Statistical Inference Project - Part 1

Mariame M

September 11, 2014

In this project, we will investigate the distribution of averages of 40 exponential with a rate lambda = 0.2 by running 1,000 simulations of 40 exponentials.

First, we load the required libraries, in this case ggplot2 only, and we set the seed so that we are able to reproduce our results and finally :

```
library(ggplot2)
set.seed(405)
```

We can then set the parameters and run the simulations. The raw data of the simulations is stored in a data frame 'data' and the sampling distribution of means for our simulations is stored in the data frame 'means':

```
# number of simulations
nosim <- 1000
# sample size
n <- 40
# rate lambda
lambda <- 0.2
# build a matrix of size (nosim x n)
data <- matrix(rexp(nosim * n, lambda), nosim)
# compute means for each simulation
means <- apply(data, 1, mean)
```

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

The theoritical mean (mu) of the exponential distribution with lambda rate 0.2 is $\frac{1}{\lambda} = \frac{1}{0.2} = 5$. We compute the mean of each simulation and then compute the observed mean (xbar). We can see that both values are pretty close, which can make us make the hypothesis that our sample mean is a good estimate of the population mean.

```
# compute theoritical mean
mu <- 1 / lambda
mu
```

```
## [1] 5
```

```
# compute observed mean
xbar <- mean(means)
xbar
```

```
## [1] 4.986
```

2. Show how variable it is and compare it to the theoretical variance of the distribution.

Given that the standard deviation σ of the exponential distribution is $\frac{1}{\lambda}$, the theoritical variance of our sampling distribution is the squared standard error given by the formula $\frac{\sigma}{\sqrt{n}}$. The below code computes the theoritical and the observed variance for our distribution. We see indeed that both values are very close (0.625 vs 0.635):

```
var.theoretical <- (1/lambda)^2 / n
var.theoretical
```

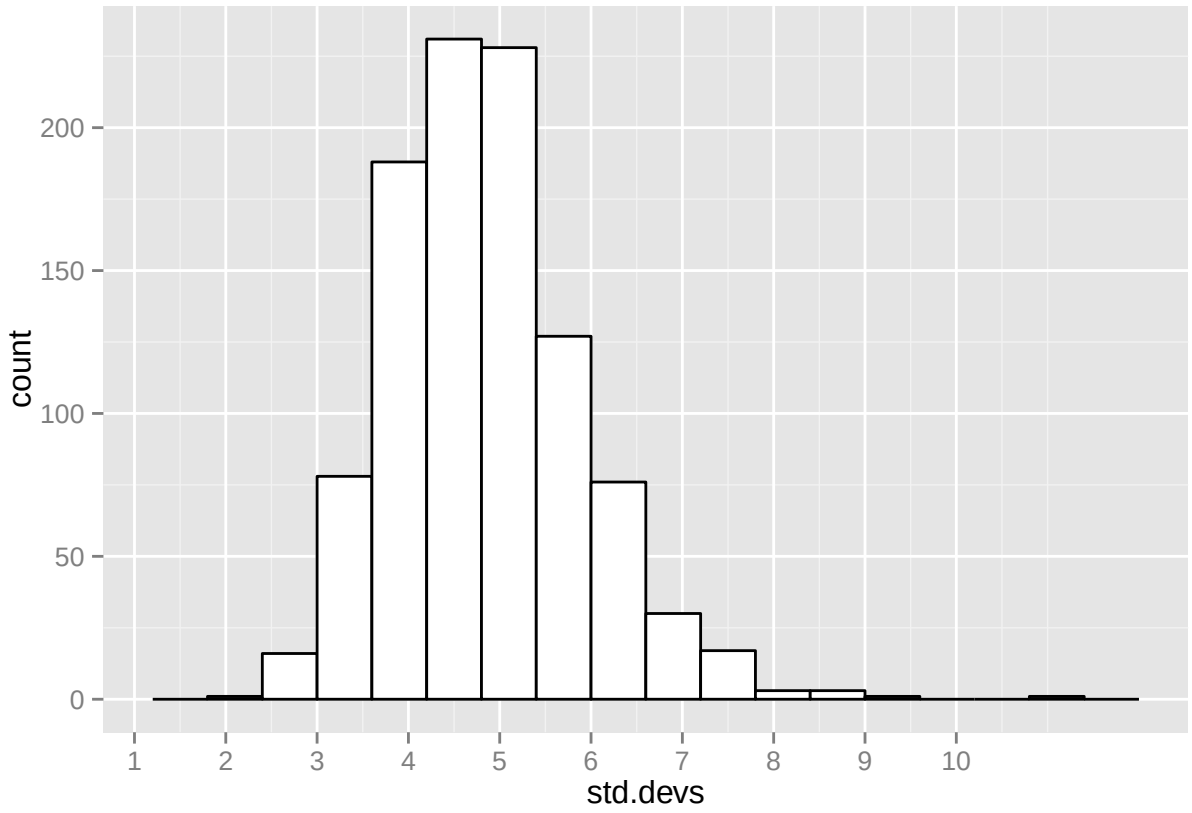
```
## [1] 0.625
```

```
var.observed <- (sd(means))^2
var.observed
```

```
## [1] 0.6351
```

Furthermore, if we look at the distribution of standard deviations, we can see that the standard deviations are slightly right-skewed with a center at about the theoritical standard deviation 1/lambda, i.e. 5:

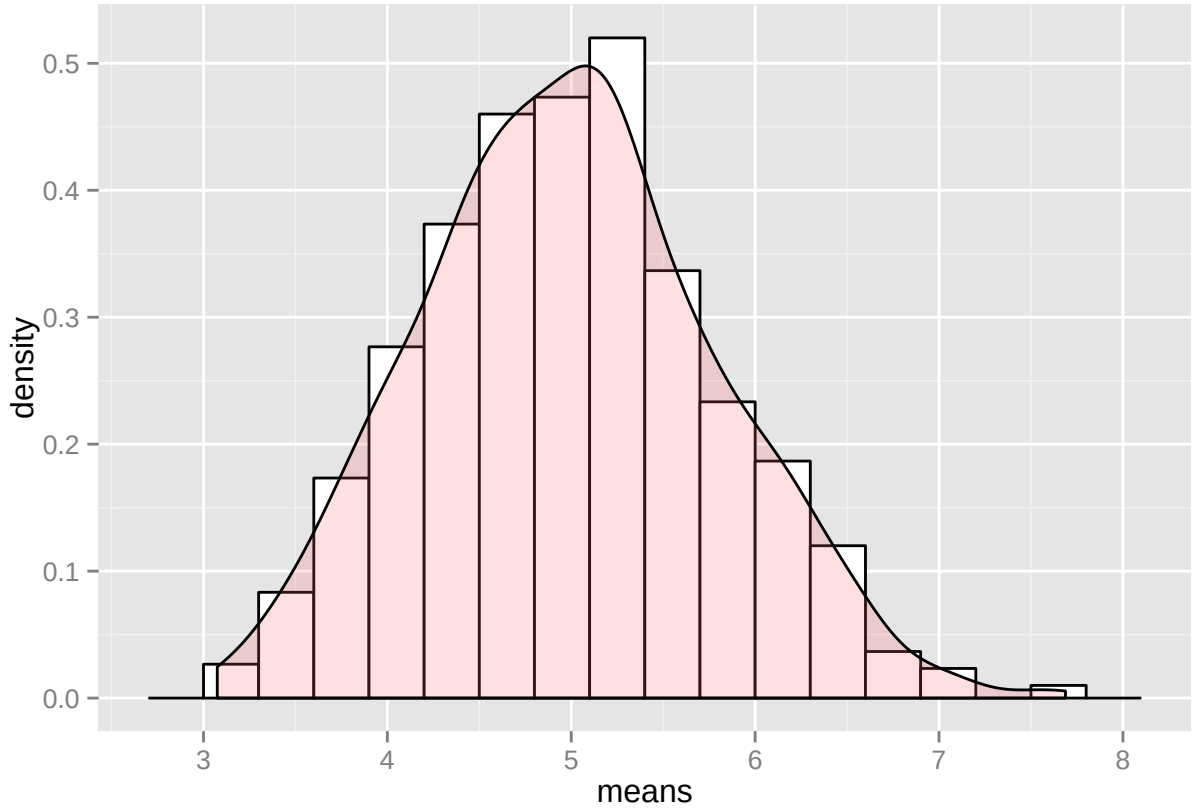
```
# draw distributions of standard deviations
std.devs <- apply(data, 1, sd)
df1 <- cbind.data.frame(simul.id=(1:nosim), std.devs)
# histogram overlaid with kernel density curve
ggplot(df1, aes(x=std.devs)) +
    geom_histogram(binwidth=0.6, colour="black", fill="white") +
    scale_x_continuous(breaks=(0:10))
```



3. Show that the distribution is approximately normal.

Drawing the histogram along with the density function for our sampling distritbion of means, we can see that the distribution is nearly normal and centered at 5, which is the theoritical population mean:

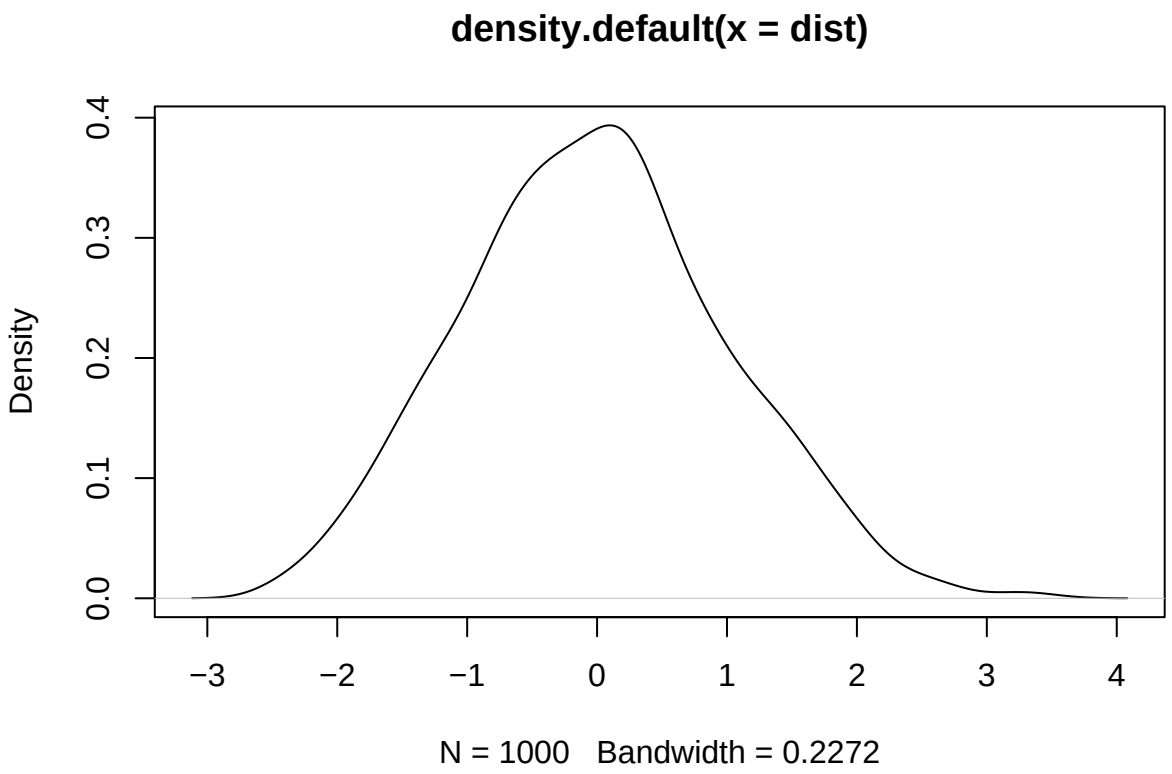
```
# store data in a data frame to use with ggplot
df2 <- cbind.data.frame(simul.id=(1:nosim), means)
# histogram overlaid with kernel density curve
ggplot(df2, aes(x=means)) +
    geom_histogram(aes(y=..density..), binwidth=0.3, colour="black", fill="white") +
    geom_density(alpha=.2, fill="#FF6666") +    # overlay with transparent density plot
    scale_x_continuous(breaks=(0:10))
```



Similarly, we can plot the density function of the differences between our simulated means and the theoritical mean.

```
# compute standard
std.err <- (1/lambda) / sqrt(n)

# distribution of differences in  means vs theoritical mean
dist <- (means - (1/lambda)) / std.err
d <- density(dist)
plot(d)
```



This plot shows that the distribution of differences in means is bell-shaped and centered at zero.

4. Evaluate the coverage of the confidence interval for 1/lambda.

We previously computed our observed mean (xbar) and our standard error (std.err). Based on these two values, we can compute a 95% confidence interval as follows:

```
# compute confidence interval
conf.int <- xbar + c(-1,1) * qnorm(0.975) * std.err
conf.int
```

```
## [1] 3.437 6.536
```

We see that the interval indeed contains our theoritical mean 5.