Sampling data from an exponential distribution: case study

JoeriW

Wednesday, March 11, 2015

Overview

We'll investigate the distribution of 40 variables drawn from an exponential distribution (with a lambda of 0.2) by comparing the sample mean to the theoretical mean, show how variable the sample is (via variance) and compare it to theoretical variance and demonstrate the obtained distribution is approximately normal

Preparation

Load the required packages:

```
library(ggplot2)
```

Simulation

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. Lambda will be set **0.2** for all the simulations. The number of simulations is **1000**

set the parameters:

```
set.seed(4000); nosim = 1000;lambda = 0.2;noexp = 40
```

create a data frame containing the simulated data:

```
dat <- data.frame(sim.means =
c(apply(matrix(rexp(nosim*noexp,lambda),nosim),1,mean)))
```

Sample Mean versus Theoretical Mean:

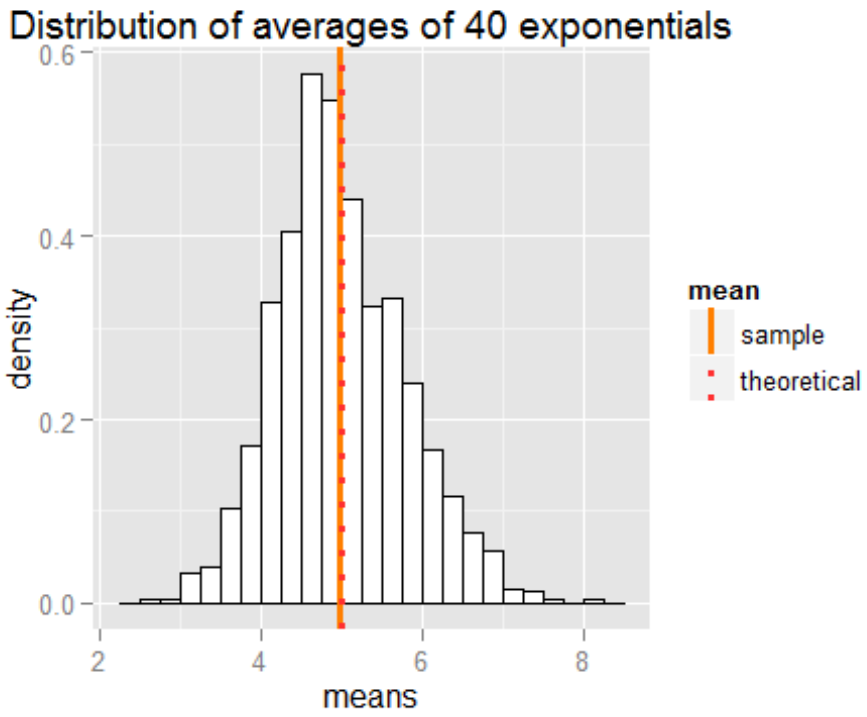
calculate the sample mean and the theoretical mean and store them in a data.frame (in order to plot the legend in ggplot)

```
sample.mean <- mean(dat$sim.means)
theo.mean <- 1/lambda
vlines.df <-data.frame(mean =
c("theoretical","sample"),value=c(theo.mean,sample.mean))
```

plot the obtained distribution and add the theoretical and sample mean on the graph:

```
g1 <- ggplot(data = dat, aes(x = sim.means)) +
geom_histogram(aes(y=..density..), binwidth=.25, colour="black",
fill="white")
g1 <- g1 +
geom_histogram(aes(y=..density..),binwidth=.25,colour="black",fill="white")
g1 <- g1 + geom_vline(data = vlines.df,aes(xintercept = value,colour =
mean,linetype = mean),show_guide = T,size = 1.1)
```

```
g1 <- g1 + scale_linetype_manual(values=c("solid", "dotted")) +
scale_colour_manual(values = c("darkorange1","firebrick1"))
g1 <- g1 + labs(title="Distribution of averages of 40 exponentials") +
xlab("means")
g1
```



When lambda is **0.2**, the theoretical mean of the exponential equals **5** (= 1/lambda). The sample mean of our simulations excercise equals **4.9884633**.

Sample Variance versus Theoretical Variance

Calculate the variance of the sample data

```
sample.var <- var(dat$sim.means)
```

The theoretical variance of the distribution of 40 exponentials equals the population variance (1/lambda^2) divided by the sample size. So given a lambda of **0.2** and a sample size of **40**:

theoretical variance= 0.625
theoretical standard deviation= 0.7905694

We compare this to the variance and standard devation of the sample:

sample variance= 0.6335151
sample standard deviation= 0.7959366

The theoretical variance and sample variance,as expected, are very close to each other.

Distribution

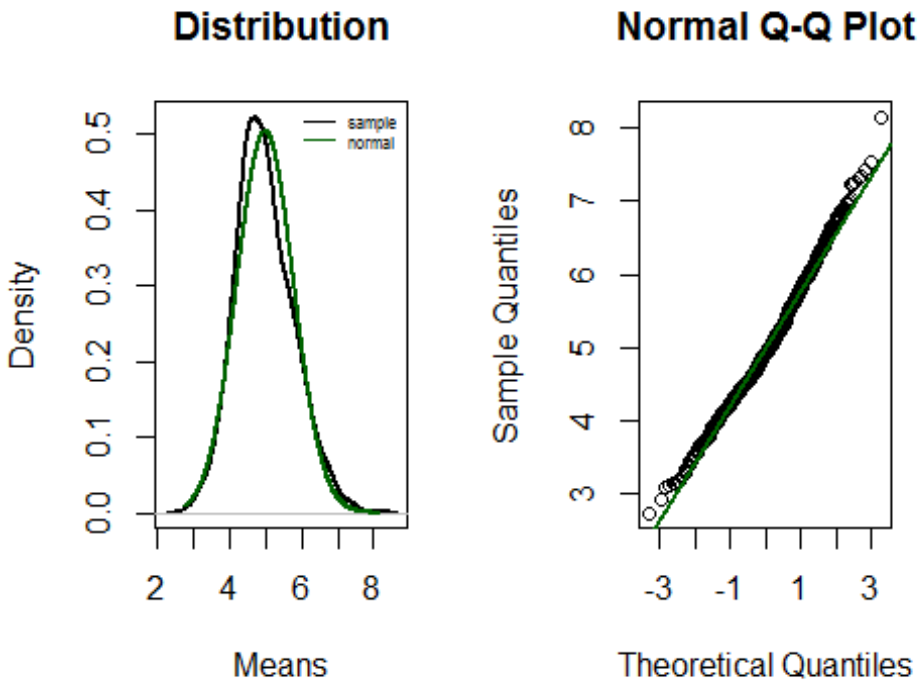
create a density plot and a q-q plot

```r
par(mfrow=c(1,2),mar = c(5,4,4,2))
plot(density(dat$sim.means),xlab = "Means",ylab = "Density",main =
"Distribution",lwd = 2)

lines(seq(min(dat$sim.means),max(dat$sim.means),0.01),dnorm(seq(min(dat$sim
.means),max(dat$sim.means),0.01),1/lambda,(1/lambda)/sqrt(noexp)),col="dark
green",lwd=2)

legend("topright",c("sample","normal"),lty = 1,lwd = 1.5,col =
c("black","darkgreen"),bty = "n",cex = 0.5)

qqnorm(dat$sim.means)
qqline(dat$sim.means,col = "darkgreen",lwd = 2)
```



Both plots confirm that the distribution of the sampled data is approximately normal.The density plot is rather self-explanatory in the sense that is plots the normal distribution over the distribution of the sample.The q-q plot plots the quantiles of the normal distribution agains the the quantiles of the sampled data. The linearity of the points suggests that the sample data is normally distributed.

Conclusion

The simulations are a clear application of the Central Limit Theorem, which states that the distribution of averages of independent and identically distributed variables becomes that of a normal distribution (given that the sample size is large enough).