Course Project: Statistical Inference - Part 1

iyermobile

Dec/21/2014

Repository URL : [https://github.com/iyermobile/Statistical-Inference (https://github.com/iyermobile/Statistical-Inference)]

Project Goal:

This is the project for the statistical inference class. In it, you will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

About Course Objective:

The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter. The mean of exponential distribution is `1/lambda` and the standard deviation is also also `1/lambda`. Set `lambda = 0.2` for all of the simulations. In this simulation, you will investigate the distribution of averages of 40 exponential(0.2)s. Note that you will need to do a thousand or so simulated averages of 40 exponentials.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s. You should

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.
2. Show how variable it is and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Solution:

Evaluate the coverage of the confidence interval for `1/lambda`: $\bar{X} \pm 1.96 \frac{s}{\sqrt{n}}$ (This only needs to be done for the specific value of `lambda`).

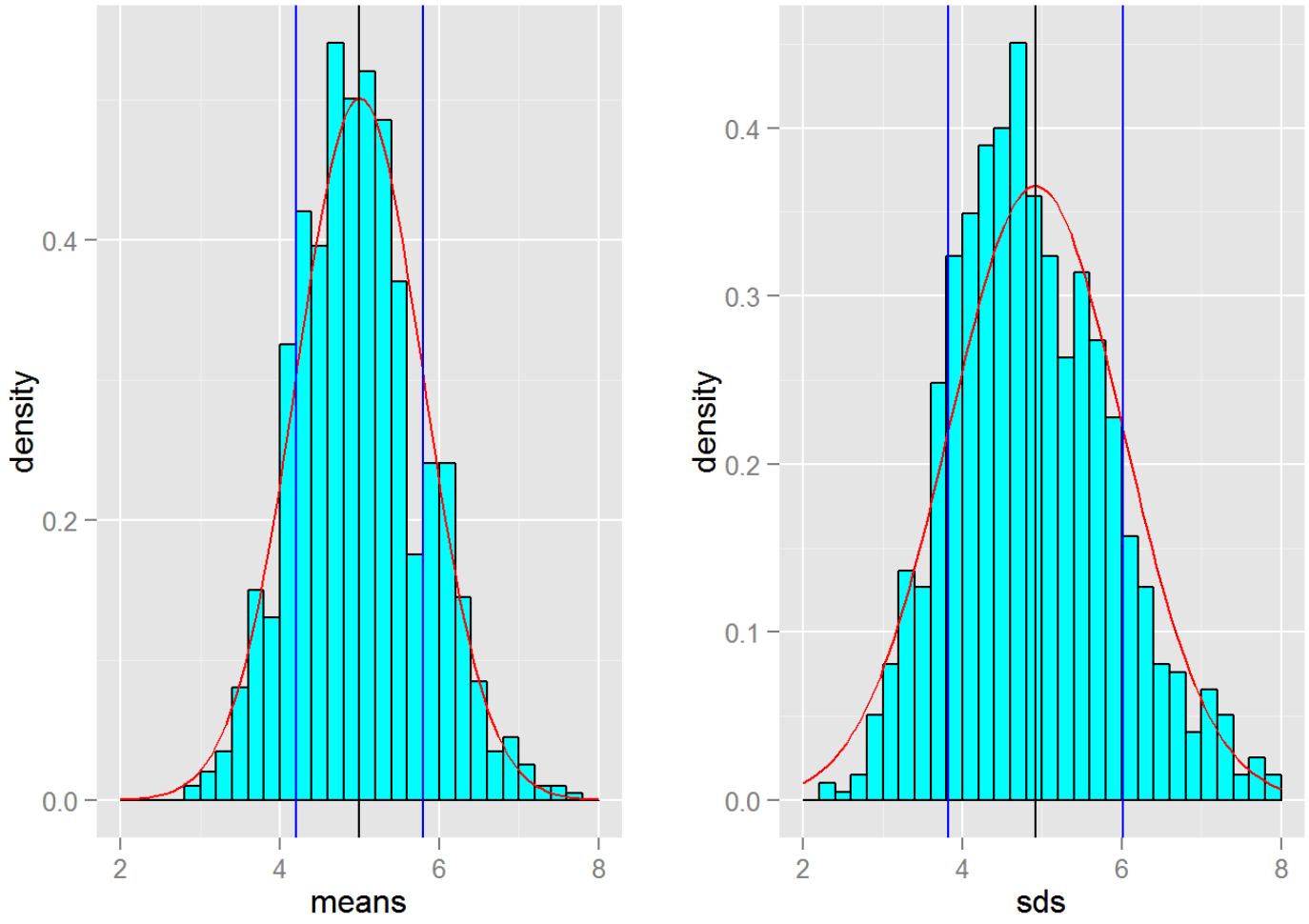
First, generate the data: 1000 trials of 40 exponentially distributed random variables. The mean and standard deviation of the set of 40 variables for each trial is stored in `dataset`.

```
set.seed(100)
lambda = 0.2
n = 40
trials = 1000
data = matrix(rexp(n*trials, lambda), nrow = trials, ncol = n)
dataset = data.frame(means = apply(data, 1, mean), sds = apply(data, 1, sd))
```

The theoretical mean is `1/lambda` = 5 while the mean over all trials is 4.9997019. The trial means are distributed following the CLT, so they should be normally distributed around 5 with a standard deviation of $5/\sqrt{40} = 0.7905694$. Experimentally, the standard deviation of the means is 0.7959461. The distribution

is shown below, on the left, and it is approximately normally distributed because of the CLT. On the right, I examine the distribution of the standard deviation of the 1000 trials of the 40 exponential random variables. Theoretically, the standard deviation is `1/lambda` = 5 and because of the CLT, I expect the same normally distributed experiments around 5 with a standrad deviation of 0.7905694. From the experimental data, the standard deviations are centered around 4.9185188 with a standard deviation of 1.0918739. The plot below demonstrates that they standard deviations are approximately normally distributed in agreement with the CLT. The standard deviation of the exponential distribution is much more variable than the mean.

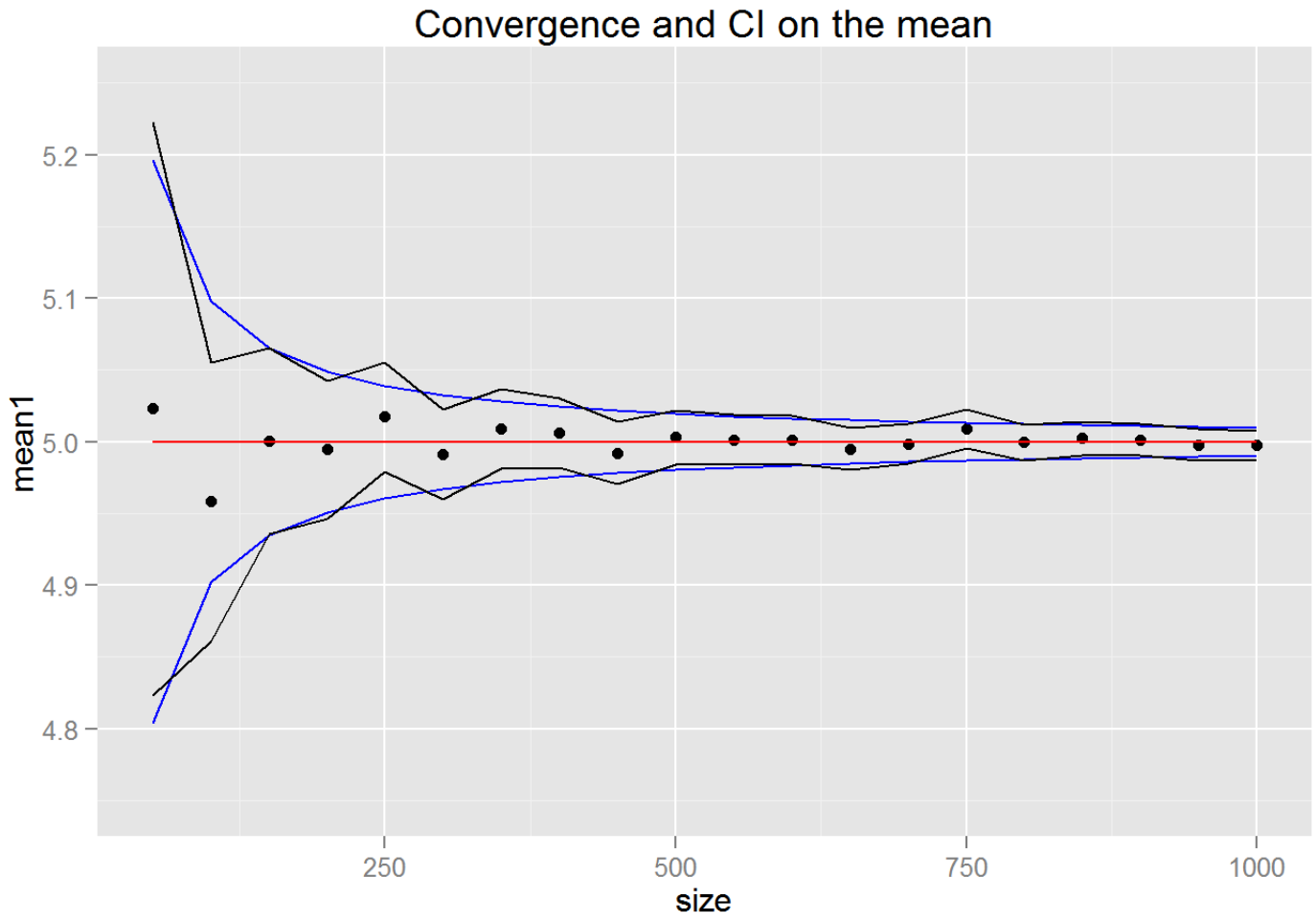
```
## Loading required package: gridExtra
## Loading required package: grid
```



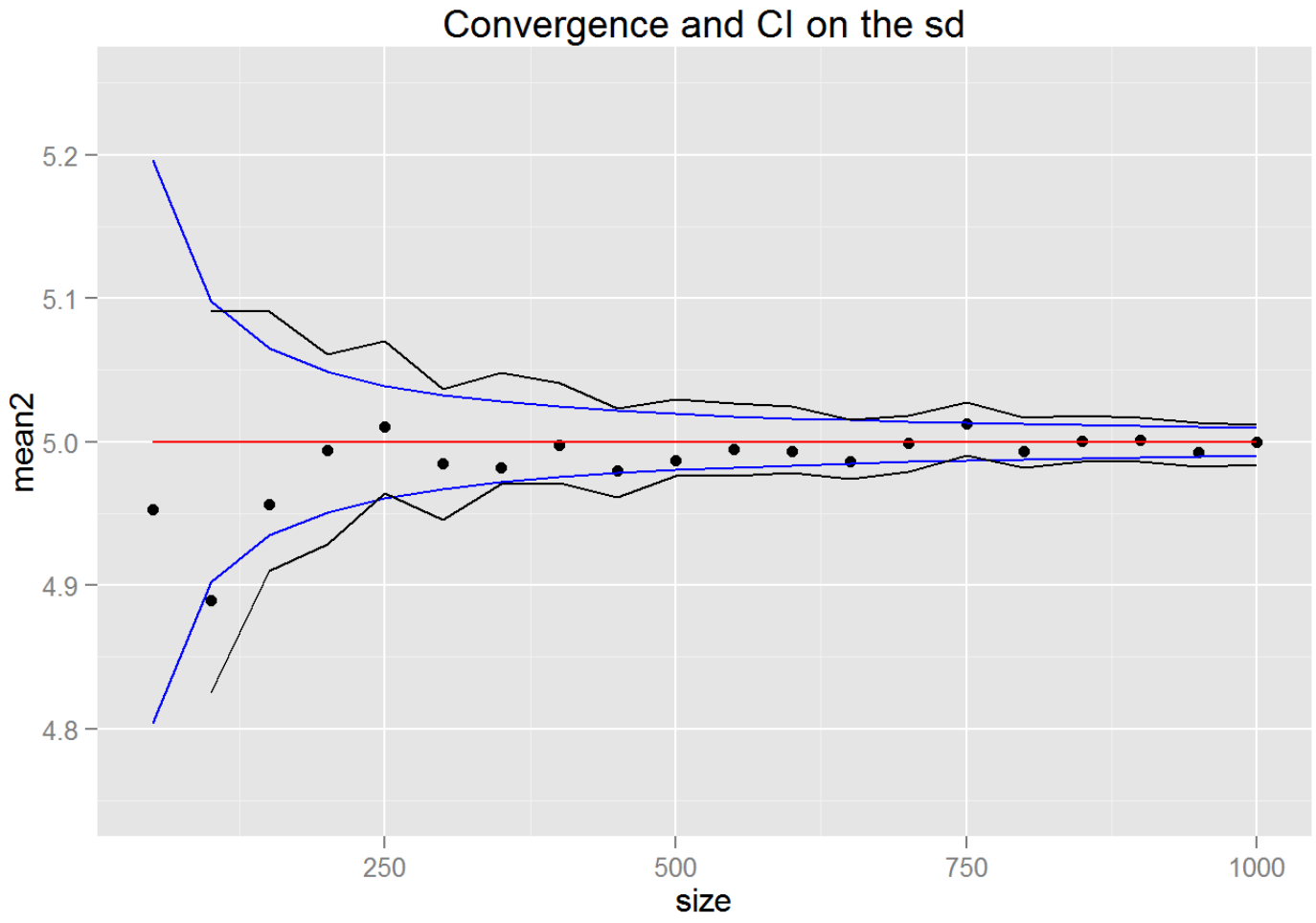
Since the above has been computed, we compute the convergence of the exponential distribution's mean by repeating 1000 trials of an increasing number of exponential random variables and plotting the confidence interval around the mean.

```
dataset = data.frame()
for (n in (1:20)*50) {
  data = matrix(rexp(n*trials, lambda), nrow = trials, ncol = n)
  means = apply(data, 1, mean)
  sds = apply(data, 1, sd)
  dataset = rbind(dataset, data.frame(means = means, sds = sds, size = n))
}
datamean = aggregate(. ~ size, dataset, mean)
datasd =   aggregate(. ~ size, dataset, sd)
dataBySize = cbind(datamean, datasd)
names(dataBySize) <- c("size", "mean1", "mean2", "null", "sd1", "sd2")
dataBySize$null <- NULL
```

The means are represented below. The points are the experimental means and the experimental CI in shown as black lines. The theoretical mean is shown in red and the theoretical CI is shown in blue.



The standard deviations are represented below. The points are the experimental sd and the experimental CI in shown as black lines. The theoretical sd is shown in red and the theoretical CI is shown in blue.


Convergence and CI on the sd
mean2
5.2
5.1
5.0
4.9
4.8
250
500
750
1000
size