Statistical interference project1

MS

January 14, 2016

The mean and variance of a exponential distribution

Overview

This project simulates 1000 data series of 40 data-points using the exponential distribution. The mean and variance of each data series is plotted, their distribution is analysed for normal distribution.

Simulation

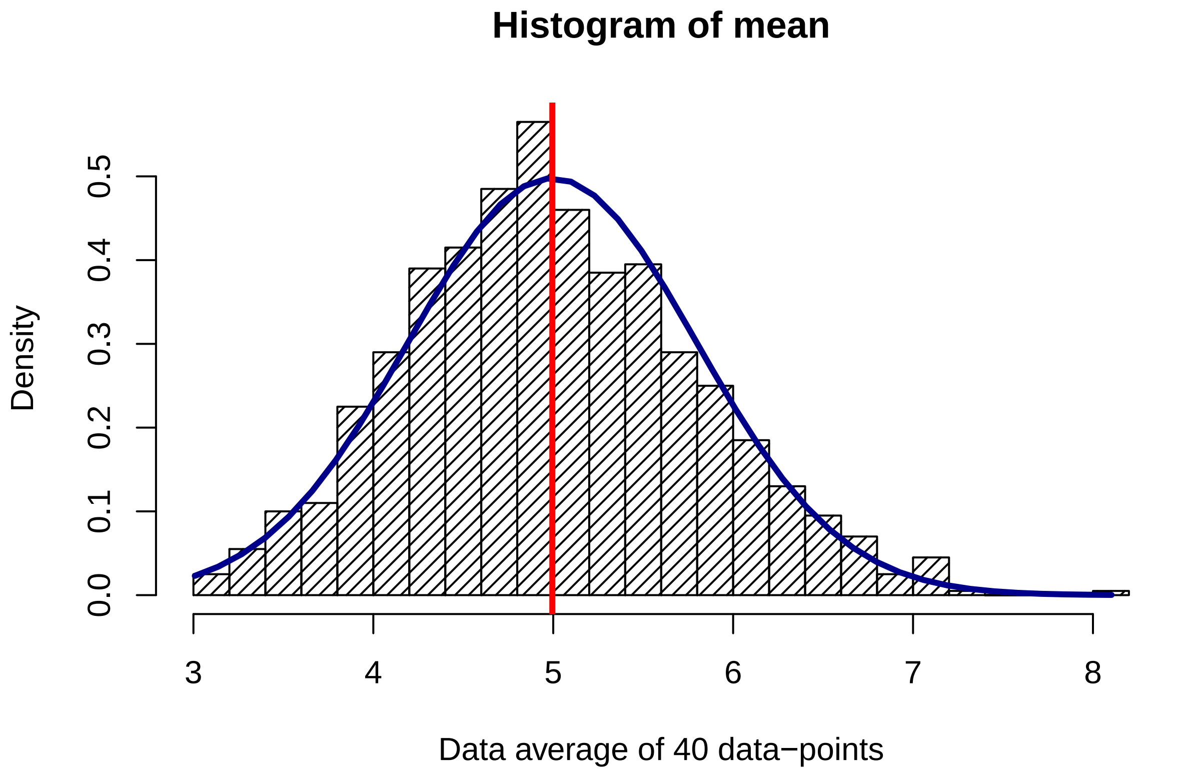
1000*40 data-points is simulated with the exponential distribution with lampda = 0.2. These data points are divided into 1000 data-series with 40 data-points each.

```r
#Set seed to make reproducible simulations
set.seed(31415)
#Set variables
n <- 40
Nsim <- 1000
lampda <- 0.2
#Simulate the data and store each simulation set in rows in a matrix
simData <- matrix(rexp(Nsim*n, lampda), Nsim, n)
```

Sample Mean versus Theoretical Mean

The sample mean is calculated for each data-series and plotted using a histogram.

```r
#Calculate mean of the entire data-set
popMean <- mean(simData)
#Calculate mean of each data series
simMean <- apply(simData,1,mean)
#Calculate the standard error
simSE <- sd(simMean)
#Plot histogram of means
hist(simMean,
     density=20, breaks=20, prob=TRUE, freq = FALSE,
     xlab="Data average of 40 data-points",
     main="Histogram of mean")
x<-seq(min(simMean),max(simMean),length=40)
y<-dnorm(x,popMean,sd=simSE)
lines(x, y, col="darkblue", lwd=3)
#Add line to show the population mean.
abline(v=mean(popMean),col="red", lwd=3)
```



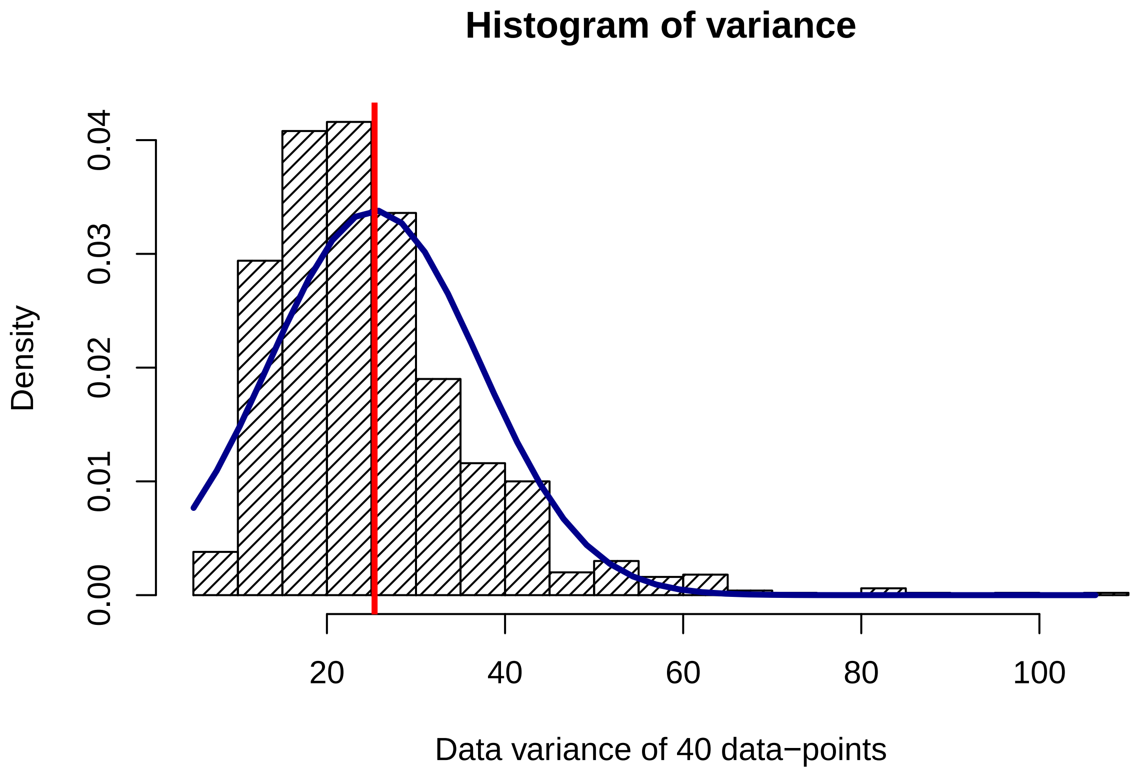
The red line is placed on the full-data set mean (40,000 data points). It is thereby close to the population mean and is calculated to 4.99.

The standard error is calculated as the standard deviation of the means and is 0.8. Using the population mean and the standard error the normal distribution is superimposed on the graf (blue line).

Sample Variance versus Theoretical Variance

The sample variance is calculated for each data-series and plotted using a histogram.

```
#Calculate variance of each data series
simVar <- apply(simData,1,var)
#Calculate the variance of the population
popVar <- sd(simData)^2
#Plot histogram of means
hist(simVar,
     density=20, breaks=20, prob=TRUE, freq = FALSE,
     xlab="Data variance of 40 data-points",
     main="Histogram of variance")
 x<-seq(min(simVar),max(simVar),length=40)
 y<-dnorm(x,popVar,sd=sd(simVar))
 lines(x, y, col="darkblue", lwd=3)
#Add line to show the population variance.
abline(v=mean(popVar),col="red", lwd=3)
```



The red line is placed on the full-data set variance (40,000 data points). It is thereby close to the population mean and is calculated to 25.35.

Using the population variance and the standard error, the normal distribution is superimposed on the graf (blue line).

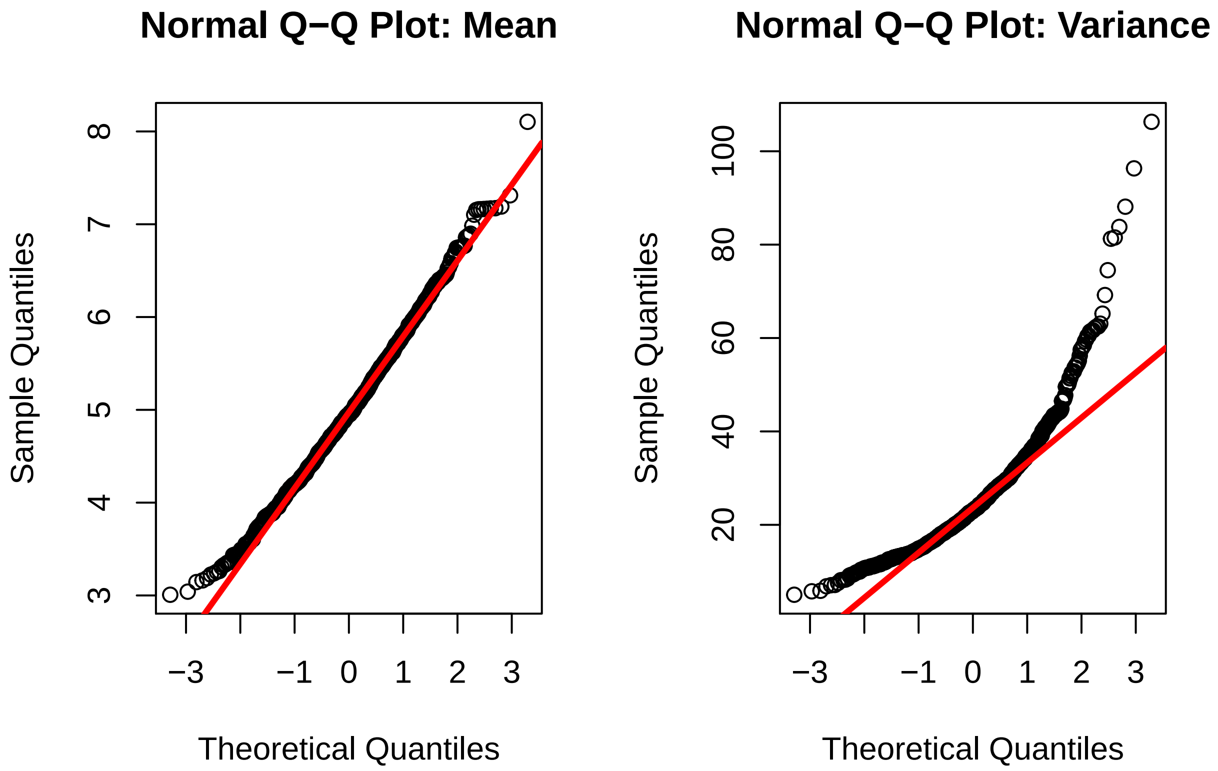
Distribution

The plotted normal distributions correlates well with the histograms. However, the variance distribution looks skewed. Skewness can be emphasized or deminished by the choice of breaks in histogram. Hence, to investigate the distributions are normal, a Shapiro-Wilk test is performed with the null-hypothesis that it is a normal distribution and a required alpha level of 1%.

```r
#Calculate the p-value for a Shapiro-Wilk test
meanSWPValue<-shapiro.test(simMean)$p.value
varSWPValue<-shapiro.test(simVar)$p.value
```

The p value is 0.0012933 and $2.0950152 \times 10^{-27}$ for the mean and variance distributions respectively. Hence, it is <0.01 and we reject the null hypothesis. Thus according the Shapiro-Wilks test the distributions are NOT normal. However, the test might not be appropriate, as it is skewed toward rejecting the null-hypothesis at large n. To better validate the distributions we look at the QQ-plots:

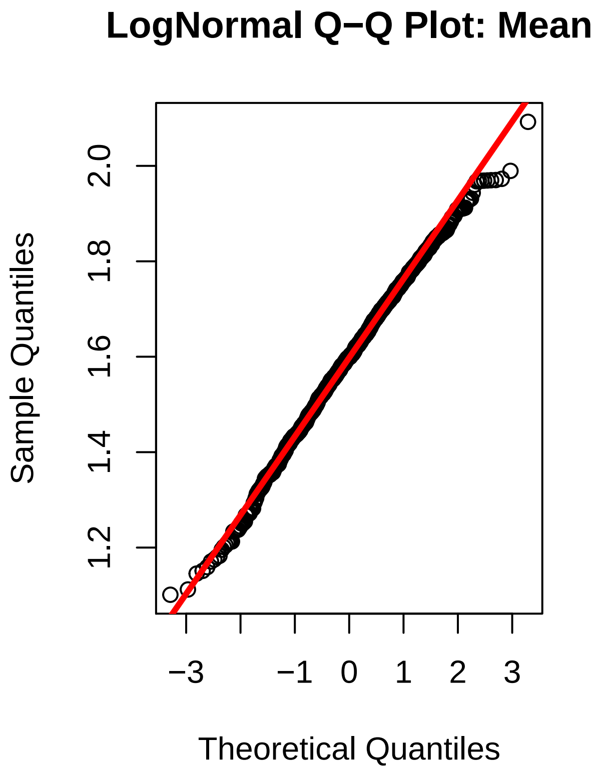
```r
#Plot the QQ-plots side by side
par(mfrow=c(1,2))
qqnorm(simMean, main ="Normal Q-Q Plot: Mean");qqline(simMean, col = 2, lwd = 3)
qqnorm(simVar, main ="Normal Q-Q Plot: Variance");qqline(simVar, col = 2, lwd = 3)
```

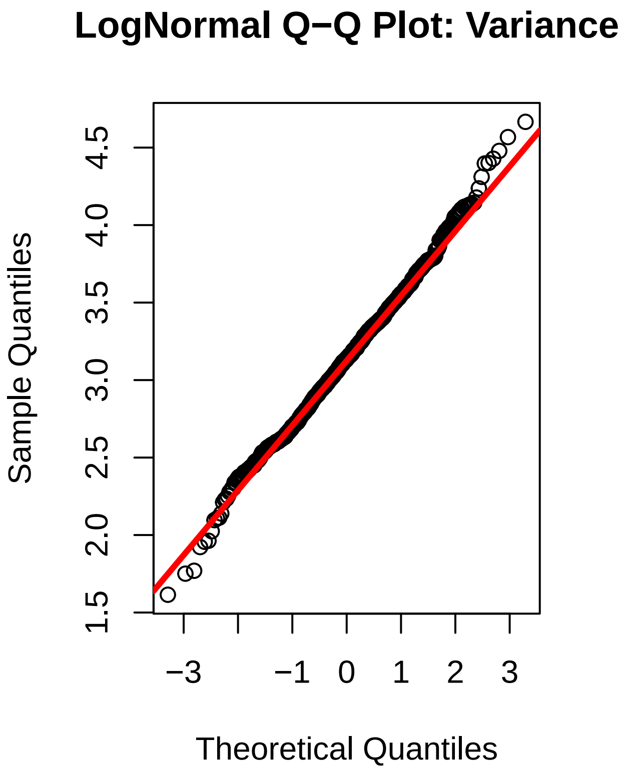


The mean distribution lies close to the theoretical line while the variance distribution show systematic deviation. Hence, we conclude that the mean distribution could be normal while the variance distribution is not.

To do a final test the distributions are tested against a LogNormal distribution - similar to above.

```r
#Calculate the p-value for a Shapiro-Wilk test
logvarSWPValue<-shapiro.test(log(simVar))$p.value
logmeanSWPValue<-shapiro.test(log(simMean))$p.value
#Plot the QQ-plots side by side
par(mfrow=c(1,2))
qqnorm(log(simMean), main ="LogNormal Q-Q Plot: Mean");qqline(log(simMean), col = 2, lwd = 3)
qqnorm(log(simVar), main ="LogNormal Q-Q Plot: Variance");qqline(log(simVar), col = 2, lwd = 3)
```





The p value are 0.0482768 and 0.0487002 for the mean and variance distributions respectively. Now both p values are >0.01. Thus, we fail to reject the null hypothesis and instead accept it: "The mean and variance distributions are Log-Normal distributed". This is also emphasized by the QQ-plots. For the mean, the deviation is only slightly closer to the straight line, however, for the variance, the points now lie close to the straight line.

The log-normal distribution often emerge when we have a finit limit on the values our data can take (e.g. can't be negative, as in this example). As we move away from the limit, the distribution becomes more like a normal distribution. Hence, it makes sense that our distributions are log-normal.