Simulation Exercise

Fritz Lin

1 Overview

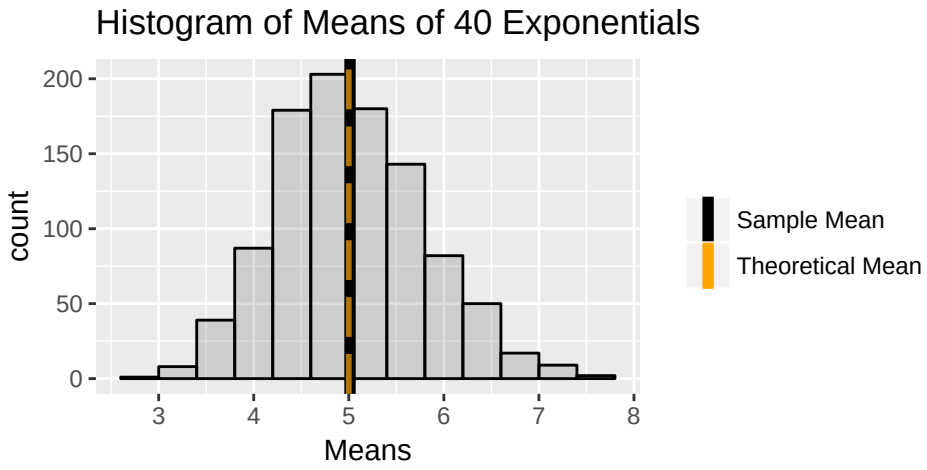
In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution has the following probability density function with parameter λ:

$$p(x) = \lambda e^{-\lambda x} \qquad \text{for } x \geq 0$$

2 Simulations

The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter. The mean of the exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. We set $\lambda = 0.2$ for all of the simulations. We will investigate the distribution of averages of 40 exponentials, that is $n = 40$. We will do a thousand simulations, that is $nosim = 1000$.

Below we have plotted the result of our simulations in the form of a histogram of means from 40 exponentials. Furthermore, we overlayed the simulated mean of 1000 means and the theoretical mean for comparison. As can be seen, they are more or less identical. We will investigage it more thoroughly in the next section. The code for the simulation and to generate the plot can be found in Appendix A.



3 Sample Mean versus Theoretical Mean

Let us now compare the sample mean to the theoretical mean of the exponential distribution.

```
# sample Mean
sampleMean <- mean(simMeans$x); sampleMean
```

```
## [1] 5.012806
```

```
# theoretical Mean
theoMean <- 1/lambda; theoMean
```

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

As shown the sample mean and the theoretical mean are very close. As we have set a sufficiently large sample size n, it was expected that the sample statistic should converge to the population statistic given the Law of Large Numbers (LLN).

4 Sample Variance versus Theoretical Variance

LLN states that the sample statistics of iid samples are all consistent estimators for their population counterparts. We have seen it for the sample mean being consistent. We now take a look at the sample variance.

```
# sample Variance
sampleVar <- var(simMeans$x); sampleVar
```

```
## [1] 0.6095236
```

```
# theoretical Variance
theoVar <- (1/lambda)^2 / n; theoVar
```

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

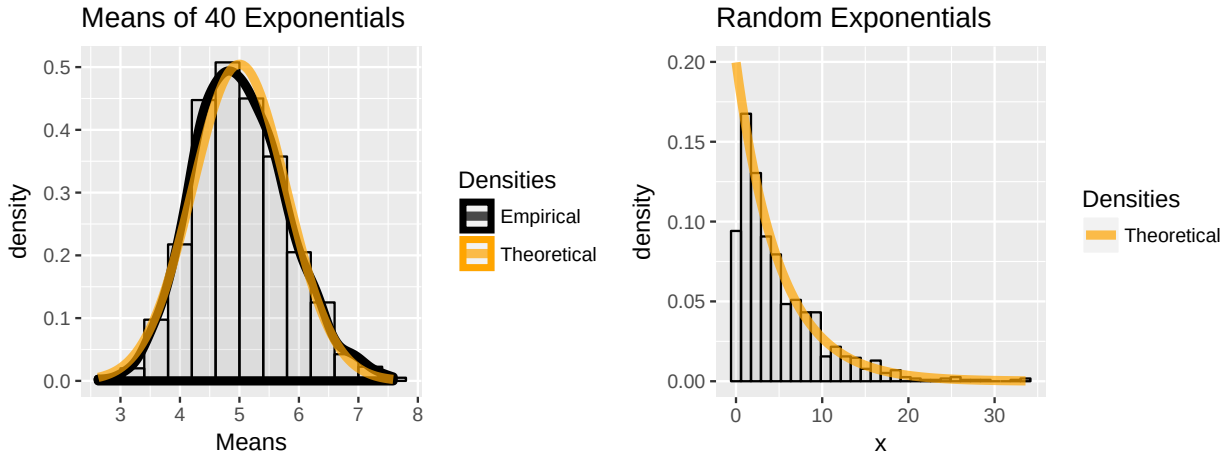
Again, values are close. So the sample variance is also consistent.

5 Distribution

Now we will show that the sample distribution is approximately normal. This is stated by the Central Limit Theorem (CLT), one of the most important theorems in statistics. In particular, CLT states that the distributions of means of iid variables approaches the normal, as sample size n increases. We will show it with regard to our sample distribution in several ways.

5.1 Visualization

First we do a visual examination of the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.



On the left side of the graph above we have plotted the histogram from before, but also overlayed the sample density distribution as well as the theoretical density. As we can see, the distribution of sample means convert

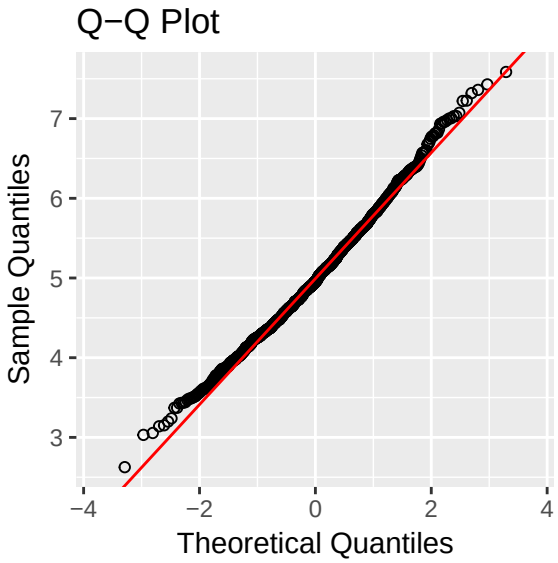
to the normal distribution. On the right side we have plotted the histogram of 1000 randomly generated numbers from the exponential distribution with $\lambda = 0.2$. The theoretical density function is also drawn.

If we compare both graphs, we see that regardless where the random numbers were drawn from, the distribution of iid samples with a sufficiently size of n will convert to the normal. In our case we have drawn 40 exponentials 1000 times, calculated their mean, resulting in the distribution similar to the normal.

Code for the above graph can be found in Appendix B

5.2 Q-Q Plot

The quantile-quantile (Q-Q) plot is a graphical method for determining if two data sets come from populations with a common distribution such as the normal. For example, if we perform a statistical analysis that assumes normal distribution for the random variable, we can use a Normal Q-Q plot to check that assumption. Let us do it for our simulated data.



The red line represents the quantiles for the normal (the theoretical) distribution. The black circles are the quantiles of our normalized simulated means. If the circles lie on the red line, we would say it is approximately normal. From around -1.5 to 1.5 we do see this behavior. The tails however divert from the normal distribution. The code for the Q-Q plot can be found in Appendix C

5.3 Confidence intervals for sample mean

Next we calculate the confidence intervals for our sample mean and check how many times our simulated means fall in between. It should be around 95%.

```
CI <- sampleMean + c(-1, 1) * qnorm(0.975) * 1/lambda/sqrt(n); CI
```

```
## [1] 3.463318 6.562293
```

```
coverage <- mean(simMeans$x > CI[1] & simMeans$x < CI[2]); coverage
```

```
## [1] 0.954
```

6 Appendices

6.1 Appendix A

```r
library(ggplot2)

# Settings
set.seed(83813)
n <- 40; nosim <- 1000
lambda <- 0.2

# Simulation
dat <- data.frame(
  x = c(apply(matrix(rexp(nosim * n, lambda), nosim), 1, mean), rexp(nosim, lambda)),
  what = factor(rep(c("Mean", "Obs"), c(nosim, nosim)))
)
simMeans <- subset(dat, what == "Mean")
simExp <- subset(dat, what == "Obs")

# plotting
ggplot(simMeans, aes(x)) +
  geom_histogram(binwidth = 0.4, alpha = .2, color = "black") +
  labs(title = "Histogram of Means of 40 Exponentials", x = "Means") +
  # sample mean
  geom_vline(aes(xintercept = mean(simMeans$x), color = "Sample Mean"), size = 2) +
  # theoretical mean
  geom_vline(aes(xintercept = 1/lambda, color = "Theoretical Mean"), size = 1,
             linetype = "longdash", alpha = 0.7) +
  scale_color_manual("", breaks = c("Sample Mean", "Theoretical Mean"),
                     values = c("black", "orange"))
```

6.2 Appendix B

```
library(gridExtra)

# plotting
g1 <- ggplot(simMeans, aes(x)) +
  geom_histogram(binwidth = 0.4, alpha = .1, color = "black", aes(y = ..density..)) +
  geom_density(size = 2, aes(color = "Empirical")) +
  stat_function(fun = dnorm, args = list(mean=1/lambda, sd=1/lambda/sqrt(n)),
    aes(color = "Theoretical"), size = 2, alpha = 0.7) +
  scale_color_manual("Densities", breaks = c("Empirical", "Theoretical"),
    values = c("black", "orange")) +
  labs(title = "Means of 40 Exponentials", x = "Means")

g2 <- ggplot(simExp, aes(x)) +
  geom_histogram(alpha = .1, color = "black", aes(y = ..density..)) +
  stat_function(fun = dexp, args = (mean=lambda),
    aes(color = "Theoretical"), size = 2, alpha = 0.7) +
  scale_color_manual("Densities", breaks = c("Empirical", "Theoretical"),
    values = c("orange")) +
  labs(title = "Random Exponentials")

grid.arrange(g1, g2, ncol=2)
```

6.3 Appendix C

```
# Q-Q plot
y <- quantile(simMeans$x, c(0.25, 0.75))
x <- qnorm(c(0.25, 0.75))
slope <- diff(y) / diff(x)
int   <- y[1] - slope * x[1]

ggplot(simMeans, aes(sample = x)) +
  stat_qq(shape = 1) +
  geom_abline(intercept=int, slope=slope, color="red") +
  labs(title = "Q-Q Plot", x = "Theoretical Quantiles", y = "Sample Quantiles") +
  coord_cartesian(xlim = c(-3.75, 3.75))
```