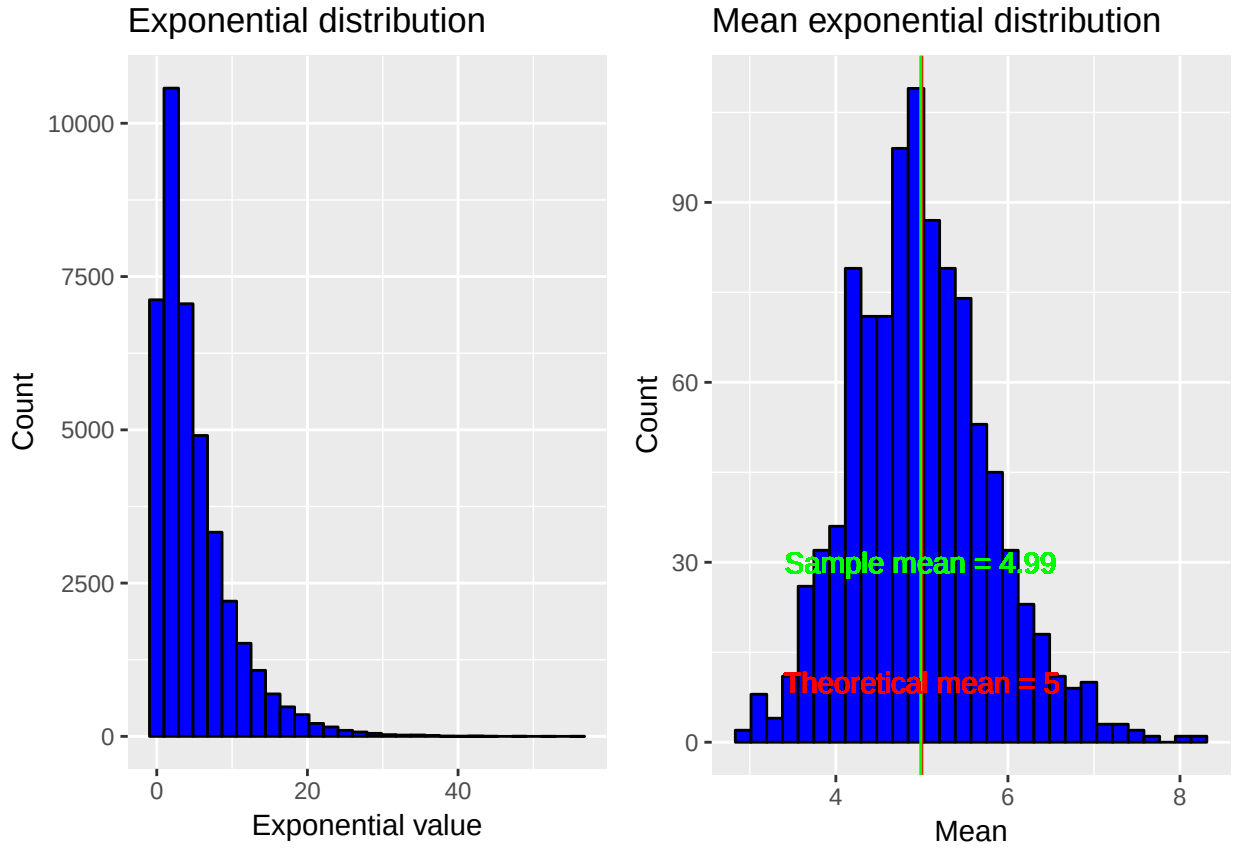
The first graph represents a histogram of the random samples generated from a exponential distribution. The second graph represents the mean of 40 exponential generated samples using the same parameters. The mean of the mean sample distribution has a 95% confidence interval of 4.9370807, 5.035936 where the theoretical mean is 5.

```
mean.plot <- ggplot(rms, aes(x= rms))+
  geom_histogram(fill="blue", colour="black", bins = 30)+
```

```
  scale_x_continuous()+
  labs(title="Mean exponential distribution", x="Mean", y="Count")+
  geom_vline(xintercept = mu, colour="red")+
  geom_vline(xintercept = mean(as.matrix(rms)), colour="green")+
  geom_text(aes(x=mu, label=paste0("Theoretical mean = ", mu), y=10), colour="red")+
  geom_text(aes(x=mean(as.matrix(rms)), label=paste0("Sample mean = ", round(mean(as.matrix(rms)),2)),

exp.plot <- ggplot(dis.matrix.stacked, aes(x= dis.matrix.stacked))+
  geom_histogram(fill="blue", colour="black", bins = 30)+
  scale_x_continuous()+
  labs(title="Exponential distribution", x="Exponential value", y="Count")

grid.arrange(exp.plot, mean.plot, ncol = 2)
```

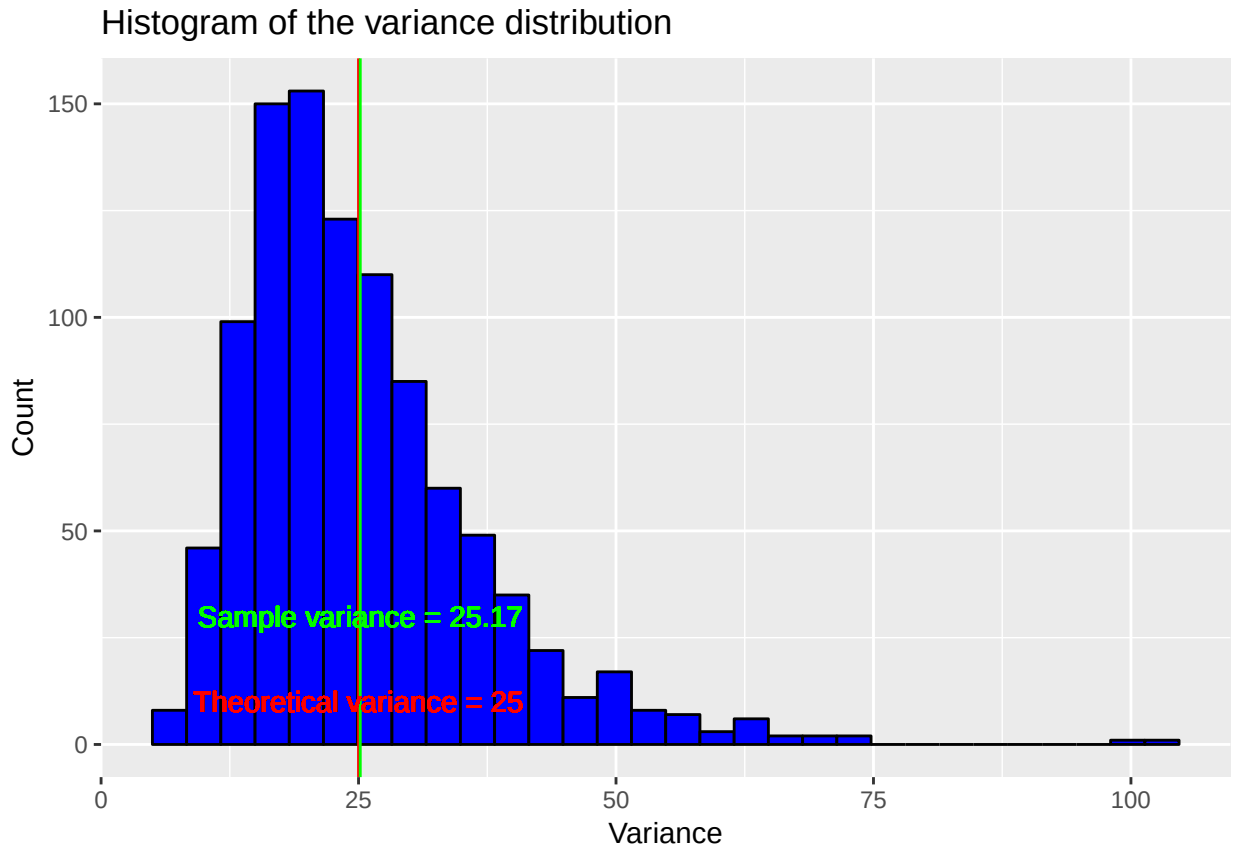


Sample Variance versus Theoretical Variance

The following graph represents the variance of the means of the exponential distribution. Unlike the distributions of the mean, it will not always approximate to another distribution. The difference between the theoretical distribution and the sample distribution is -0.165914.

```
ggplot(sds, aes(x = sds^2))+
  geom_histogram(fill="blue", colour="black", bins=30)+
  scale_x_continuous()+
  labs(title="Histogram of the variance distribution", x = "Variance", y = "Count")+
  geom_vline(xintercept = vars, colour = "red")+
```

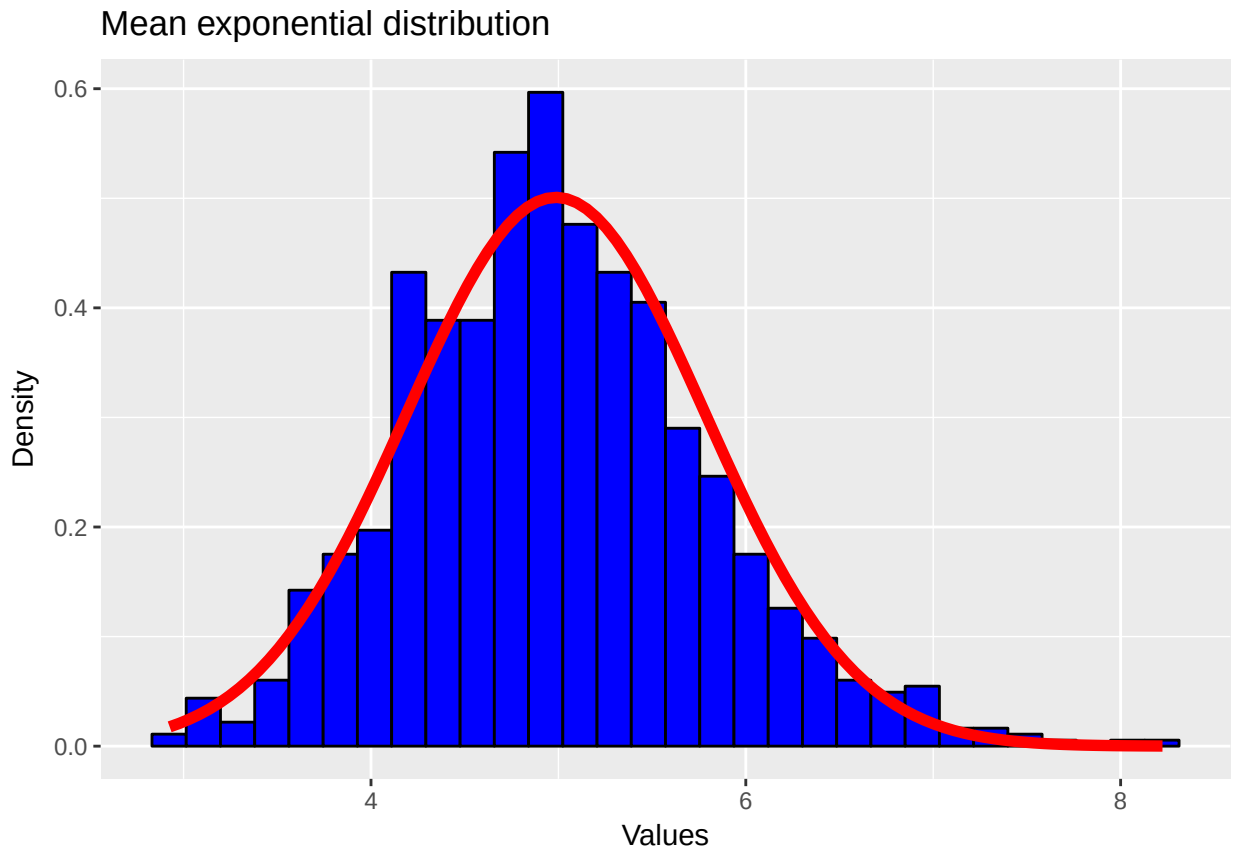
```
  geom_vline(xintercept = mean(sds^2), colour="green")+
  geom_text(aes(x = vars, label=paste0("Theoretical variance = ", vars), y=10), colour="red")+
  geom_text(aes(x = mean(sds^2), label=paste0("Sample variance = ", round(mean(sds^2),2)), y=30), colou
```



Distribution

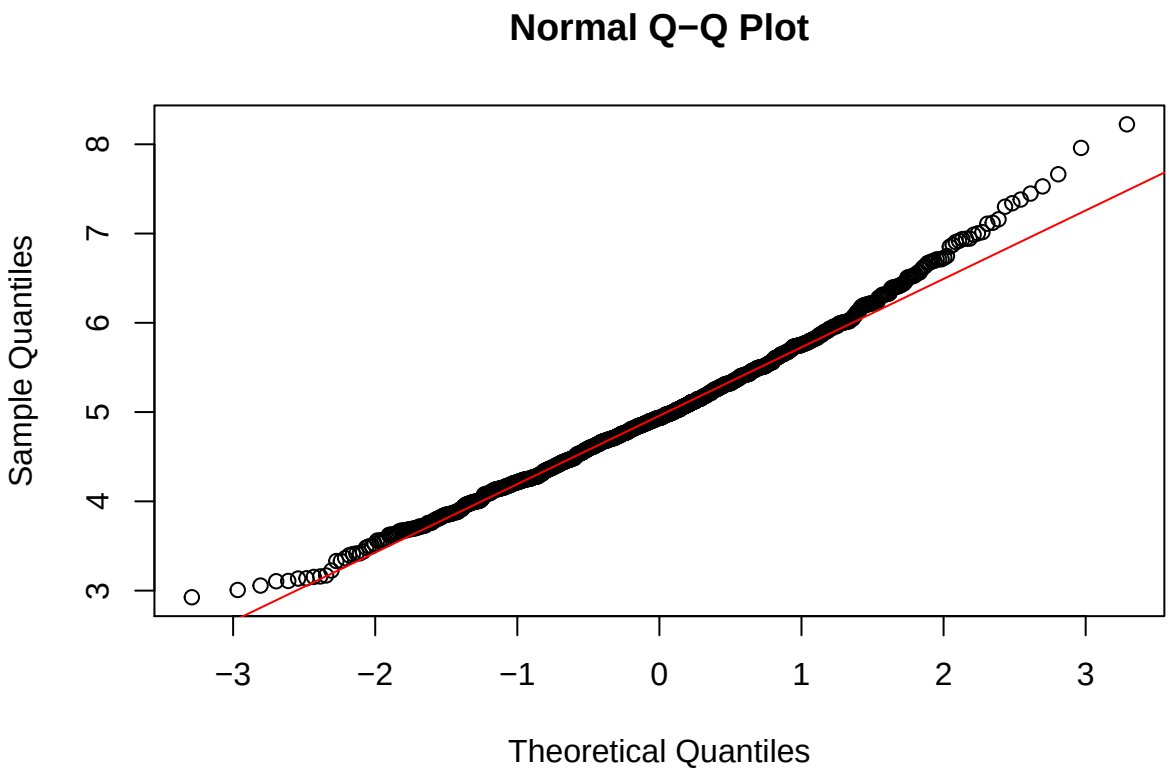
We can see that the distribution appears approximately normally distributed as a result of the Central Limit Theorem. The red graph below represents a normal distribution with the parameters of the simulation results N ~(mean = 4.9865083, sd = 0.7965177).

```
ggplot(rms, aes(x= rms))+
  geom_histogram(aes(y=..density..), fill="blue", colour="black", bins = 30)+
  stat_function(fun = dnorm, args = list(mean = mean(rms$mean), sd = sd(rms$mean)), lwd = 2, col = 'red
  scale_x_continuous()+
  labs(title="Mean exponential distribution", x="Values", y="Density")
```



The Q-Q plot below indicates that the results are approximately distributed. Only the extreme values tend to deviate from the expected normal distribution results.

```
qqnorm(rms$mean)
qqline(rms$mean, col="red")
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
3
4
5
6
7
8
−3
−2
−1
0
1
2
3