Statistical Inference Course Project

Overview

This is a Coursera Statistical Inference project which aims to compare the distribution of the average of 40 exponentials with lambda = 0.2 simulated 1000 times against the Central Limit Theorem (CLT). The CLT states that in general, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

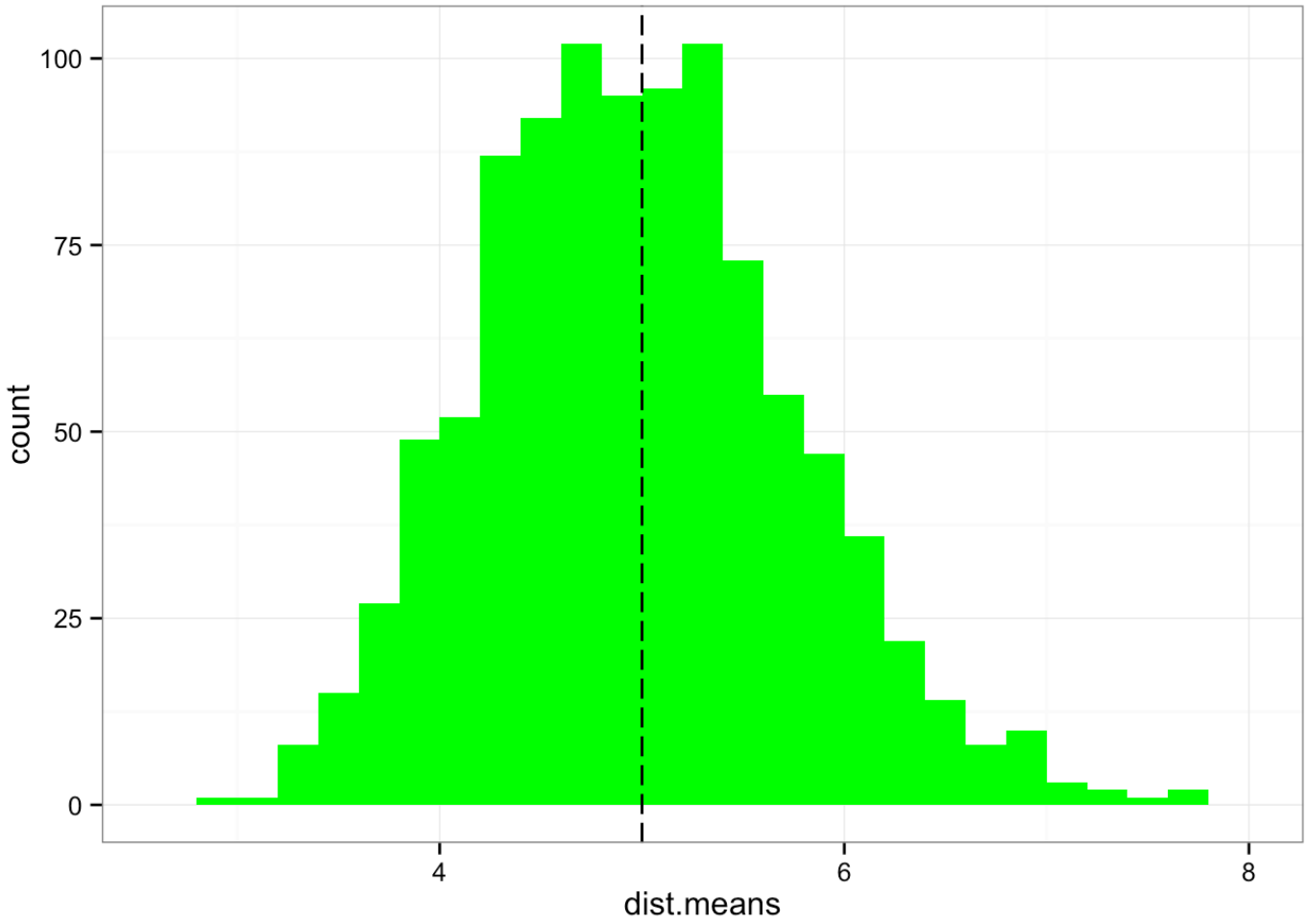
Setting up the Simulations

```
library(ggplot2)

set.seed(1234)
nosim = 1000
sampleSize = 40
lambda = 0.2
dist.means = (apply(matrix(rexp(nosim * sampleSize, lambda), nosim), 1, mean))
```

Sample Means vs. Theoretical Means

```
theo_mean = 1 / lambda
qplot(dist.means, geom="histogram", fill=I("green"), binwidth=0.2) + geom_vline(xi
ntercept=theo_mean, linetype="longdash") + theme_bw()
```



We know that an exponential distribution looks nothing like the normal distribution. And yet, we see from the above chart that after a 1000 simulations of the averages of 40 exponentials, the distribution of averages begins to approach that of a normal distribution. The theoretical mean (vertical dashed line), given by 1 / lamda is 5, falls near the center of our simulated distribution of sample means, which is 4.97.

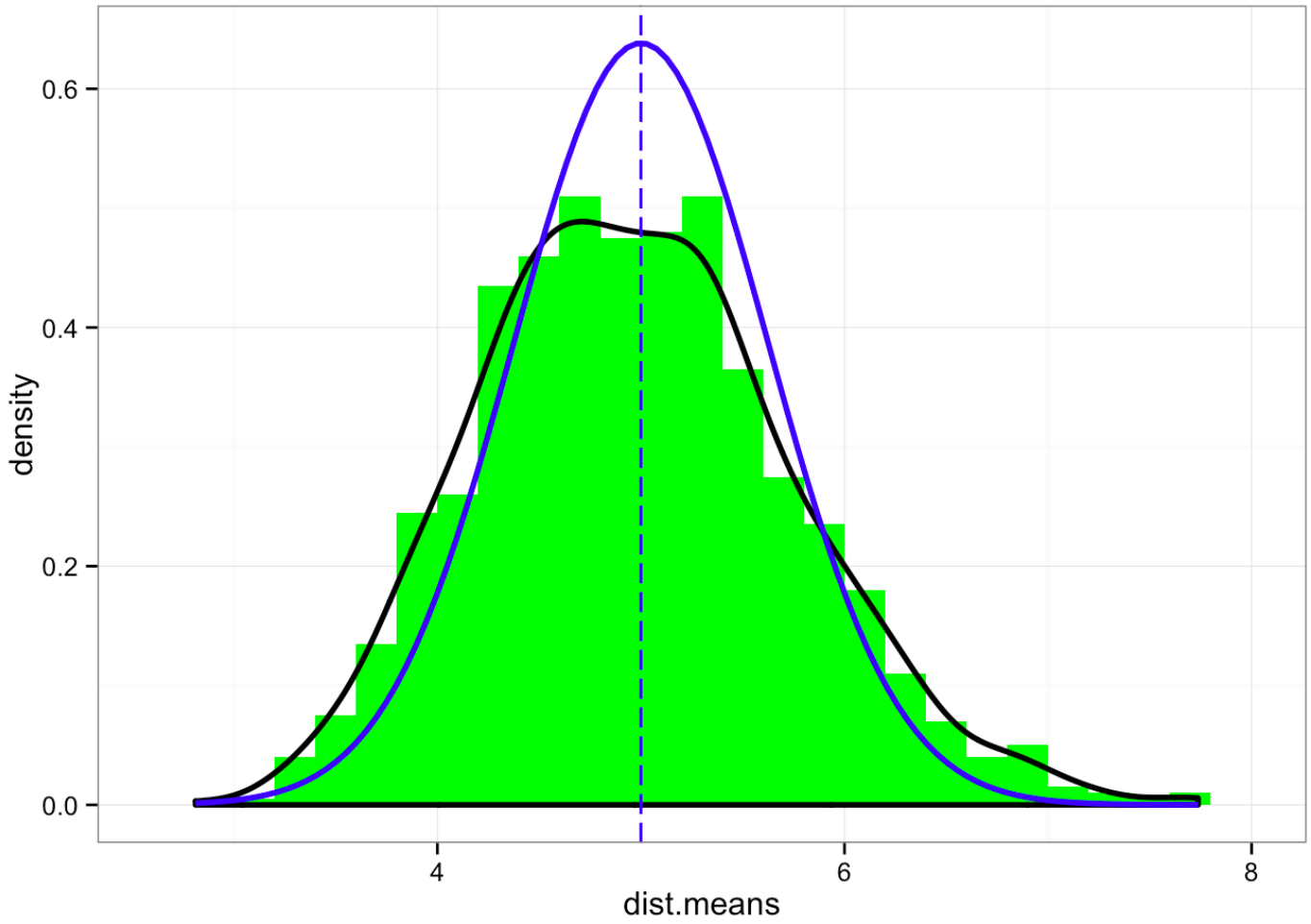
Sample Variance vs. Theoretical Variance

```
theo_var = (1 / lambda^2) / sampleSize
```

The theoretical variance, given by (1 / lamda^2) / n is 0.625, is fairly close to our sample variance, 0.595.

Distribution

```
ggplot(data.frame(dist.means), aes(x=dist.means)) + geom_histogram(aes(y=..densit
y..), fill="green", binwidth=0.2) + geom_density(color="black", size=1) + geom_vli
ne(xintercept=theo_mean, linetype="longdash", color="blue") + theme_bw() + stat_fu
nction(fun = dnorm, args = list(mean = 5, sd = 0.625), color="blue", size=1)
```




```
ci = mean(dist.means) + c(-1,1) * 1.96 * sd(dist.means)/sqrt(sampleSize)
theo_ci = theo_mean + c(-1,1) * 1.96 * sqrt(theo_var)/sampleSize
```

As we can see from the above chart, the fitted distribution of the samples (black line) looks approximately normal (blue line). The 95% confidence interval of the sample distribution is 4.74 to 5.21, which is fairly close to the 95% confidence interval for the theoretical distribution, which is 4.96 to 5.04.