Statistical Inference Course Project: Part 1

by Maxim Podkolzine

Problem: The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter. The mean of exponential distributio `n` 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`.

Let's begin our research by doing 1000 simulations of 40 exponentials and collecting their means.

```
lambda = 0.2
n = 40
set.seed(0)
sample_means = NULL
for (i in 1:1000) {
  sample_means = c(sample_means, mean(rexp(n, lambda)))
}
```

Q1: Now let's find the mean over all simulations and compare it to the theoretical mean.

```
mu_hat = mean(sample_means)
mu_hat
```

```
## [1] 4.989678
```

```
mu_true = 1 / lambda
mu_true
```

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

Q2: Now the same for the variance.

```
sigma2_hat = var(sample_means)
sigma2_hat
```

```
## [1] 0.6181582
```

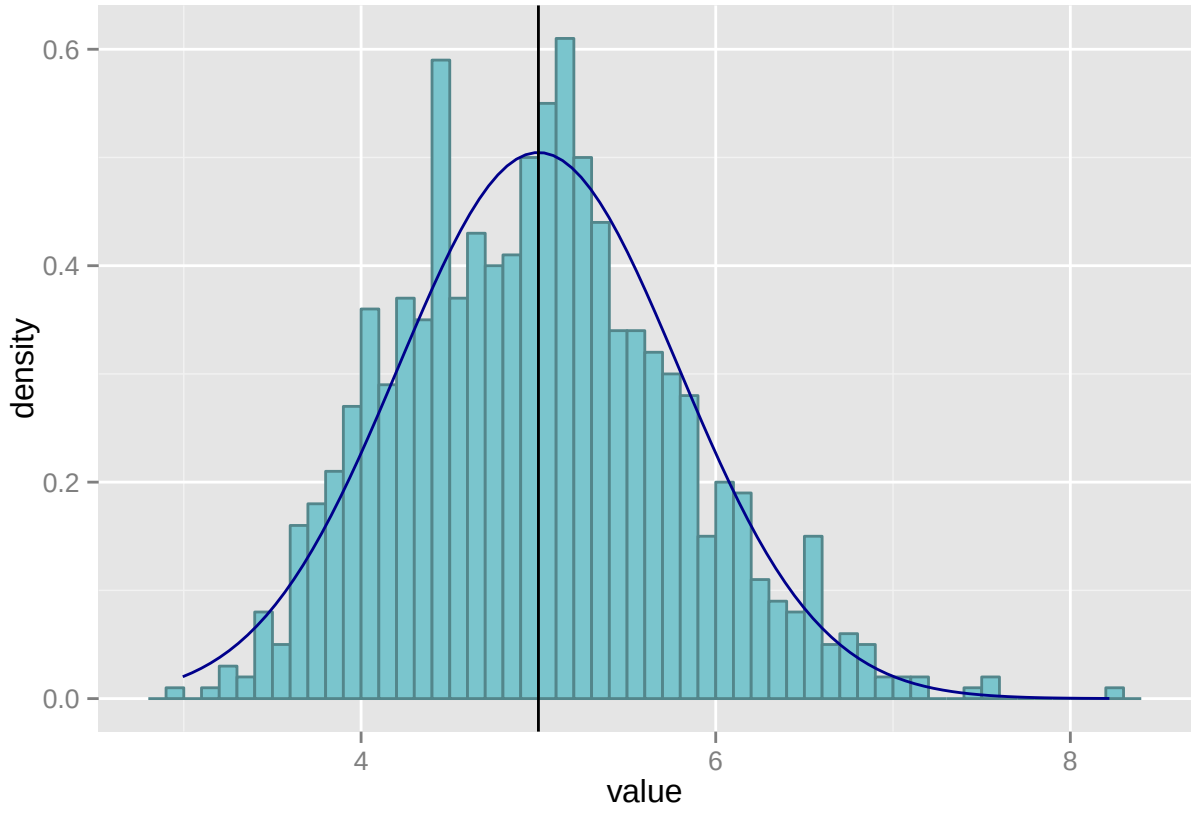
```
sd = 1 / lambda
sigma2_true = sd ^ 2 / n
sigma2_true
```

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

Q3: Finally let's make a couple of plots to compare the observed distribution with the theoretical normal.

The density histogram:

```
ggplot(data=data.frame(value=sample_means), aes(x=value)) +
  geom_histogram(aes(y=..density..), fill=I("cadetblue3"), binwidth=0.1, color=I("cadetblue4")) +
  stat_function(fun=dnorm, arg=list(mean=mu_true, sd=sqrt(sigma2_true)), color=I("darkblue")) +
  geom_vline(xintercept=mu_true, color=I("black"))
```



And the QQ plot:

```
ggplot(data.frame(value=sample_means), aes(sample=value)) +
  geom_point(stat="qq", color=I("darkblue"), alpha=0.3)
```