Statistical Inference Course Project

Aiyu Li

01/16/2021

Synopsis

This is a project for Statistical Inference class. There are two parts in this project.

- Part 1: simulation Exercise: investigate the distribution of averages of 40 exponentials and a thousand simulations in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations.
- Part 2: analyze the ToothGrowth data in the R datasets package; provide a basic summary of the data; Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose and provide the compared conclusions.

R environment and reproducibility

```
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18363)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.6.3  magrittr_1.5    tools_3.6.3     htmltools_0.5.0
##  [5] yaml_2.2.1      stringi_1.4.6   rmarkdown_2.3   knitr_1.30
##  [9] stringr_1.4.0   xfun_0.17       digest_0.6.25   rlang_0.4.7
## [13] evaluate_0.14
```

```r
set.seed(2021)
```

Load required libraries

```r
library(stats)
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite
```

```r
library(dplyr, warn.conflicts = F)
```

Part 1 (P1) Simulation Exercise

P1 Objectives

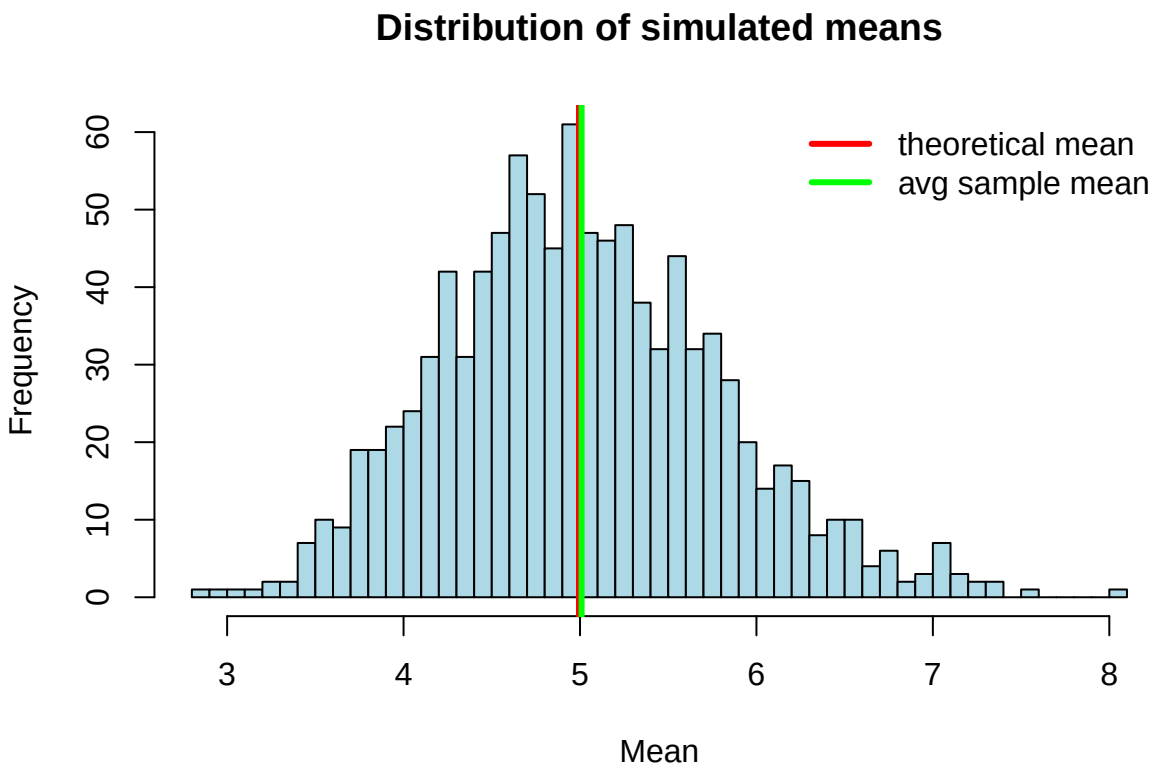
- Show the sample mean and compare it to the theoretical mean of the distribution.
- Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
- Show that the distribution is approximately normal.

P1: Setup variables and create simulation dataset

```r
# investigate the distribution of averages of 40 exponentials
n <- 40
# set lambda = 0.2 for all of the simulations
lambda <- 0.2
# need to do a thousand simulations
nosim <- 1000
# Create simulations dataset
simdata <- matrix(rexp(n*nosim, lambda), nrow = nosim, ncol = n)
```

P1: Compare sample means to theoretical means of the distribution

```r
mns <- apply(simdata,1,mean)
hist(mns,col="light blue",breaks=50, xlab = "Mean",main = "Distribution of simulated means")
abline(v=1/lambda,col="red",lwd=4)
abline(v=round(mean(mns),3),col="green",lwd=3)
legend('topright', c("theoretical mean","avg sample mean"), col=c("red", "green"), lty=c(1,1), lwd=c(3,3
```



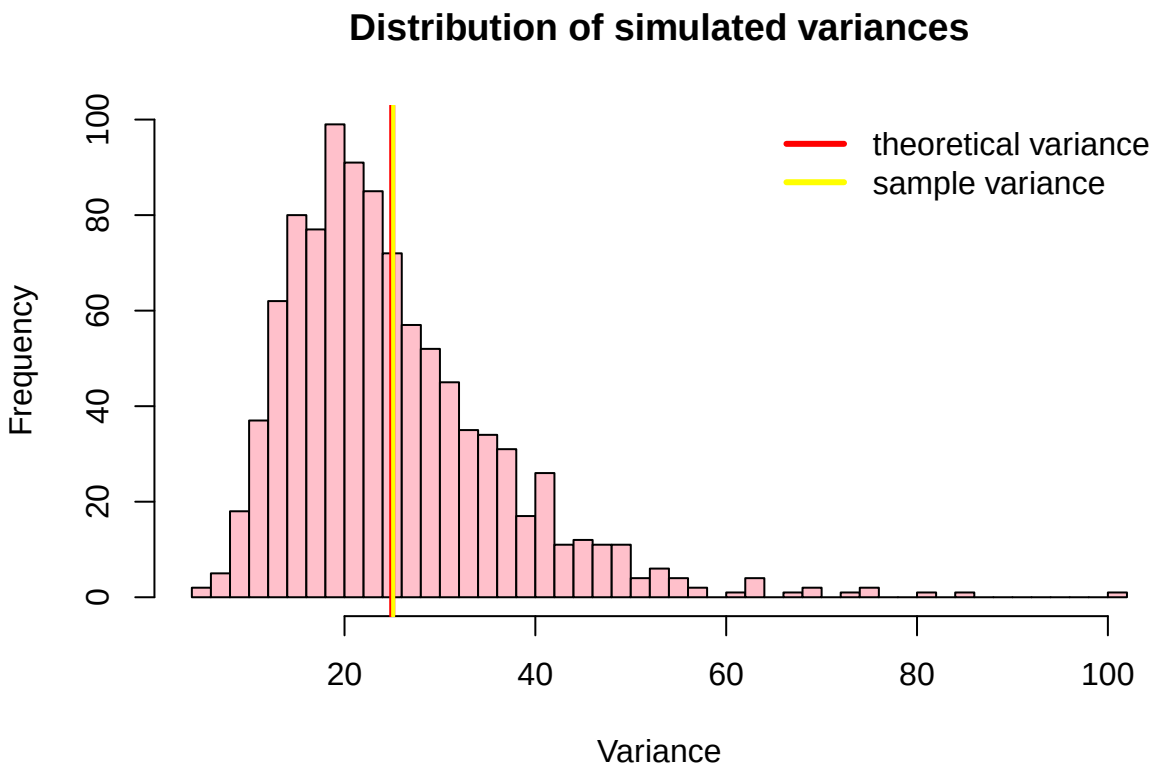

```
comMeans<-paste('sample mean =', round(mean(mns),3), ', theoretical mean=', 1/lambda, sep = "", collaps
comMeans
```

```
## [1] "sample mean =5.009, theoretical mean=5"
```

Above histogram shows: the sample mean is very closed to theoretical mean of distribution when having enough samples of the exponential distribution.

P1: Compare sample variance to the theoretical variance of the distribution

```
vars <- apply(simdata,1,var)
hist(vars,col="pink",breaks=50, xlab = "Variance", main = "Distribution of simulated variances")
abline(v=1/lambda^2,col="red",lwd=3)
abline(v=round(mean(vars),3),col="yellow",lwd=2)
legend('topright', c("theoretical variance","sample variance"), col=c("red", "yellow"), lty=c(1,1), lwd=
```




```
comVars<-paste('sample variance =', round(mean(vars),3), ', theoretical variance=', 1/lambda^2, sep = "
comVars
```

```
## [1] "sample variance =25.096, theoretical variance=25"
```

Above histogram shows: the sample variance is very closed to theoretical variance of distribution when having enough samples of the exponential distribution.

P1: Compare the results to a normal distribution

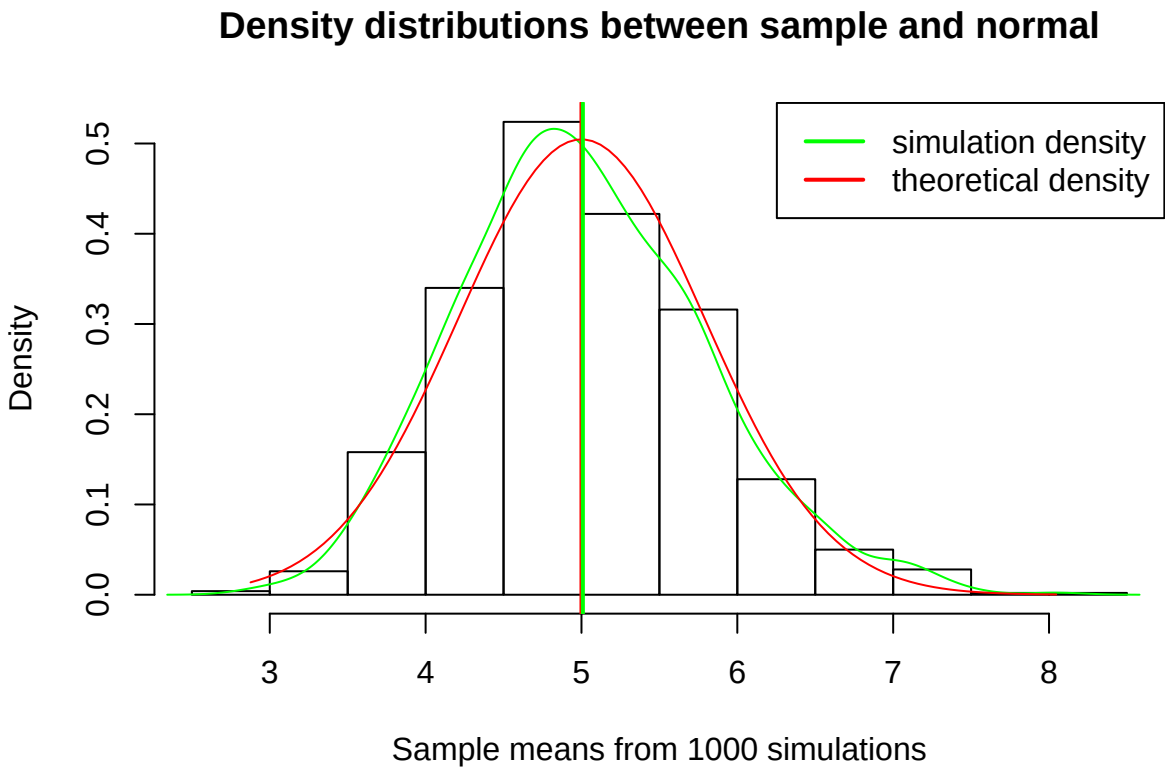
```
## step 1 draw the histgram
hist(mns,  prob=T,
     main="Density distributions between sample and normal",
     xlab="Sample means from 1000 simulations")

# step 2 draw simulated sample mean density distribution
lines(density(mns), col="green", lty=1)


# step 3 draw therotical mean desnity distibution
x <- seq(min(mns), max(mns), length=100)
y <- dnorm(x, mean=1/lambda, sd=(1/lambda/sqrt(n)))
lines(x, y, col="red", lty=1)
```

```r
# step 4 draw both sample mean and theoretical Mean
abline(v=1/lambda, col='red', lwd=2)
abline(v=round(mean(mns),3),col="green",lwd=2)

# step 5 insert legends
legend('topright', c("simulation density", "theoretical density"),
       lty=c(1,1), col=c("green", "red"), lwd=c(2,2))
```



Above histogram with overlay of sample/normal curves shows: the two distribution curves are very similar and normally distributed.

P1 Conclusion

- The sample mean is very close to the theoretical mean of the distribution.
- The sample variance is very close to the theoretical variance of the distribution.
- The sample distribution is approximately normal.