Statistical Inference: Course Project Part1

Author: Jens Berkmann

Introduction

In this report we investigate the distribution of a random variable (RV) Y which is obtained by averaging n=40 iid exponentially distributed random variables X which are parameterized by the parameter $\lambda = 0.2$. The distribution is given by the formula $f(x) = \lambda \exp(-\lambda x)$ for $x >= 0$ and zero elsewhere. The mean of an exponentially distributed RV with parameter λ is $1/\lambda$ where the variance is $1/\lambda^2$. For the parameter choice in this report the exponentially distributed RV has a mean of $1/\lambda = 5$ and a standard deviation of $\sqrt{(}1/\lambda^2) = 5$. The theoretical mean and variance of the averaged RV Y is given by E[Y]=5 and STD[Y]=$\sqrt{(}25/40) = 0.79$. The distribution of the RV Y follows a Gamma distribution. However, the distribution will be investigated by simulation.

Generating data

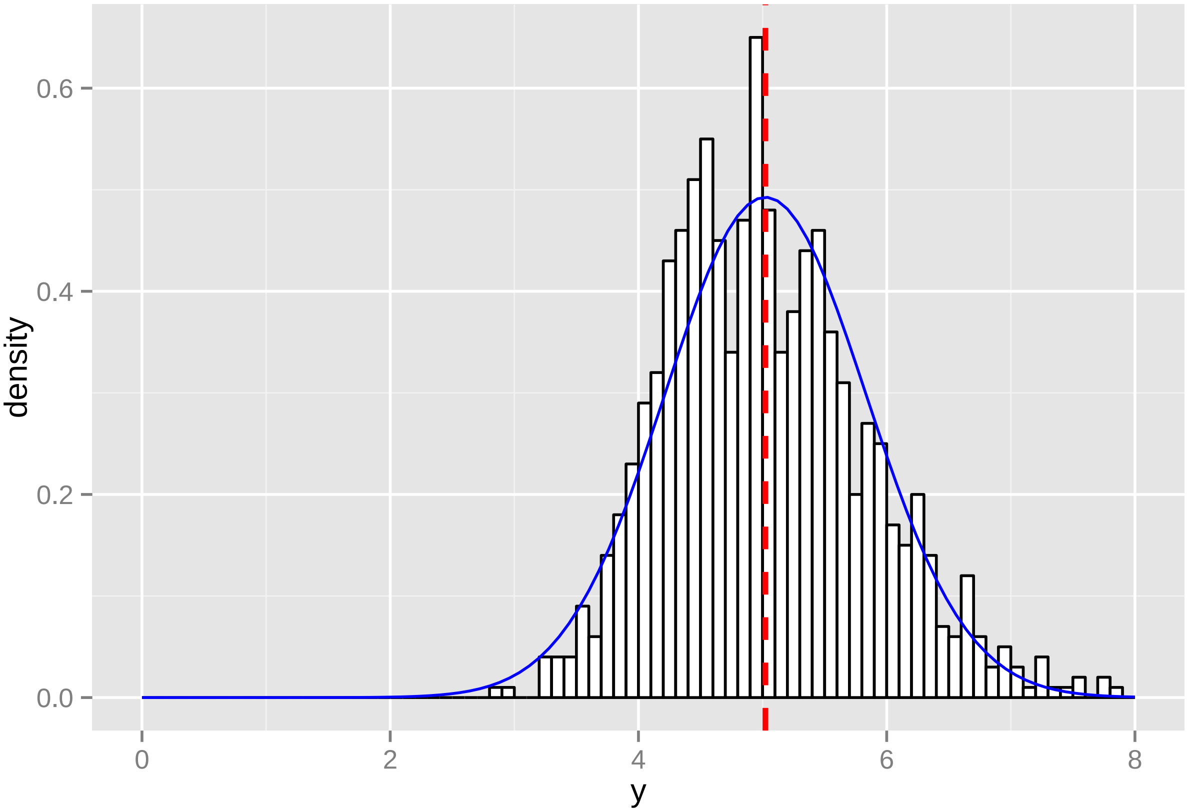
Let's first generate 1000 averaged RVs Y.

```
library(knitr)
lambda    <- 0.2
n         <- 40
nosim     <- 1000
theor_mean<- 1/lambda
theor_sd  <- 1/lambda/sqrt(n)
set.seed(123456)
data      <- matrix(rexp(nosim*n,lambda),nosim) # nosim rows
mean_data <- apply(data,1,mean)                 # 1st dimension (rows) to be retained
emp_mean  <- mean(mean_data)
emp_sd    <- sd(mean_data)
```

Histogram and Gaussian Approximation

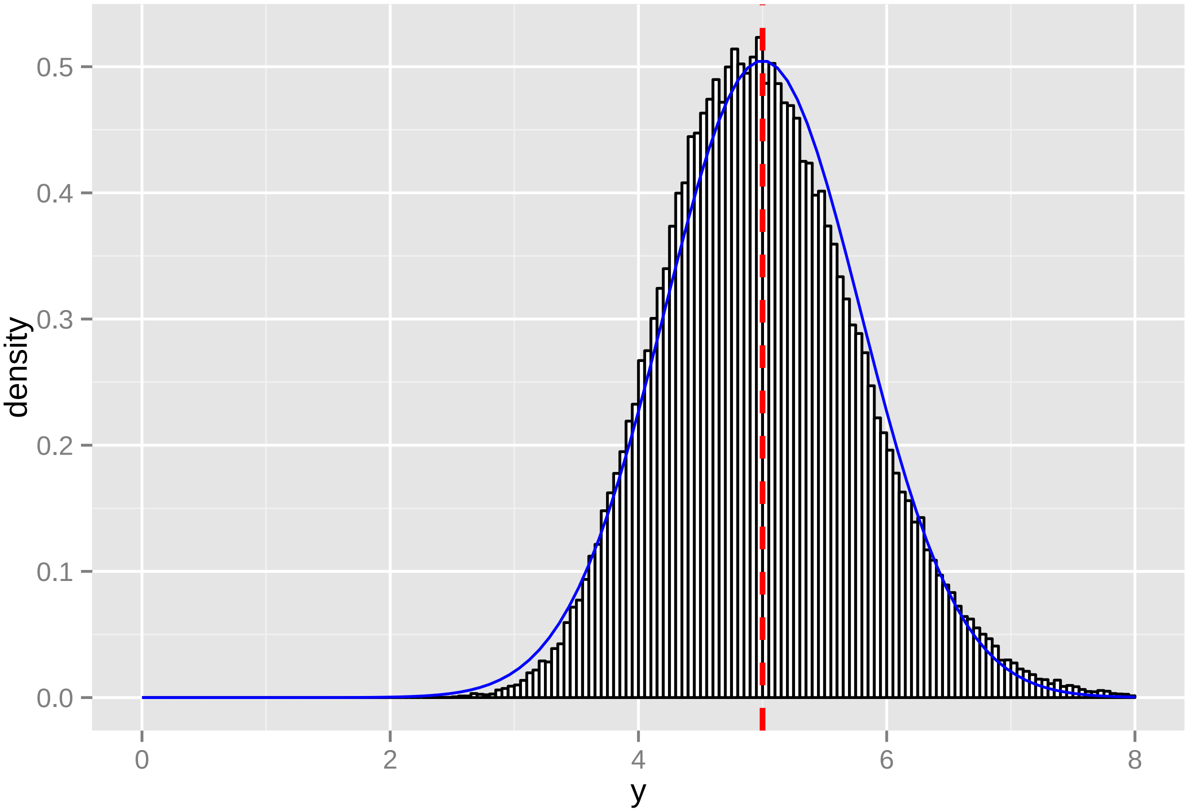
Based on this simulated data the empirical mean is m=5.0229 while the empirical standard deviation is s=0.8098 which is already fairly close to the theoretical expectations. Let's now plot the histogram of the data together with a Gaussian distribution having the same empirical mean and standard deviation as the simulated data.

```
library(ggplot2)
df <- data.frame(y=mean_data)
p<-ggplot(df, aes(x=y)) +
        geom_histogram(
            aes(y=..density..),
            binwidth=0.1,
            colour="black", fill="white") +
            geom_vline(aes(xintercept=mean(y, na.rm=T)),
            color="red", linetype="dashed", size=1) +
            stat_function(fun = dnorm, args=list(mean=emp_mean,sd=emp_sd), colour = "blue") +
            xlim(0,8)
print(p)
```



From this figure one could rush to the conclusion that the Gaussian distribution nicely approximates the simulated data. Before concluding, let us increase the number of simulated RVs Y by 2 decades to *nosim*=100000 and plot again histogram and Gaussian pdf.

```
nosim     <- 100000
theor_mean<- 1/lambda
theor_sd  <- 1/lambda/sqrt(n)
set.seed(123456)
data      <- matrix(rexp(nosim*n,lambda),nosim) # nosim rows
mean_data <- apply(data,1,mean)                 # 1st dimension (rows) to be retained
emp_mean  <- mean(mean_data)
emp_sd    <- sd(mean_data)
df <- data.frame(y=mean_data)
p<-ggplot(df, aes(x=y)) +
        geom_histogram(
            aes(y=..density..),
            binwidth=0.05,
            colour="black", fill="white") +
            geom_vline(aes(xintercept=mean(y, na.rm=T)),
            color="red", linetype="dashed", size=1) +
            stat_function(fun = dnorm, args=list(mean=emp_mean,sd=emp_sd), colour = "blue") +
            xlim(0,8)
print(p)
```



The empirical mean and standard deviation are now m=4.9997 and s=0.7903, respectively, which are closer to the theoretical values as expected. The histogram is much smoother now and shows a fairly well fit to the blue Gaussian pdf. However, still some discrepancy of the pdf under consideraton to a Gaussian pdf is seen. In particular, the pdf is slightly non-symmetric around the mean, is slightly varying less to the left of the mean and slightly varying more to the right of the mean compared to the Gaussian approximation. Only if n would be further increased, the pdf would ultimately be Gaussian distributed according to the central limit theorem. Still, for most cases of interest (also in this case n=40) the approximation as a Gaussian pdf might be sufficiently valid.

Convergence of Confidence Interval

By building the mean of n=40 iid exponentially distributed RVs we have actually constructed a particular estimator of the mean of a exponentially distributed RV. The question arises how well this estimator works. For answering this question we evaluate the coverage of the confidence interval for $1/\lambda = 5$, i.e. we assess by repeated simulation of the mean estimator in action what percentage of the estimated means fall into the confidence interval $\bar{X} \pm 1.96\frac{S}{\sqrt{n}}$. If the central limit theorem would sufficiently apply already for n=40, the pdf of the mean-estimated RV Y would be approximately Gaussian and the confidence intervall corresponds to a confidence level of 0.95. This means that roughly 95% of all estimated means fall into the confidence intervall around the true mean $1/\lambda$. In the previous section we have seen that the RV Y looks approximately Gaussian. Let us now evaluate what percentage of means fall into the confidence interval by running the code snippet below. We simulate again N=1000 mean estimations $\hat{X} = Y$. We also determine the empirical standard deviation from the data for each of N experiments and compute the confidence intervals lower and upper limits. Finally, we estimate the probability that the estimated means fall within the confidence interval(s).

```
nosim      <- 1000
data       <- matrix(rexp(nosim*n,lambda),nosim) # nosim rows
mean_data <- apply(data,1,mean)                  # 1st dimension
sd_data    <- apply(data,1,sd)
lolim      <- mean_data - 1.96*sd_data/sqrt(n)
uplim      <- mean_data + 1.96*sd_data/sqrt(n)
conf1       <- mean(lolim < 1/lambda &  uplim > 1/lambda)
```

When running the code we obtain a confidence level of 0.926% which is close to the expected confidence level of 0.95 but not
sufficiently close. We would like to check whether the estimated confidence level was just by accident smaller than 0.95 possibly because too few experiments (N=1000) were done. Therefore we repeate the expirement with N=100000 and obtain a confidence level of 0.9252%. We conclude that the observed confidence level does indeed not reach the expected level 0.95. We attribute this behaviour to the fact that the Gaussian assumption is still not good enough for n=40 in order to accurately predict the confidence level of the mean estimation.