The Exponential Distribution and the Central Limit Theorem

Overview

In this report, we investigate the exponential distribution in relation with the Central Limit Theorem (CLT). In particular, we compare the distribution of a sample of exponentials with the distribution of averages of exponential random variables. We investigate the properties of this distrubution by comparing the simulated sample mean and variance with their corresponding theoretical values, and show the averages to be approximately normally distributed, in agreement with the CLT.

Simulations

The data generated is a sample from the distribution of averages of 40 exponentials. For the sake of the experiment, we set the rate parameter, λ, to 0.2, and firstly generate a random sample of 1000 exponentials. We then draw a sample of 1000 averages of 40 iid exponentials. This simulated data is arranged as a matrix of 1000 rows and 40 colums, where each row contains the values from one of the samples of 40 random exponential values. Next, we take the average from each random sample, thus obtaining a sample from the distribution of averages of 40 random exponentials.

```
set.seed(2791) # allow reproducibility
nosim <- 1000 # number of simulations
n <- 40 # sample size
lambda <- 0.2 # rate parameter, fixed
x <- rexp(nosim, rate = lambda) # a sample of 1000 random exponentials
head(x)
```

```
## [1] 12.2485907  0.6410901  1.2725927  4.9849586  0.1333218 13.5017177
```

```
data <- matrix(rexp(nosim * n, rate = lambda), nosim)
avgs <- apply(data, 1, mean) # a sample of 1000 averages of 40 random exponentials
head(avgs)
```

```
## [1] 5.327338 5.018687 5.006336 4.356090 5.745851 4.569238
```

We now explore the properties of the distribution of averages of exponentials by looking at the obtained simulated sample of averages.

Sample Mean versus Theoretical Mean

Since the sample mean is an unbiased estimator, the expected value of the distribution of averages of iid variables is the same as the population mean which is equal to $\frac{1}{\lambda}$ for exponential variables with rate λ. We compare the population mean with the simulated sample mean and see that they exhibit very little difference.

```
sample.mean <- mean(avgs)
mu <- 1 / lambda
```

The simulated sample mean is 4.957 and population mean is 5. They are indeed pretty close to each other:

```
round(abs(sample.mean - mu),3)
```

```
## [1] 0.043
```

We show this graphically by plotting the density of averages and highlighting the two compared means:

```
g <- ggplot(data.frame(xhat = avgs), aes(x = xhat)) +
  geom_density() +
  geom_vline(xintercept = sample.mean) +
  geom_vline(xintercept = mu, col = "red") +
  theme(axis.title.x=element_blank(),axis.title.y=element_blank())
plot(g)
```

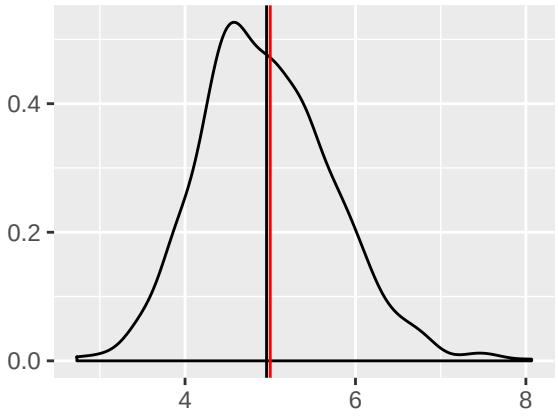

Figure 1: Distribution of averages of 40 exponential random variables with rate $\lambda = 0.2$. The vertical black line indicates the sample mean, and the red one is the theoretical value of the mean $\mu = 1/\lambda$.

Sample Variance versus Theoretical Variance

Under the central limit theorem, the variance of the sample mean is the variance of the population ($\frac{1}{\lambda^2}$) divided by the sample size of each random draw. Similarly to the mean, we compare this theoretical value with the empirical one (the sample variance of the simulated averages):

```
sample.variance <- var(avgs)
variance <- 1/(n*lambda^2)
vars <- data.frame(sample_variance = sample.variance, th_variance = variance)
colnames(vars) <- c("Sample Variance", "Theoretical Variance")
```

Here are the two values:

Sample Variance	Theoretical Variance
0.596	0.625

The difference between the two values is 0.029. The simulated variance is smaller than the expected one. For the purpose of the experiment, these two values seems to be close enough. This indicates the choice of the sample size to be a reasonably large one to empirically confirm the variance predicted by the central limit theorem.

Distribution of Averages: Comparison with the Exponential and Normal Distributions

First, let's compare the summaries of the original sample of 1000 random exponentials and the sample of 1000 averages of 40 exponentials:

```
summary(x)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00137  1.57600  3.56800  5.06500  6.86600 34.60000
```

```
summary(avgs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.730   4.421   4.902   4.957   5.444   8.068
```

which shows clear differences except the mean value, as predicted by the CLT.

Fig.3 in the Appendix shows a plot of the sample of 1000 exponentials, together with the density of the generating distribution. Except the lowest values, the sample looks not unexpectedly drawn from the exponential distribution, which is different from the distribution of a large collection of averages of 40 exponentials. We show this distribution in Fig. 2, compared with the CLT distribution which is $N(1/\lambda, 1/\sqrt{40}\lambda)$.

```
g <- ggplot(data.frame(xhat = avgs), aes(x = xhat)) +
  geom_histogram(alpha = 0.5, binwidth = 0.2, aes(y = ..density..), color = "black") +
  stat_function(fun = dnorm, args = list(mean = mu, sd = sqrt(variance)),
                size = 1.5, col = "red") +
  theme(axis.title.x=element_blank(),axis.title.y=element_blank())
plot(g)
```

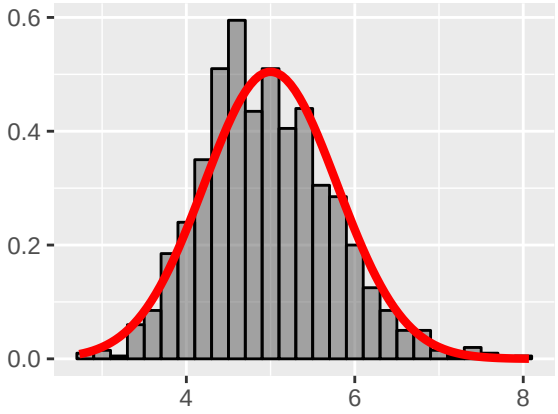

Figure 2: Distribution of averages of 40 exponential random variables with rate $\lambda = 0.2$. The theoretical mean and standard deviation of this distribution are respectively $\mu = 1/\lambda$ and $\sigma = 1/\sqrt{40}\lambda$. To illustrate the CLT, the distribution is compared to a normal distribution $N(\mu, \sigma)$.

The simulated averages look approximately normally distributed. To further confirm this, we plot the quantiles from the normalised sample averages against the standard normal quantiles in Fig.4 in the Appendix. The normal QQ plot confirms the sample of averages is approximately normally distributed. As the sample size n increases, we expect it converge to $N(1/\lambda, 1/\sqrt{n}\lambda)$, as predicted by the Central Limit Theorem.

Appendix

Simulated data vs Exponential Distribution

```
g <- ggplot(data.frame(x = x), aes(x = x)) +
  geom_histogram(binwidth = .4, aes(y = ..density..), color = "black") +
```

```
  stat_function(fun = dexp, args = list(rate = lambda), size = 1.5, col = "red") +
  theme(axis.title.x=element_blank(),axis.title.y=element_blank())
plot(g)
```

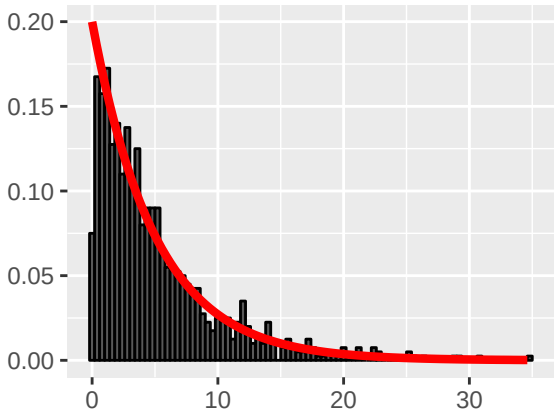

Figure 3: Histogram of the simulated data overlaid with the theoretical exponential distribution (red curve, $\lambda = 0.2$).

Distribution of Averages: QQ plot

The sample of averages is first rescaled so that the mean and standard deviation are respectively 0 and 1. Their quantiles are then compared to the standard normal quantles:

```
z <- (avgs - mu)/sqrt(sample.variance)
qqnorm(z)
qqline(z)
```

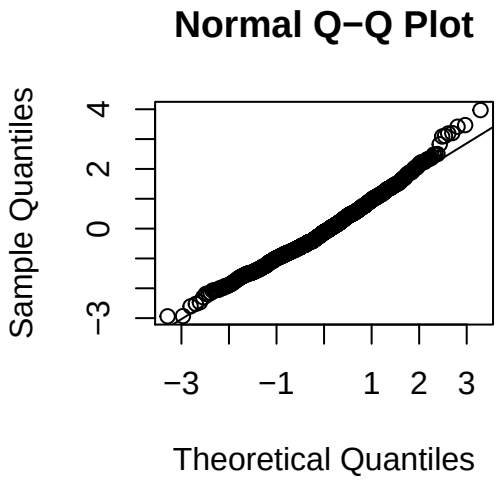

Figure 4: QQ plot of the distribution of the simulated averages of 40 exponential random variables. The straight line is the standard normal reference.