The Central Limit Theorem in Practice

Robert Sizemore

9/4/2019

Overview

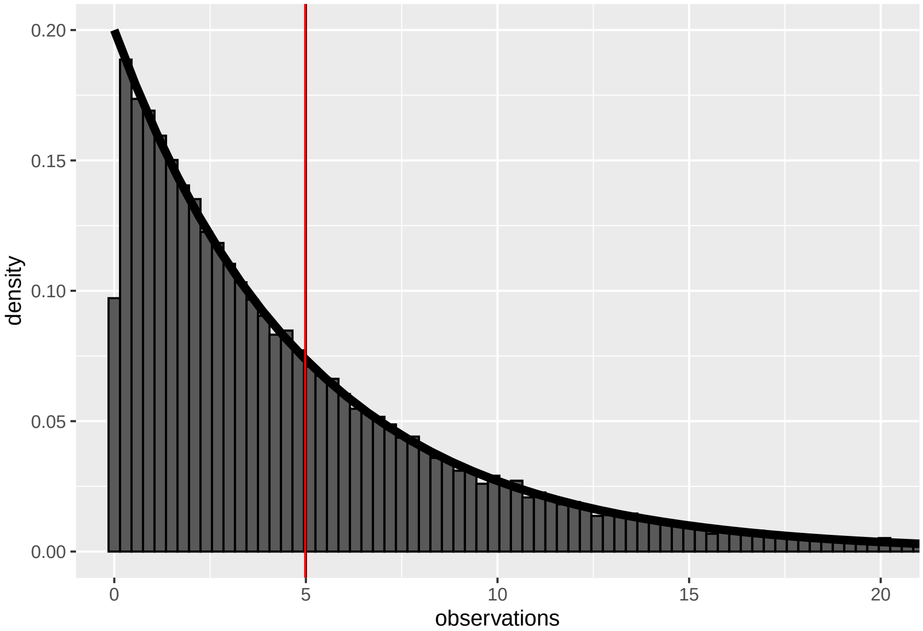
In this document we use are going to explore the results of the Central Limit Theorem. In particular, we are going to use sample statistics from randomly generated exponential data to estimate the true parameters of the exponential distribution. For this analysis, we draw 40,000 observations from a simulated exponential distribution with $\lambda = 0.2$. We group these observations into 1000 samples of 40 exponentials and store it in a 1000x40 matrix. We can then compute the sample mean from the rows of this matrix.

```r
library(datasets)
library(ggplot2)
set.seed(153)
obs <- rexp(40000,rate=0.2)
smpls <- matrix(data=obs,nrow=1000,ncol=40)
mns <- apply(smpls,1,mean)
```

Sample Mean vs Theoretical Mean

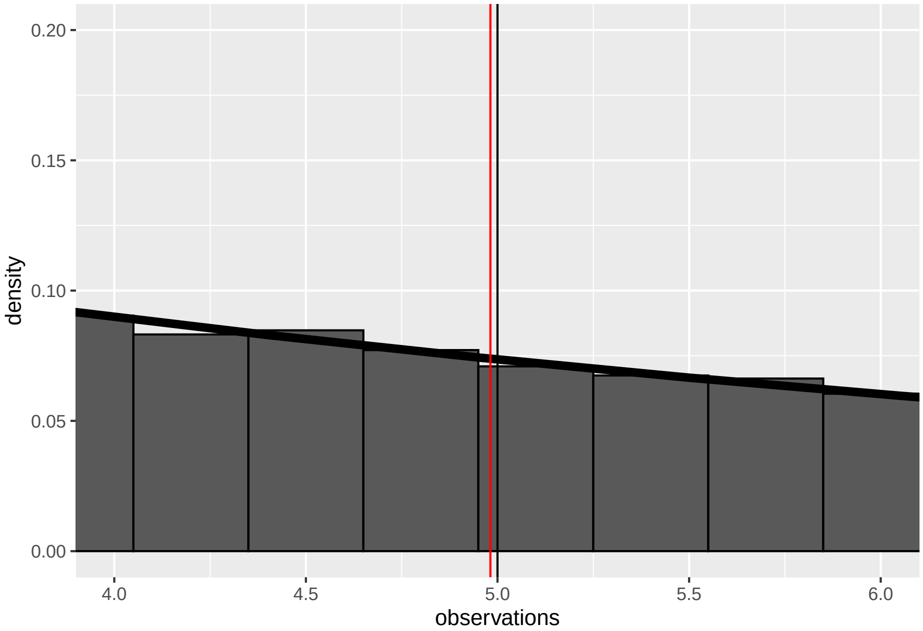
From the Central Limit Theorem, we know that the distribution of the sample mean $\bar{X}_n$ converges to $N(\mu, \sigma^2/n)$ as we take large numbers of samples. Here, μ and σ are the population mean and standard deviation and n is the sample size, in this case $n = 40$ and $\mu = \sigma = 1/\lambda = 5$. From this, we should that expect that our sample distribution should be approximately normal. In particular, the mean of the sample distribution should be approximately the same as the population mean μ. We take a look at the distribution of the observed data, together with vertical lines denoting the expected and sample means:

```r
data1 <- data.frame(obs = obs)
g1 <- ggplot(data1, aes(x = obs))
g1 <- g1 + geom_histogram(binwidth=.3, colour = "black", aes(y = ..density..))
g1 <- g1 + stat_function(fun = dexp, size = 2, args = list(rate=0.2))
g1 <- g1 + geom_vline(xintercept = 5, color  = "black")
g1 <- g1 + geom_vline(xintercept = mean(mns), color = "red")
g1 <- g1 + labs(x = "observations")
g1  + coord_cartesian(xlim=c(0,20))
```



We see that the average of our sample means $\bar{X}$ is very close the theoretical mean $\mu = 5$, as expected. Their line markers overlap and are essentially indisguishable, so we zoom in around the mean to see the separation:

```
g1 + coord_cartesian(xlim=c(4,6))
```



Sample Variance vs Theoretical Variance

Recall from the CLT that our sample distribution is approximately $N(\mu, \sigma^2/n)$. So, we should also be able to estimate the population standard deviation from the calculated sample standard deviation. We expect a sample variance of

$$S^2 \approx \sigma^2/n \implies \sigma \approx \sqrt{40S^2}.$$

In this case, there isn't a convenient graphical way to compare the sample and theoretical variances, so we simply compute the value using the var function in R:

```
sqrt(40*var(mns))
```

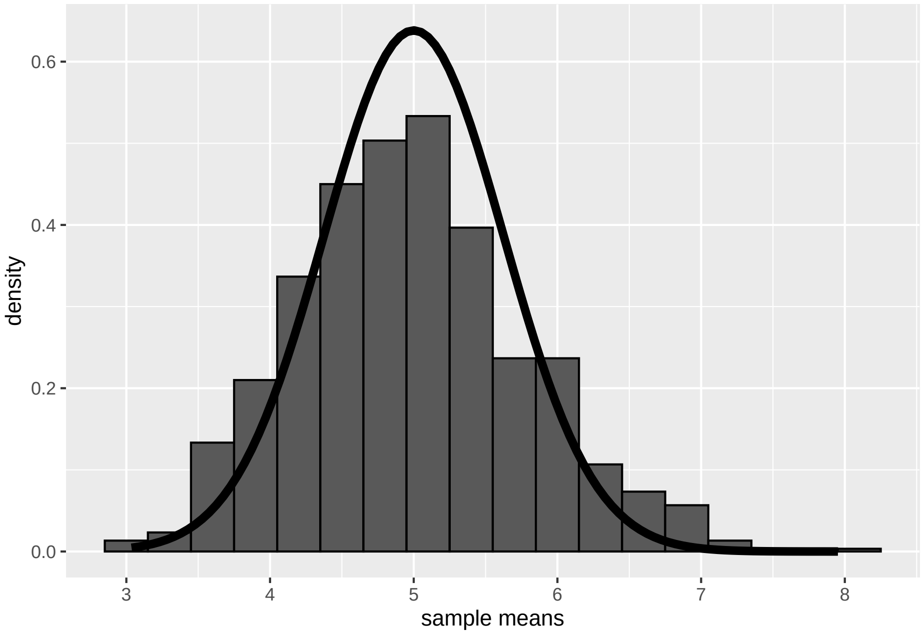
```
## [1] 4.995671
```

This is very close to the known population standard deviation $\sigma = 5$.

The Sample Distribution

Finally, we wish to take a closer look at the sample distribution formed by our simulated data. We expect our data to approximate the normal distribution with mean μ and variance σ^2/n, where $\mu = \sigma = 5$ and $n = 40$. Below is a histogram of the sample means, with a density curve for the theoretical distribution superimposed.

```
data3 <- data.frame(means = mns)
g3 <- ggplot(data3, aes(x = mns))
g3 <- g3 + geom_histogram(binwidth=.3, colour = "black", aes(y = ..density..))
g3 <- g3 + stat_function(fun = dnorm, size = 2, args = list(mean = 5, sd = 25/40))
g3 + labs(x = "sample means")
```



We see that our data seems to roughly approximate the distribution given by the curve, as the CLT predicts. We expect that as we include more observations in each sample, our sample distribution should converge to a better approximation of a normal distribution.

Conclusion

In this document we tested the predictions of the Central Limit theorem by showing that sample distribution is approximately normal. Furthermore, the calculated mean and variance of the sample distribution can be used to infer the population mean and variance with surprising accuracy.