Simulation exercises

21/09/2014

Simulation

In this project we will simulate averages of exponential distribution with parameter $\lambda = 0.2$. The mean and the stardard deviation are both $\mu = \sigma = \lambda = 0.2$. The number of outcomes that we will generate is $n = 40$, and we will generate 1000 averages of 40 exponential distributed outcomes. A seed is set so that the project is fully reproducible. The 1000 means are stored in a data frame called x.

According to the central limit theorem, the mean of the 1000 averages will be the mean of the distribution λ, and the standard deviation will be $\frac{\sigma}{\sqrt{(n)}}$, where σ is the standard deviation of the distribution.

The mean

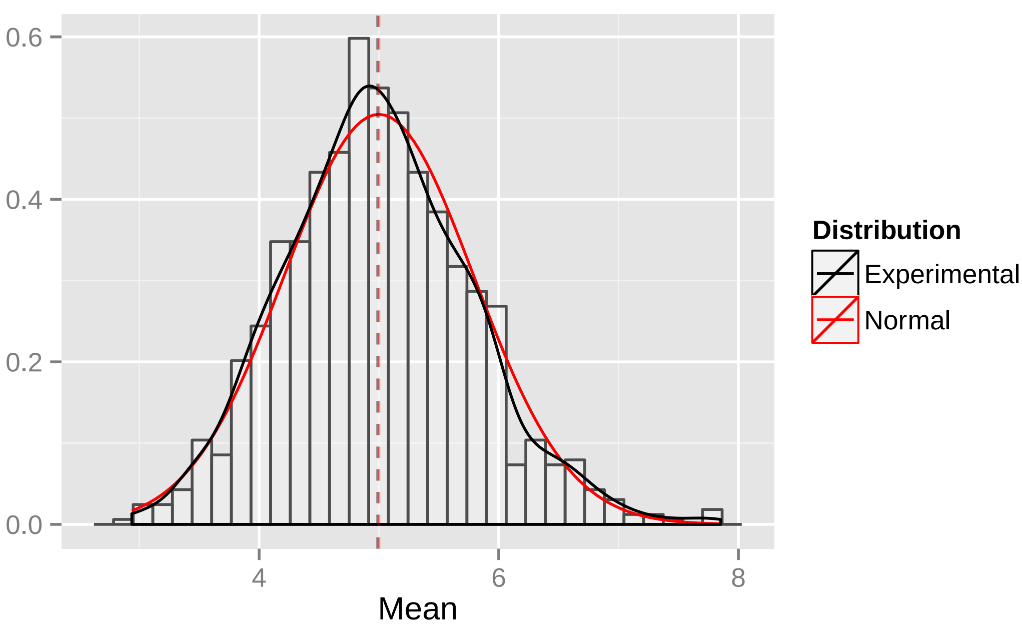
The theoretical mean is `tmean = 5`, and the experimental value is `mean(x$means) = 4.9897`. As we can see, both values are similar. We can compute the relative percent error $100 * |\frac{\mu - \bar{\mu}}{\mu}|$, which yields the value `abs(tmean-mean(x$means))/tmean*100 = 0.2057%`. The relative percent error is small enough for us to conclude that $\bar{\mu} \approx \mu$.

Variance

We can compare the theoretical and experimental variances in the same way as we did for the mean. The values are `tvar = 0.625` for the theoretical variance and `var(x$means) = 0.6037` for the experimental one. The relative percent error is `abs(tvar-var(x$means))/tvar*100 = 3.4141%`, again small enough.

Is the distribution approximately normal?

In order to see if the distribution is approximately normal, we can plot the density of the experimental distribution of the 1000 averages and compare it to the density of a normal distribution with parameters $\mu = 5$ and $\sigma = 0.7906$.



The plot shows the histogram of the experimental values and their density distribution in black. In red, a normal distribution with the theorical values obtained by the central limit theorem is shown. As we can see, the distributions are similar. The differences are due to the fact that we have simulated 1000 averages. If the number of averages were higher, the experimental distribution would get closer to the normal distribution. The plot also shows the means of both distributions, in dashed lines, which appear superimposed since their value is almost the same.

Coverage of the confidence interval

If we evaluate the confidence interval for each average, $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{(n)}}$, we will obtain 1000 95% confidence intervals. This means that, according to the central limit theorem, 95% of the intervals should contain the actual value of the mean $\mu = 0.2$. If we construc the intervals and calculate the percentage of them that contain the mean we get

```
confs <- confs <- data.frame(low = x$means -1.96*tsd, up = x$means +1.96*tsd)
mean(tmean > confs[, 1] & tmean < confs[, 2])
```

```
## [1] 0.951
```

which is a value close to the 95% theoretical value that we should obtain.

Appendix

Code

This code produces all the project outcomes and is fully reproducible.

```
# Simulation
lambda <- 0.2
n <- 40
set.seed(1899)
x <- data.frame(means = replicate(1000, mean(rexp(n, lambda))))

# Theoretical values
tmean <- 1/lambda
tvar <- 1/(lambda^2*n)
tsd <- 1/(lambda*sqrt(n))

# Where the distribution is centered and compare to the theoretical value
mean(x$means) # Experimental value
tmean # Theoretical value
abs(tmean-mean(x$means))/tmean*100 # Relative percent error

# How variable it is and compare it to the theoretical variance
var(x$means) # Experimental value
tvar # Theoretical value
abs(tvar-var(x$means))/tvar*100 # Relative percent error

# Show that the distribution is approximately normal.
# Histogram + densities + means
ggplot(x, aes(means)) + geom_histogram(aes(y = ..density..), colour = "grey30",
                                       fill = "white", alpha = .3) +
    stat_function(fun = "dnorm", args = list(mean = tmean, sd = tsd),
                  aes(colour = "Normal")) +
    geom_density(aes(colour = "Experimental")) +
    geom_vline(aes(xintercept = mean(x$means)),
               colour = "black", linetype = "dashed", alpha = .4) +
    geom_vline(aes(xintercept = tmean),
               colour = "red", linetype = "dashed", alpha = .4) +
    scale_colour_manual("Distribution", values = c("black", "red")) +
    xlab("Mean") + ylab("")

# Evaluate the coverage of the confidence interval for 1/lambda
confs <- confs <- data.frame(low = x$means -1.96*tsd, up = x$means +1.96*tsd)
# Percentage of intervals that contain the actual mean
mean(tmean > confs[, 1] & tmean < confs[, 2])
```