DATA 252 / DATA 551: Homework 4

- You are expected to complete this homework before class on Feb 17, 2020. Contents on this homework (including the assigned video and activity in the last problem) might appear on the quiz. Solutions of the homework will be posted on Moodle.

1. In *Introduction to Probability*, do questions 26 and 28 on page 222.

SOLUTION:

Q26. Rejection rate is $P(X \notin 1 \pm 0.003)$, which is given by `pnorm(1-0.003,mean=1,sd=0.002) + 1-pnorm(1+0.003, mean=1, sd=0.002)` $= 0.1336144$. Trying different values for σ, we see that a value of σ that is slightly smaller than 0.0012 will ensure a rejection rate of less than 1 percent.

```
mysd=seq(from=0.0001, to=0.002, by=0.0001)
cbind(mysd,
      round(pnorm(1-0.003,mean=1,sd=mysd) + 1-pnorm(1+0.003, mean=1, sd=mysd),3))
```

```
##         mysd
##  [1,] 0.0001 0.000
##  [2,] 0.0002 0.000
##  [3,] 0.0003 0.000
##  [4,] 0.0004 0.000
##  [5,] 0.0005 0.000
##  [6,] 0.0006 0.000
##  [7,] 0.0007 0.000
##  [8,] 0.0008 0.000
##  [9,] 0.0009 0.001
## [10,] 0.0010 0.003
## [11,] 0.0011 0.006
## [12,] 0.0012 0.012
## [13,] 0.0013 0.021
## [14,] 0.0014 0.032
## [15,] 0.0015 0.046
## [16,] 0.0016 0.061
## [17,] 0.0017 0.078
## [18,] 0.0018 0.096
## [19,] 0.0019 0.114
## [20,] 0.0020 0.134
```

Q28. We want $P(X > 290) + P(X < 240)$, which is given by `pnorm(240,mean=270,sd=10) + 1-pnorm(290, mean=270, sd=10)`, which gives a probability of only 0.0241 that the defendant was in the country when the child was conceived.

2. Suppose the weight of adult, female cats in a certain cat population approximately follows

a normal distribution with mean = 8 lbs and sd (standard deviation) = 2.2 lbs. What proportion of female cats weighs less than 10 lbs? More than 6 lbs? Between 7 and 9 lbs?

SOLUTION:

```
pnorm(10,mean=8,sd=2.2)  #less than 10
```

```
## [1] 0.8183489
```

```
1-pnorm(6,mean=8,sd=2.2) #more than 6
```

```
## [1] 0.8183489
```

```
pnorm(9,mean=8,sd=2.2) - pnorm(7,mean=8,sd=2.2) #between 7 and 9
```

```
## [1] 0.3505637
```

3. In *Introduction to Probability*, do question 29 on page 222. The **memoryless property** of the exponential distribution implies that the probability that the waiting time exceeds 8 hours if it already exceeds 4 hours is the same as the probability that the waiting time exceeds 4 hours. Calculate this probability first and then write a simulation to verify.

SOLUTION: Both questions have the same probability, $P(X > 4)$, calculated as:

```
1-pexp(4,rate=1/2)
```

```
## [1] 0.1353353
```

4. Suppose the proportion of all stocks that will go up in a given day (i.e., the closing price is higher than the opening price) can be modeled by a beta distribution with parameters shape1 = 2 and shape2 = 3.

a. Simulate 5000 values from the given beta distribution. Plot a histogram and note its shape and range. Use the mean of the simulated values to estimate the true mean of the beta distribution (which is $2/(2+3) = 0.4$). In the context of this problem, what does each simulated value represent?

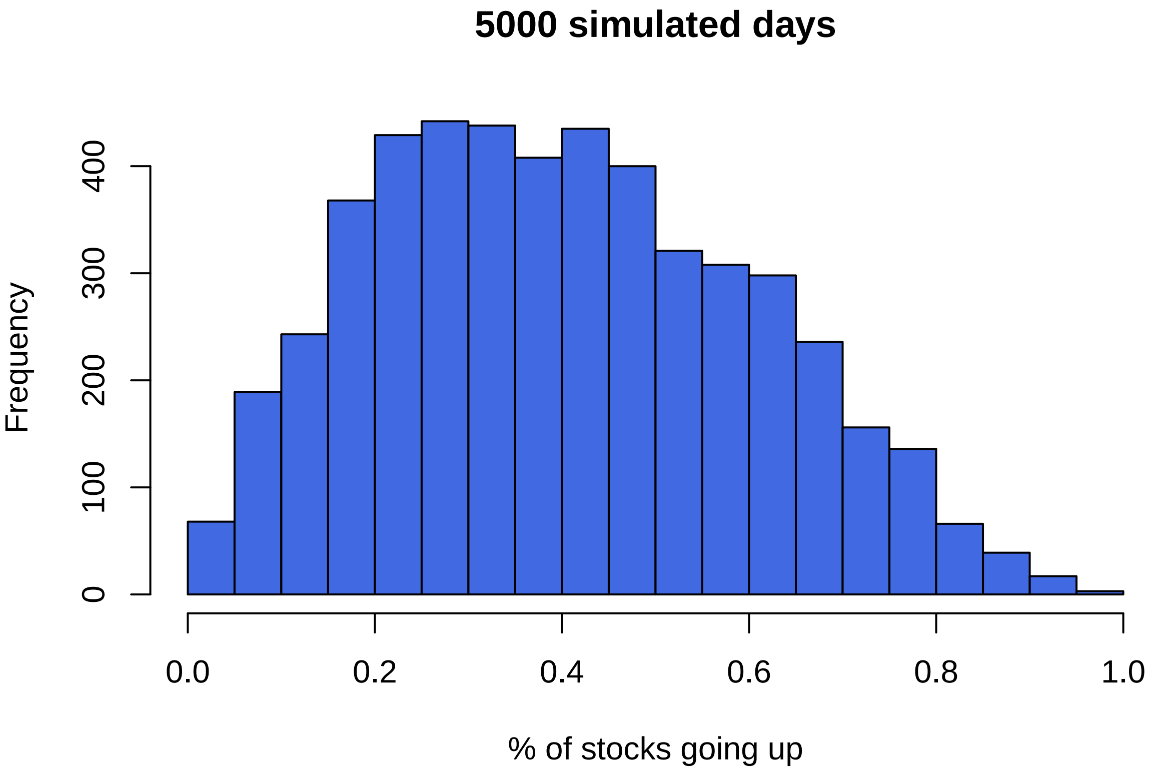
b. What is the chance that we have a really bad day, when less than 5% of all stocks will go up?

c. What is the chance that we have a really good day, when more than 75% of all stocks will go up?

d. On any given day, what is the chance that 30% to 40% of all stocks will go up?

SOLUTION: In the simulation, each number corresponds to a simulated day and represents the percent of stocks that go up in that day. The shape of this beta distribution is right skewed.

```
mysim=rbeta(5000,shape1=2,shape2=3)
hist(mysim,nclass=20,main='5000 simulated days',
    xlab='% of stocks going up',col='royalblue')
```



The probabilities in b to d are calculated below.

```
pbeta(0.05,shape1=2,shape2=3)
```

```
## [1] 0.01401875
```

```
1-pbeta(0.75,shape1=2,shape2=3)
```

```
## [1] 0.05078125
```

```
pbeta(0.4,shape1=2,shape2=3) - pbeta(0.3,shape1=2,shape2=3)
```

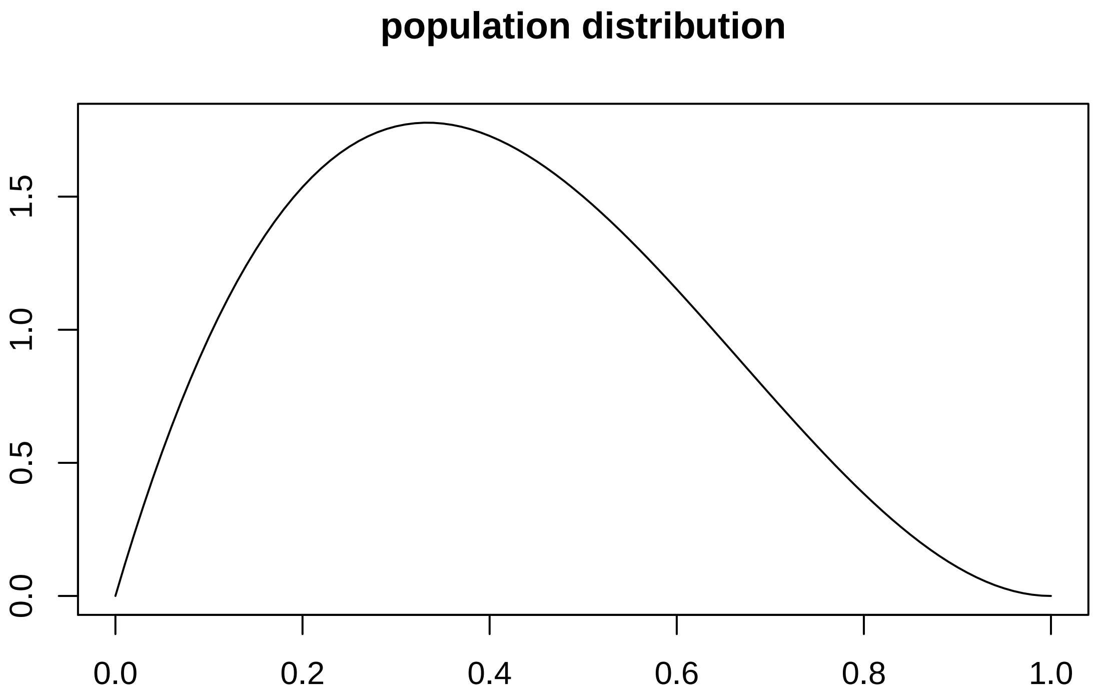
```
## [1] 0.1765
```

5. Watch this video https://www.youtube.com/watch?v=JNm3M9cqWyc&t=352s that explains the Central Limit Theorem. Next, go to http://onlinestatbook.com/stat_sim/sampling_dist/. Play with the simulation until you have a good understanding of the Central Limit Theorem. Follow the instructions on the last two slides of Monday's lecture to design an algorithm that studies the effect of the sample size n.

SOLUTION: As an example, we use a sample size of $n = 20$ and a population distribution of beta distribution with shape parameters 2 and 3. Code below shows that CLT works very well in this case. You should try to change different values of n and different population shapes to study the effect of sample size. For instance, does $n = 10$ still work? $n = 5$? What if the population distribution has a long tail or is very skewed? You can try a t distribution with parameter df=2, and try a gamma distribution with parameters shape=0.01 and rate=2.

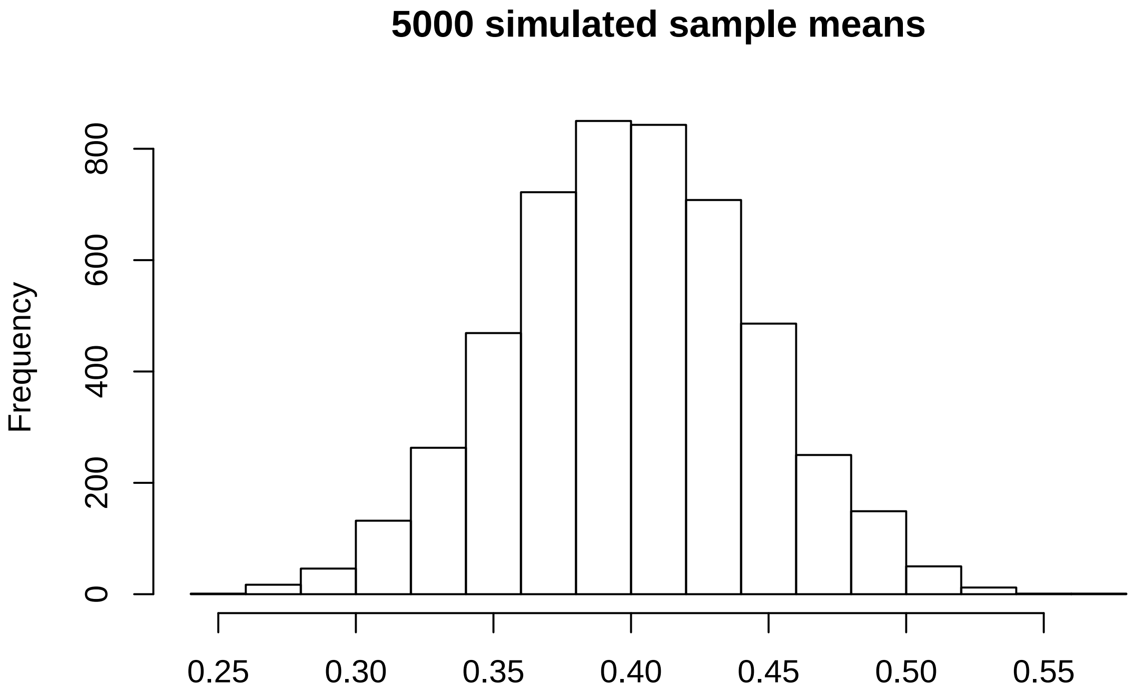
```r
nsim=5000
result=rep(NA,nsim)

n=20  #sample size
curve(dbeta(x,shape1=2,shape2=3),
      main='population distribution',
      xlab=NA,
      ylab=NA) #population distribution
```



```r
for(i in 1:nsim){
  result[i]=mean(rbeta(n=20,shape1=2,shape2=3))
}

hist(result,
     main=paste(nsim,'simulated sample means'),
     xlab=NA)
```


5000 simulated sample means
Frequency
0
200
400
600
800
0.25
0.30
0.35
0.40
0.45
0.50
0.55