Statistical Inference _ Simulation

wenlarry

Oct 15 2016

OVERVIEW

An investigation on the distribution of the averages of 40 exponentials in R and compare it with the Central Limit Theorem (CLT). This exponential distribution is simulated in R with rexp(n,lambda), where lambda is the rate parameter. Both the mean and standard deviation of the exponential distribution is 1/lambda. For another 1000 simulations, lambda is at 0.2.

The investigation shall also compare the sample mean and variance with their theoretical counterparts as well as verifying that the distributions are approximately normal. Plots and calculations, including a Confidence Interval evaluation are used to support the conclusions.

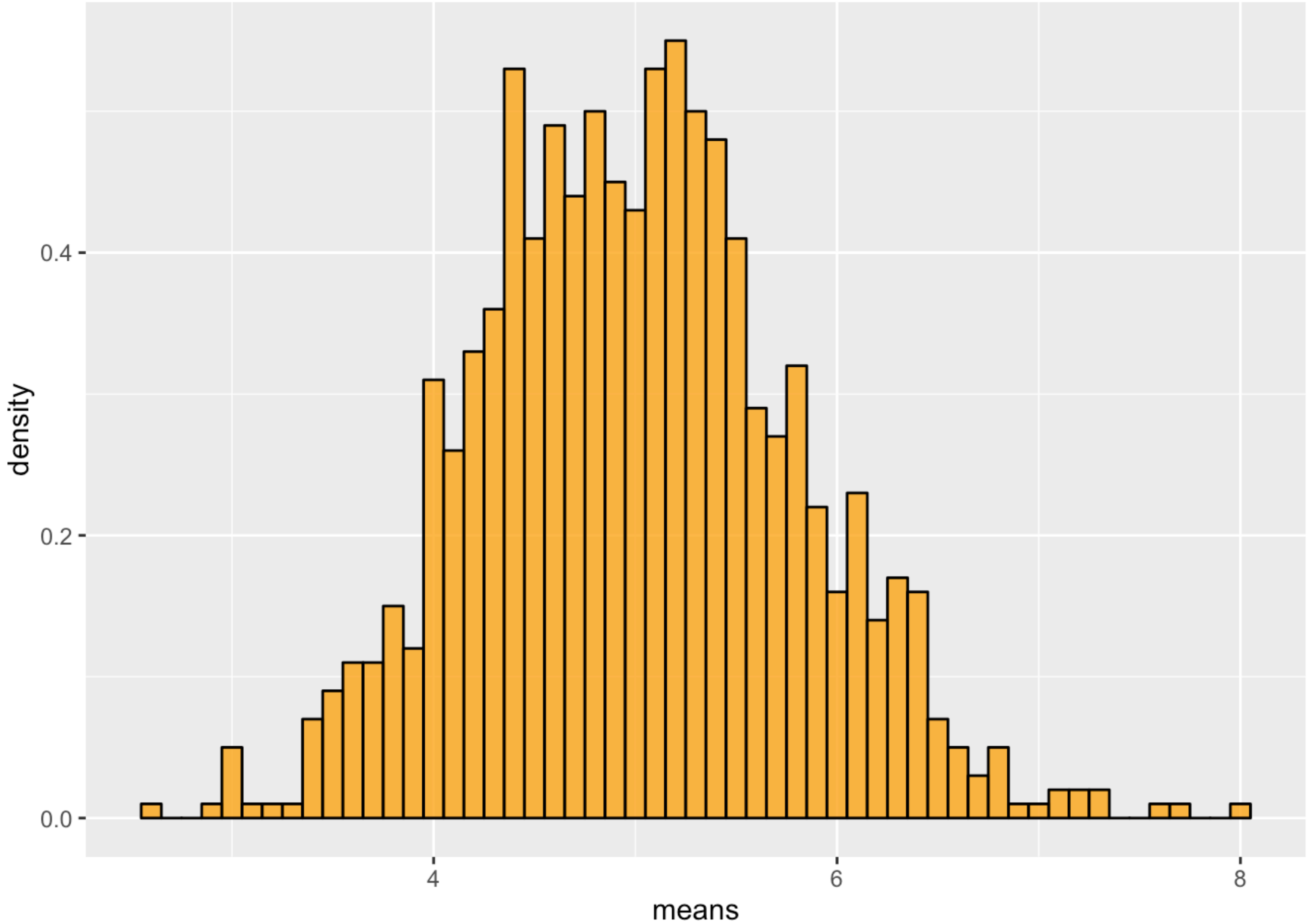
SAMPLE MEAN vs THEORETICAL MEAN

Simulation

libraries: ggplot2,dplyr,UsingR

```
set.seed (9359)
lambda<-0.2
n<-1000
s<-40
expdist<-matrix(rexp(n*s,lambda) ,n,s)
means<-rowMeans(expdist)
meanexpdist<-data.frame(means=apply(expdist,1,mean))
```

```
library(ggplot2)
ggplot(meanexpdist,aes(means))+geom_histogram(aes(y=..density..),fill="orange" ,co
lor="black",binwidth=0.1, alpha=0.8)
```



The plot shows that the simulated exponential distribution tracks the Central Limit Theorem. However, this result is based on lambda at 0.2 with 1000 simulations.

Calculation

CLT Test = Estimate - Mean of Estimate /Std Err of estimate

CLT Test = sqrt*(n)(X_bar-mu)/sd

Theoretical Mean = (1/lambda) Sample Mean = Mean (means)

```
tmean<-1/lambda
tmean
```

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

```
#5
smean<-mean(means)
smean
```

```
## [1] 5.002165
```

```
#4.964221

n<-1000
X_bar<-1/lambda
mu<-smean
sd<-sd(means)

CLT_test<-sqrt(n)*(X_bar-mu)/sd
CLT_test
```

```
## [1] -0.08743876
```

```
# -0.08743876
```

A math proof of the small difference between the theoretical mean and the sample mean as well the close fit to the CLT

SAMPLE VARIANCE vs THEORETICAL VARIANCE

Theoretical Variance = sigma^2/N

Sample Variance = Sigma^2 of the means of the exponential distribution (meanexpdist above)

```
n<-40
tsd<-1/lambda/sqrt(n)
tvar<-tsd^2
tvar
```

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

```
#0.625
ssd<-sd(means)
svar<-ssd^2
svar
```

```
## [1] 0.613057
```

```
#0.613057
```

Again, a small difference between the theoretical variance and the sample variance

Confidence Intervals Evaluation

Testing with a 95% interval for the sample mean (mu)

```
smean+c(-1,1)*qnorm(0.975)*ssd/sqrt(length(means))
```

```
## [1] 4.953636 5.050694
```

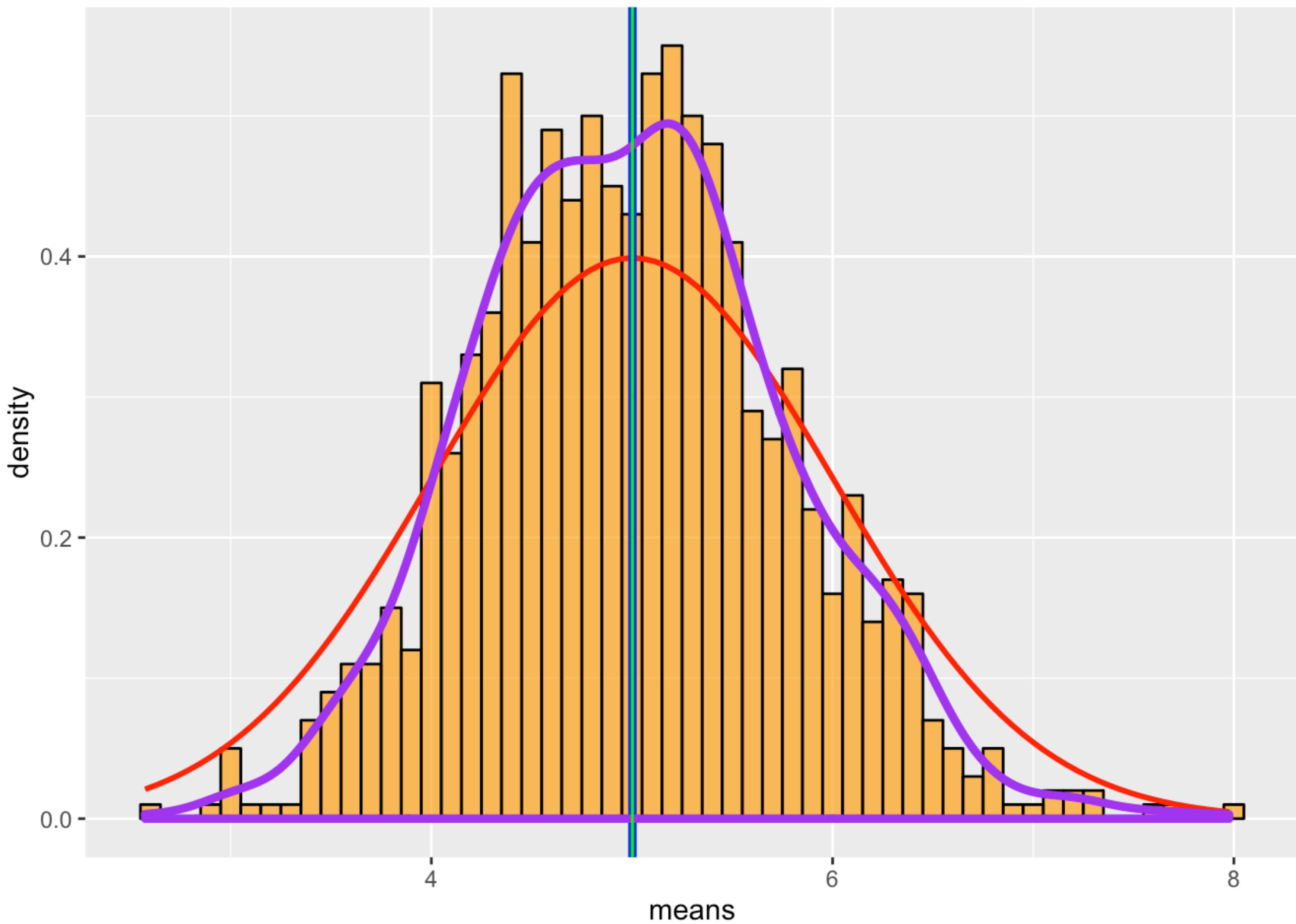
```
# 4.953636  ;  5.050694
```

About 95% of the intervals would contain the sample mean. The true mean is 5

DISTRIBUTION

Superimpose the mean and distribution of the averages of the 40 exponentials onto the first plot. Refine first plot with the mean and distribution of the 1000 simulations.

```
ggplot(meanexpdist , aes(means)) + geom_histogram( aes(y =..density..),fill = "ora
nge",color = "black",
binwidth = 0.1,alpha = 0.7) + stat_function(fun = dnorm,args = list(mean=mu),color
= "red" ,size = 1) +  geom_vline(xintercept = mu,size = 1.5,color = "blue") + geom
_density(color ="purple", size = 1.5) + geom_vline(xintercept = smean, size = 0.5,
color =
"green")
```



The graphic of the 1000 simulations vs the distribution of the averages of 40 exponentials. A pretty overlap.