Coursera Statistical Inference Project - Part 1

Father Abraham

13/09/2014

Part 1

The first task of the project is to "[i]llustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s".

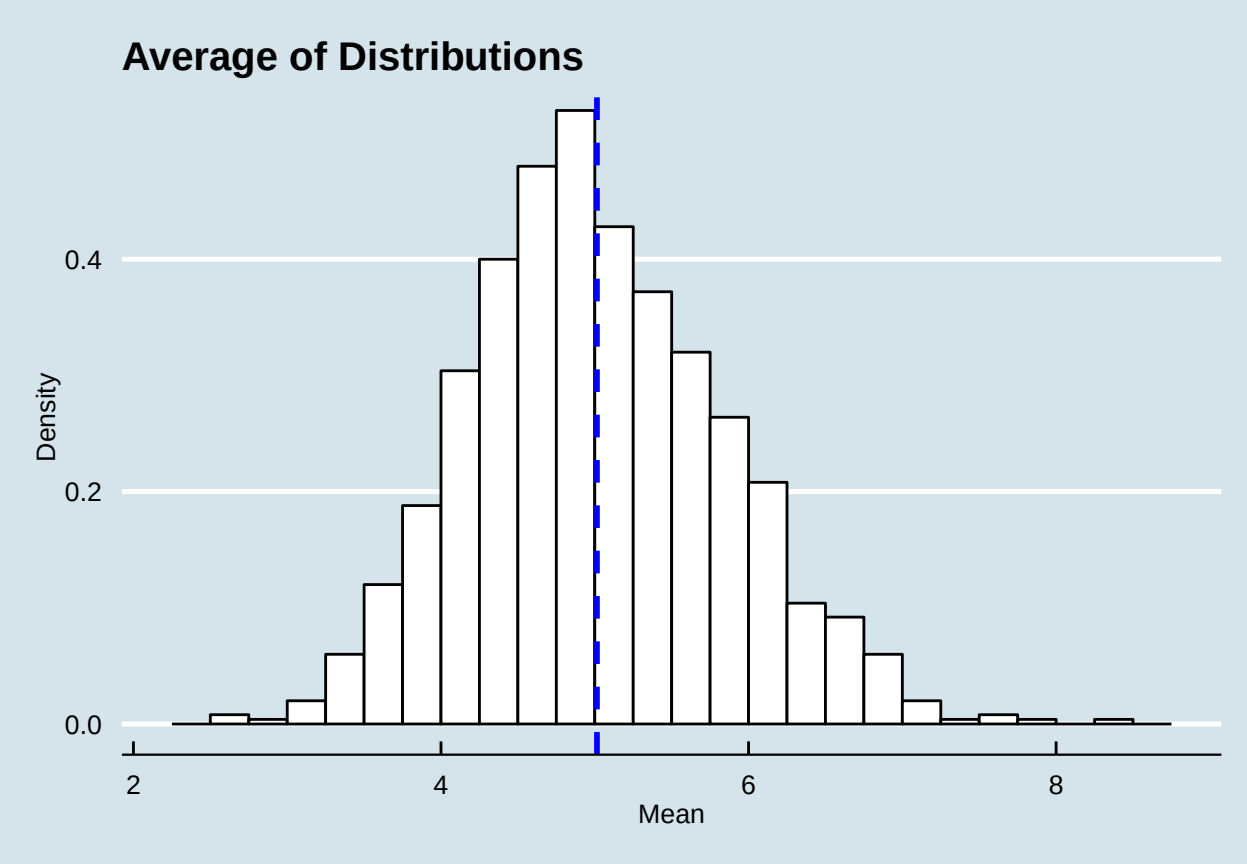
```
##set seed for predictable random numbers (!)
set.seed(23)

##satisfy project requirements, i.e. run 1000 simulations of sample size n=40
sims=1000
lambda=0.2
obs=40

##load and format data
dat1000 <- as.data.frame(replicate(sims, mean(rexp(obs, lambda))))
names(dat1000) <- c("Means")
```

This has produced the sampling distribution of the sample mean.

With the data in place we will now plot the distribution as a histogram using the following code, and we add a line to show where the distribution is centered - very close to 5, and where we would expect it, as the Central Limit Theorem tells us for a large enough number, the sample mean will converge on the population mean (which we know to be 5):



Let's have a look at some other properties of the sampling distribution:

The variance:

```
var(dat1000$Means)
```

```
## [1] 0.6862
```

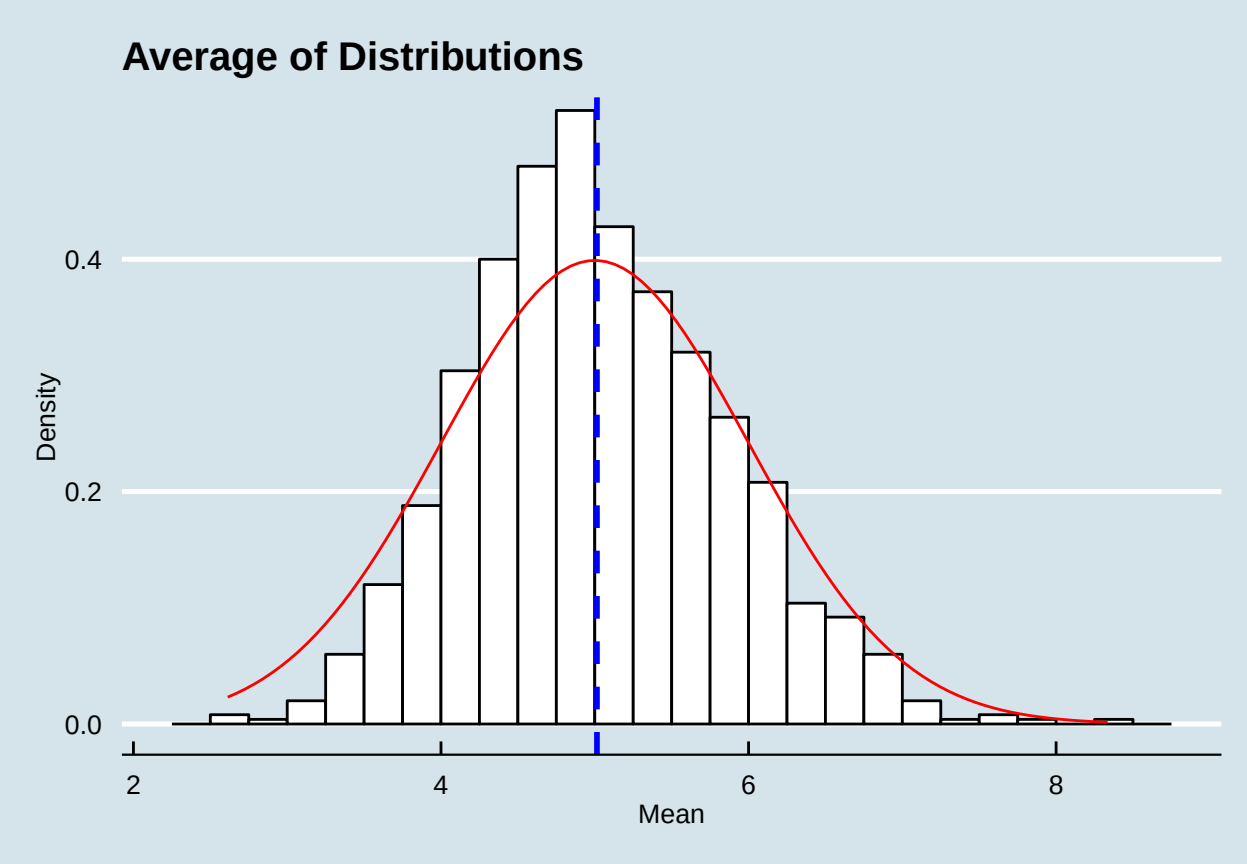
The theoretical variance of the sampling distribution is given by the square of the standard deviation of the population divided by the sample size:

```
((1/lambda)^2)/obs
```

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

Next, we plot the normal distribution on top, we know the mean is 5 from 1/lambda, so we will use that (i.e. we won't *fully* standardize, by centering the mean on 0), and for a standard normal distribution the standard deviation is 1 (Source: http://en.wikipedia.org/wiki/Normal_distribution)

```
g <- g + stat_function(fun = dnorm, colour = "red", arg = list(mean=5, sd =1))
g
```



Clearly, the distribution is very close to a normal distribution, and we have demonstrated that the Central Limit Theorem holds here.

The confidence level has both and an upper limit which we create from the population parameters we know, namely the mean and the standard deviation:

```
1/lambda + (c(-2, 2)*(1/lambda)/sqrt(40))
```

```
## [1] 3.419 6.581
```

and using sqldf we find the coverage:

```
sqldf("select count(*) from dat1000 where Means between 3.418861 and 6.581139")
```

```
##   count(*)
## 1      941
```

So we have a 94.1% confidence level.