Exponential distribution analysis

Pier Lorenzo Paracchini

Overview

The Central Limit Theorem (CLT) states that *"the distribution of averages of indipendent and identically distributed (iid) variables becomes that of a standard normal as the sample size increase."* so in this assignment it will be investigate the behaviour of the **exponential distribution** when looking at the distribution of averages of 40 exponentials (and more).

Simulations

Lets simulate an experiment with a random variable X with an **exponential distribution** with the following characteristics

- **rate parameter** $\lambda = 0.2$
  - **mean** $\mu = E[X_i] = 1/\lambda = 5$
  - **standard deviation** $\sigma = \sqrt{2 \mathrm{x} \mathrm{Var}(X_i)} = 1/\lambda = 5$
  - **standard error** (SE) $\sigma/\sqrt{2n}$

For building the distribution of averages we will run thousands of simulation and for each simulation the following steps are executed:

- run the experiment n time where $n = 40,\ 80,\ 120,\ 160,$
  - be X_i the outcome for the experiment $i = 1..n$.
- calculate the mean $\overline{X}_n$ using $X_1, X_2, .., X_{40}$
- calculate $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$

The following the code snippet can be used to generate the simulation data.

```
nosim = 1000
n = 40
clt_func <- function(x, n) sqrt(n) * (mean(x) - 5) / 5

dat <- data.frame(
  x = c(
    apply(matrix(rexp(n * nosim, 0.2), nrow = nosim), MARGIN = 1, clt_func, n),
    apply(matrix(rexp((2*n) * nosim, 0.2), nrow = nosim), MARGIN = 1, clt_func, (2*n)),
    apply(matrix(rexp((3*n) * nosim, 0.2), nrow = nosim), MARGIN = 1, clt_func, (3*n)),
    apply(matrix(rexp((4*n) * nosim, 0.2), nrow = nosim), MARGIN = 1, clt_func, (4*n))
    ),
  size = factor(rep(c(n, 2*n, 3*n, 4*n), rep(nosim, 4)))
  )
```

Sample Mean versus Theoretical Mean

Based on the **CLT** the **theoretical mean** is $\mu = 0$ and the value of the sample mean is **0.0506595** for n = 40. The **sample mean** is around the **theoretical mean** for n = 40 and it converges more and more to it increasing the size of n (n = 80, 120, 160).

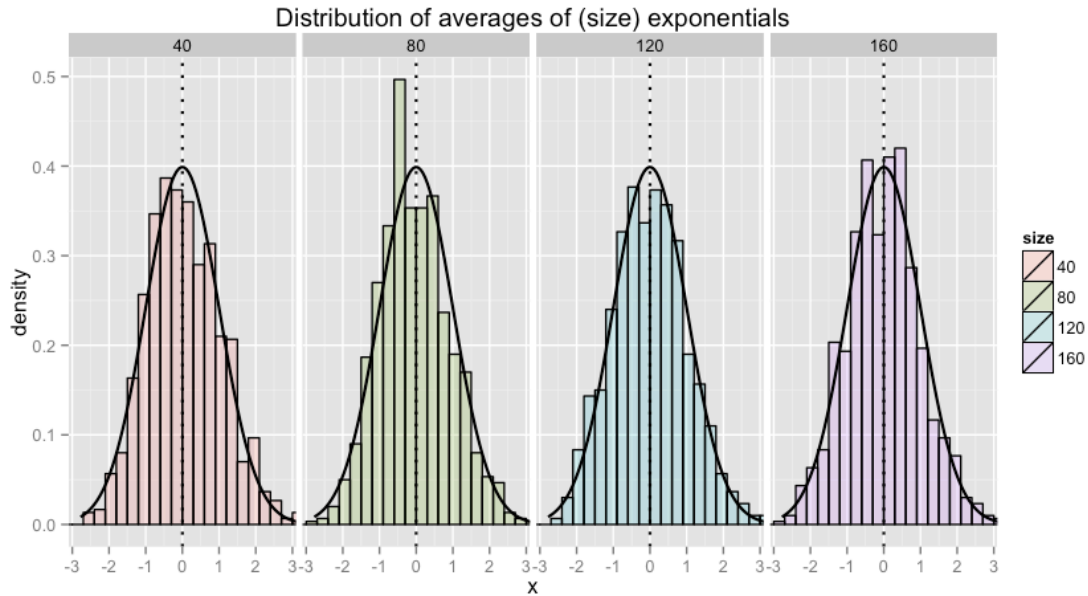Sample Mean(black) vs. Theoretical Mean(red)
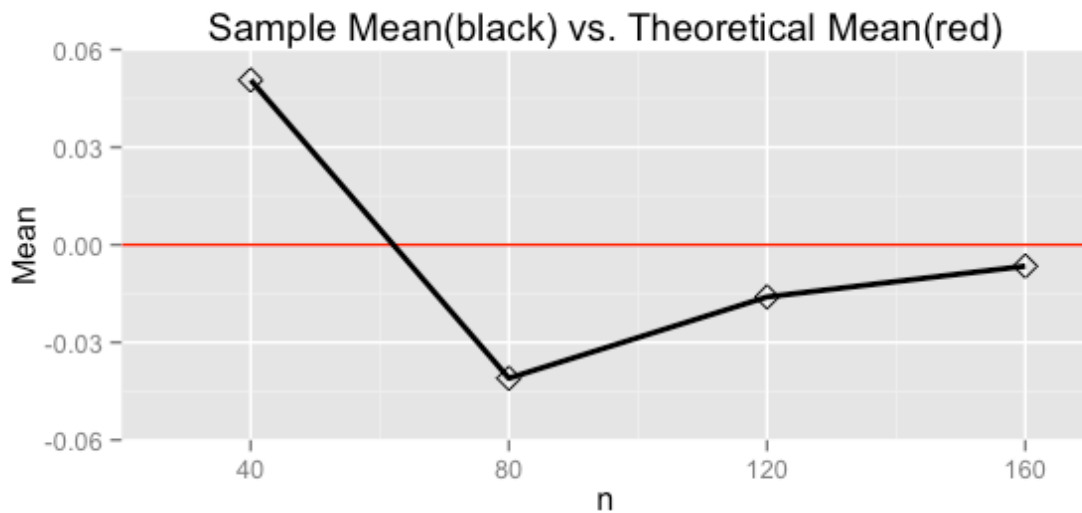
Sample Variance versus Theoretical Variance

Based on the **CLT** the **theoretical variance** is $\sigma^2 = 1$ and the value of the sample variance is **1.0607617** for n = 40. The **sample variance** is around the **theoretical variance** for n = 40 and it converges more and more to it increasing the size of n (n = 80, 120, 160).

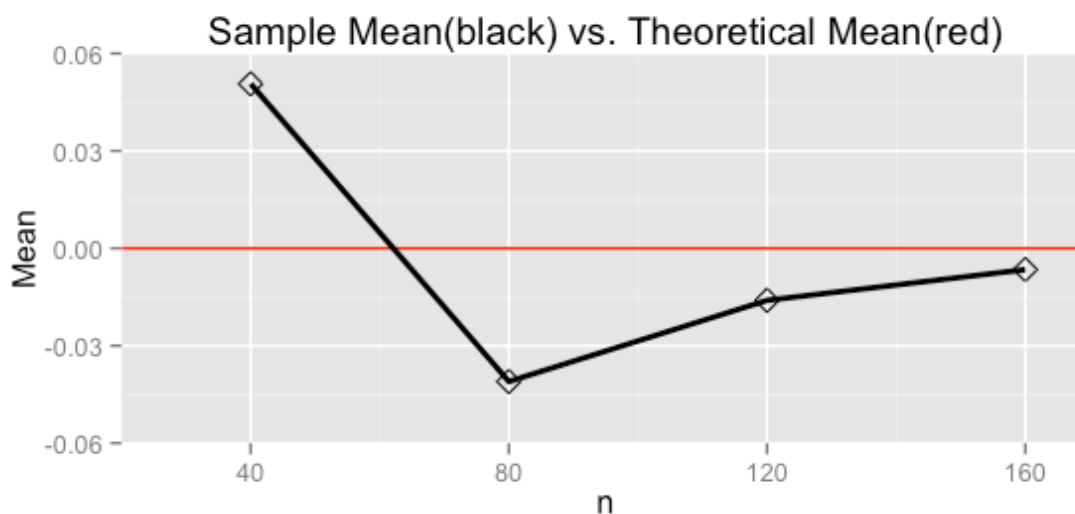Sample Variance(black) vs. Theoretical Variance(red)
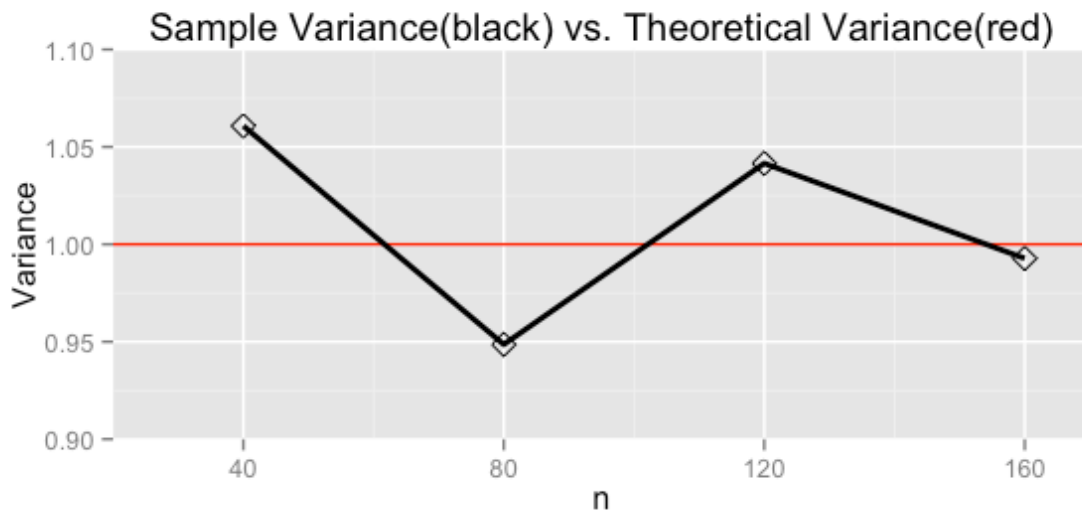
Distribution

Lets plot the distribution of the averages previously calculated over all of the simulations together with the standard normal distribution N(0,1). We can see from the plot below, the distribution of averages $(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}})$ is a good approximation of a standard normal distribution N(0,1) (black curve) as expected from the Central Limit Theorem.

Appendix

Code chunk used to generate plot in **Sample Mean versus Theoretical Mean** section.

```
meanSims <- c(
    mean(dat[dat$size == n,]$x),
    mean(dat[dat$size == (2*n),]$x),
    mean(dat[dat$size == (3*n),]$x),
    mean(dat[dat$size == (4*n),]$x)
    )
g <- ggplot(data.frame(x = c(n, 2*n, 3*n, 4*n), y = meanSims), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0, color = "red") + geom_line(size = 1) + geom_point(shape = 5, s
ize = 3)
g <- g + labs(x = "n", y = "Mean", title ="Sample Mean(black) vs. Theoretical Mean(red)")
g <- g + coord_cartesian(xlim= c(20,170), ylim=c(-0.06,0.06))
g
```


Code chunk used to generate plot in **Sample Variance versus Theoretical Variance** section.

```
varianceSims <- c(
    sd(dat[dat$size == n,]$x)^2,
    sd(dat[dat$size == (2*n),]$x)^2,
    sd(dat[dat$size == (3*n),]$x)^2,
    sd(dat[dat$size == (4*n),]$x)^2
    )
g <- ggplot(data.frame(x = c(n, 2*n, 3*n, 4*n), y = varianceSims), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 1, color = "red") + geom_line(size = 1) + geom_point(shape = 5, s
ize = 3)
g <- g + labs(x = "n", y = "Variance", title ="Sample Variance(black) vs. Theoretical Variance(red
)")
g <- g + coord_cartesian(xlim= c(20,170), ylim=c(0.9,1.1))
g
```

Sample Variance(black) vs. Theoretical Variance(red)
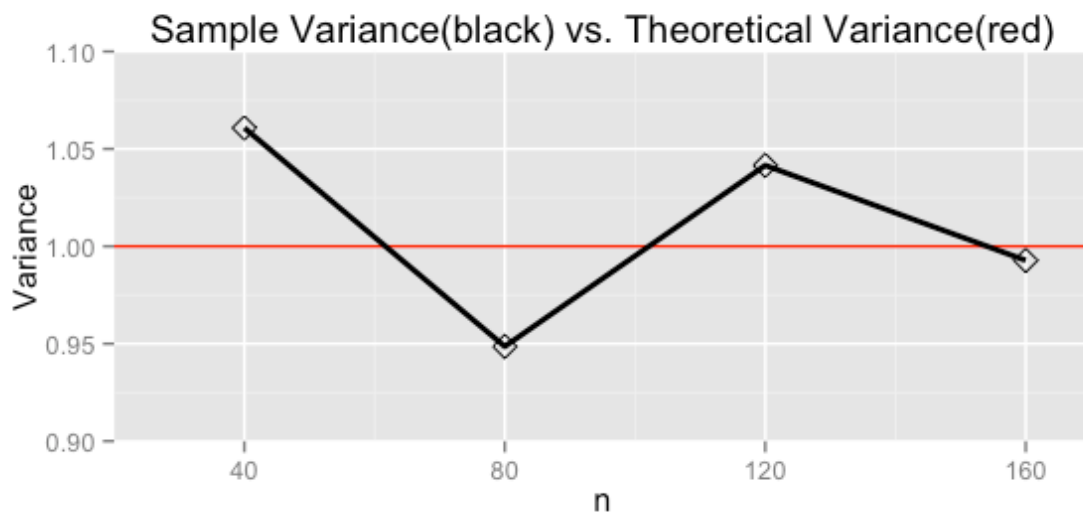
Code chunk used to generate plot in **Distribution** section.

```
g <- ggplot(dat, aes(x = x, fill = size)) + geom_histogram(alpha = .20, binwidth=.3, colour = "bla
ck", aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 0.8)
g <- g + geom_vline(xintercept=c(0,0), color="black", linetype="dotted", size = 0.8)
g <- g + coord_cartesian(xlim= c(-3.1,3.1)) + labs(title = "Distribution of averages of (size) exp
onentials")
g + facet_grid(. ~ size)
```


Distribution of averages of (size) exponentials