Statistical Interference: Course Project

Ulrich Tiedau

Monday, February 16, 2015

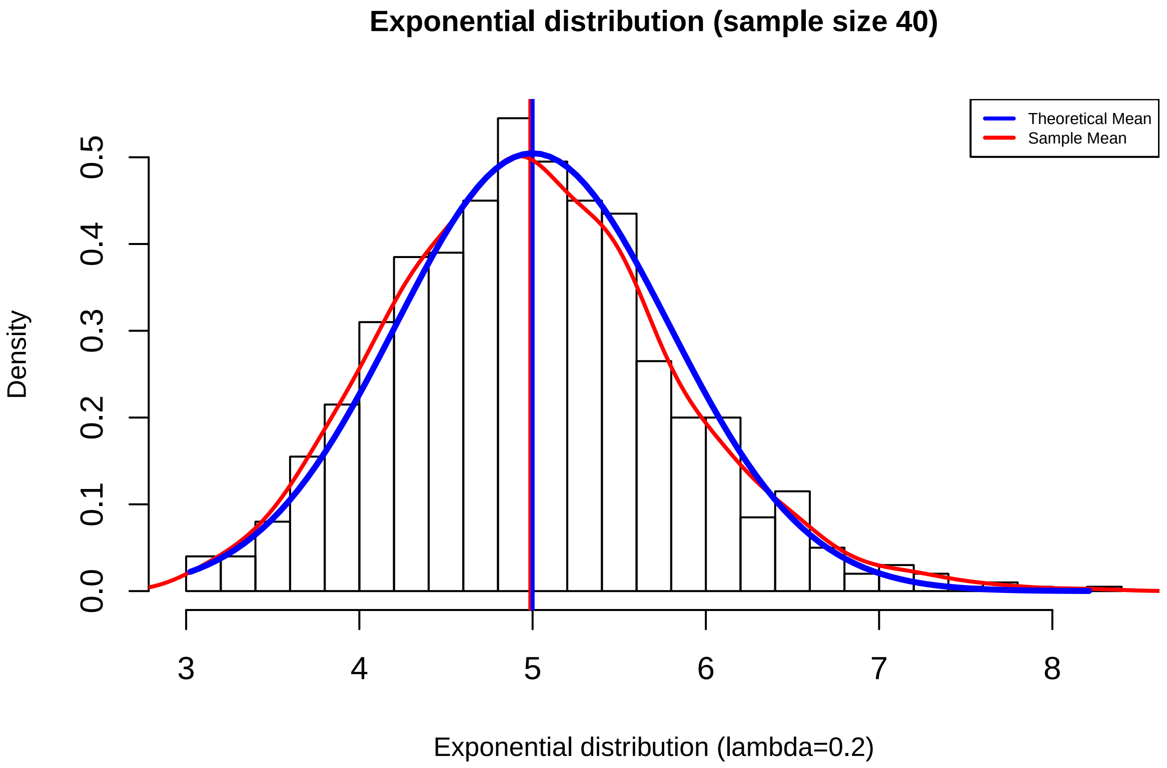
Overview

This project illustrates the properties of the exponential distribution of the mean of 40 exponentials and compares it with the Central Limit Theorem. In particular it 1) shows the sample mean and compares it to the theoretical mean of the distribution; 2) shows how variable the sample is (via variance) and compares it to the theoretical variance of the distribution; and 3) shows that the distribution is approximately normal.

Part 1: Simulations

First we define a function for running the simulations with the arguments defaulting to the values set in the assignment (λ = 0.2, sample size = 40 and number of simulations = 1000). The function simulates the exponential distribution with the R expression `rexp(nsim * nexp, rate=lambda)` and returns a vector with the sample data. Then we run the simulations and plot a histogram which shows that the distribution is fairly close to normal, although slightly skewed, with the sample mean being very close to the theoretical mean as we will investigate next:

```
run.simulations <- function (nexp = 40, lambda = 0.2, nsim = 1000) {
  v <- rowMeans(matrix(rexp(nsim*nexp, rate=lambda), nsim, nexp))
  return(v)
  }
set.seed(1000)
dat <- run.simulations()
```



Part 2: Sample Mean versus Theoretical Mean

Both the mean and the standard deviation of the exponential distribution are `1/lambda`. As in this experiment the rate parameter for all simulations is set as `lambda = 0.2`, this means that the theoretical mean is 5.0 and the theoretical standard deviation 5.0 as well. Comparing the sample mean (4.9869634) with the theoretical mean (5.0), we see that it differs by only 0.26%.

```
lambda <- 0.2
mean.theoretical <- 1/lambda
mean.sample <- mean(dat)
```

Part 3: Sample Variance versus Theoretical Variance

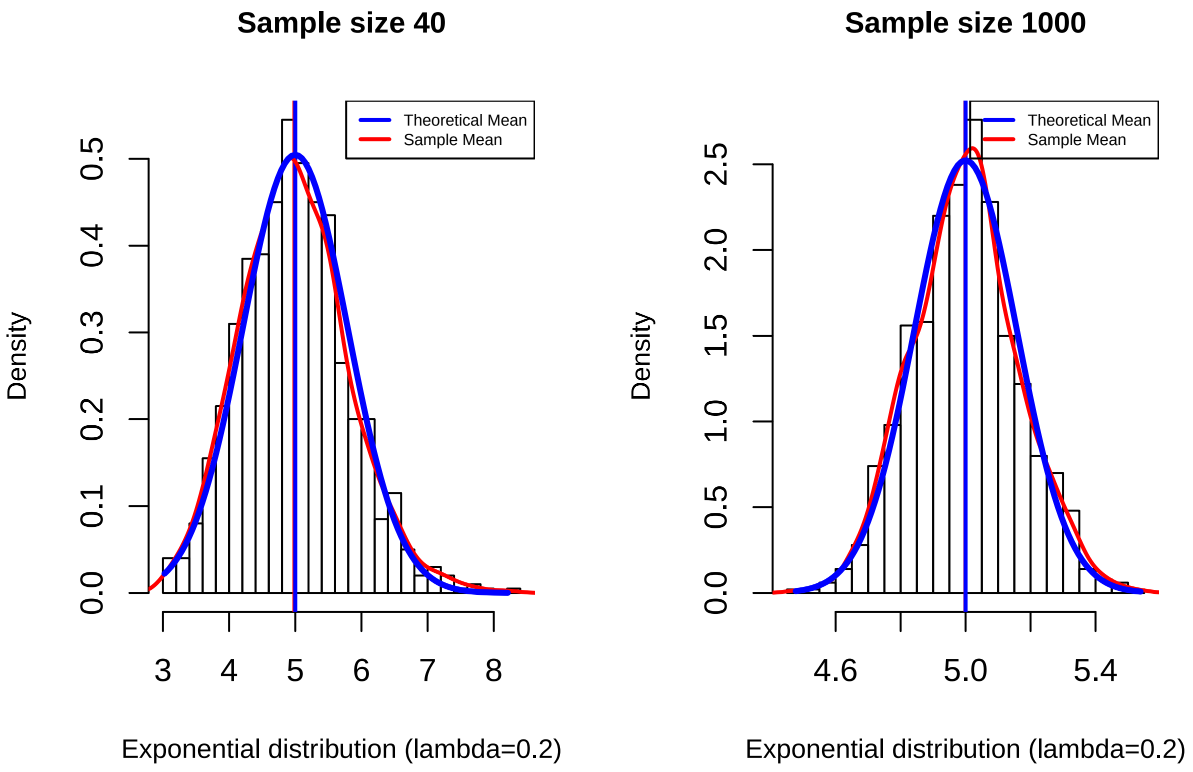
Next we investigate the variance or the distribution. It turns out that the sample variance (0.654343) differs significantly more from the theoretical variance (0.625), namely 4.69%.

```
variance.theoretical <- ((1/lambda)^2)/nexp
variance.sample <- var(dat)
```

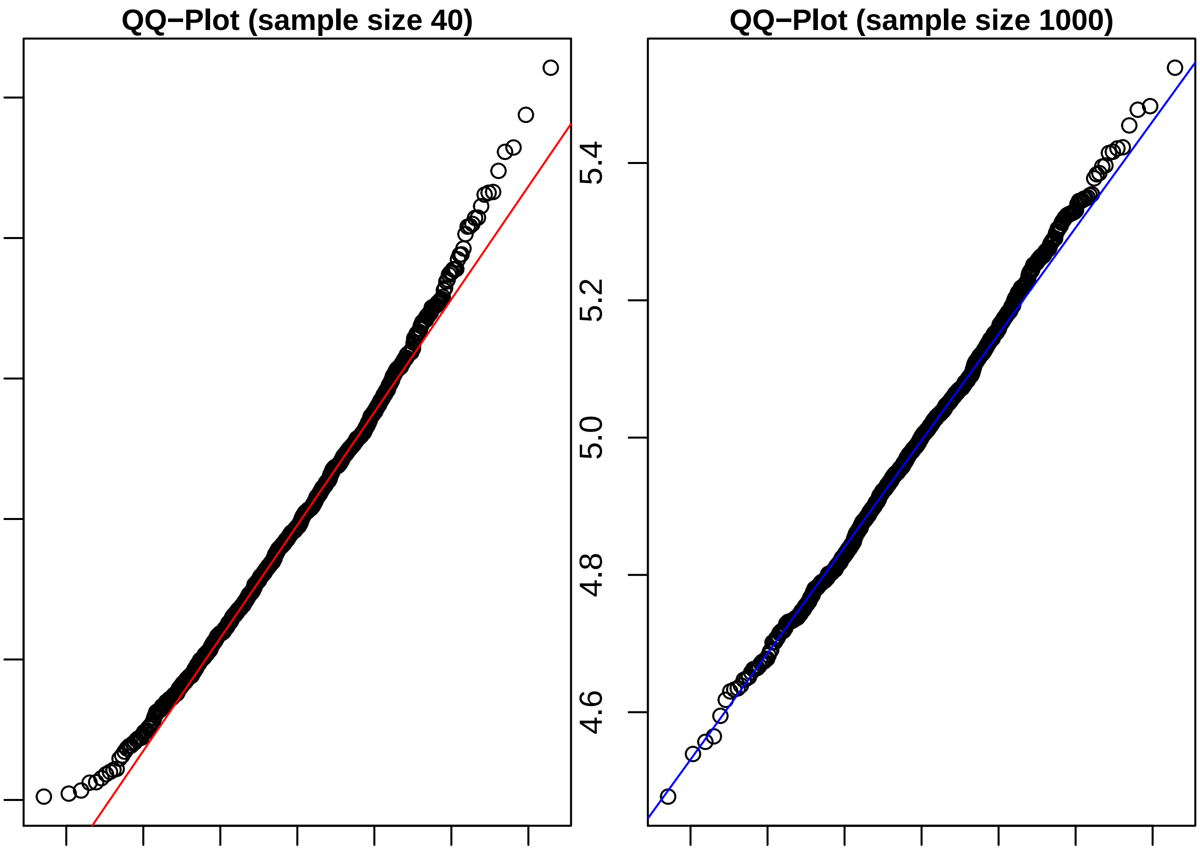
Part 4: Distribution

To see whether the distribution confirms with the Central Limit Theorem (for our purposes here stating that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases and approximates normality), we run a second simulation with an increases sample size of 1000 and compare the two histograms:

```
par(mfrow=c(1, 2)); figure(dat, lambda=0.2, nexp=40, breaks=20, main="Sample size 40", xlab="Exponential dist
dat2 <- run.simulations(nexp=1000); figure(dat2, lambda=0.2, nexp=1000, breaks=20, main="Sample size 1000", x
```



As we can see, the second simulation with the increased sample size of 1000 approximates normality even closer than the first simulation with a sample size of 40. The distribution thus confirms with the Central Limit Theorem (CLT). To further verify this finding we plot the theoretical against the sample quantiles in a QQ-Plot which allows to inspect how closely the data fits the chosen theoretical distribution:



As the quantile follow the line closely, we can conclude that the exponential distribution of the mean of 40 exponentials approximates a normal distribution. The fact that it does so even more closely for the mean of 1000 exponentials is another confirmation that the exponential distribution confirms with the Central Limit Theorem.

5. Appendix

5.1 Code used to generate the plots

```
figure <- function(x, lambda=0.2, nexp=40, breaks=20, main=main, xlab=xlab) {
  mean.theoretical <- 1/lambda # exponential distribution
  sd.theoretical <- 1/lambda # exponential distribution
  par(cex.main=0.9, cex.lab=0.8)
  hist(x, prob=TRUE, breaks=breaks, main=main, xlab=xlab)
  abline(v=mean(x), col="red", lwd=2)    # sample mean
  abline(v=1/lambda, col="blue", lwd=2) # theroretical mean
  lines(density(x), col="red", lwd=2)    # sample density
  curve(dnorm(x, mean.theoretical, sd.theoretical/sqrt(nexp)), min(x), max(x),
    col="blue", add=TRUE, lwd=3)         # theoretical density
  #rug(quantile(dat), col="red", lwd=2)
  legend("topright", c("Theoretical Mean", "Sample Mean"),
    col=c("blue", "red"), lty=c(1, 1), lwd=c(2, 2), cex=0.5)
  }
```

5.2 Code used to generate the QQ-plots

```r
par(mfrow=c(2, 1))
qqnorm(dat, main = "QQ-Plot (sample size 40)", xlab = "Theoretical Quantiles",
  ylab = "Sample Quantiles", plot.it = TRUE, datax = FALSE)
qqline(dat, col="red")
qqnorm(dat2, main = "QQ-Plot (sample size 1000)", xlab = "Theoretical Quantiles",
  ylab = "Sample Quantiles", plot.it = TRUE, datax = FALSE)
qqline(dat2, col="blue")
```