# w03

Yekta Amirkhalili

May 23nd, 2023

**Week 3 (Session 2) - May 23, 2023 (makeup class)**

**material to cover:**

---

**0. . . . From previous week: Bayes Rule + Histogram. . .**

~~**Bayes Rule**~

*Introduction*

1. Bayes First Rule:

$$P(E) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

Ex1. Suppose we have two baskets. Basket A has 3 Red and 2 White balls in it. Basket B has 3 Red and 4 White balls in it. We randomly pick one ball from Basket A and transfer it to Basket B. Then we pick a ball, randomly, out of Basket B. What is the probability of the chosen ball being white?

```
bas_A = c('r','r','r', 'w','w')
bas_B = c('r','r','r', 'w','w','w','w')

pick_white_prob <- function(basket){
    #probability of picking white
    total <- length(basket)
    basket_counts <- table(basket)
    whites <- basket_counts[names(basket_counts) == 'w']

    prob <- whites/total
    return(prob)

}

pick_red_prob <- function(basket){
    #probability of picking white
    total <- length(basket)
    basket_counts <- table(basket)
```

```
    reds <- basket_counts[names(basket_counts) == 'r']

    prob <- reds/total
    return(prob)

}
```

- Scenario 1 We pick 1 white ball from Basket A, transfer it to Basket B. Therefore, we now have 3 Red and 5 Whites in Basket B.
- Scenario 2 We pick 1 red ball from Basket A, transfer it to Basket B. Therefore, we now have 4 Red and 4 Whites in Basket B.

In any case, the final probability calculation depends on what happened in the first pick. Let's see this in action!

```
scenario <- function(num){
    #scenario 1
    if(num == 1){
        new_b <- c('r','r','r','w','w','w','w', 'w')
    }else{ #scenario 2
        new_b <- c('r','r','r','r','w','w','w','w')
    }

    return(new_b)

}
```

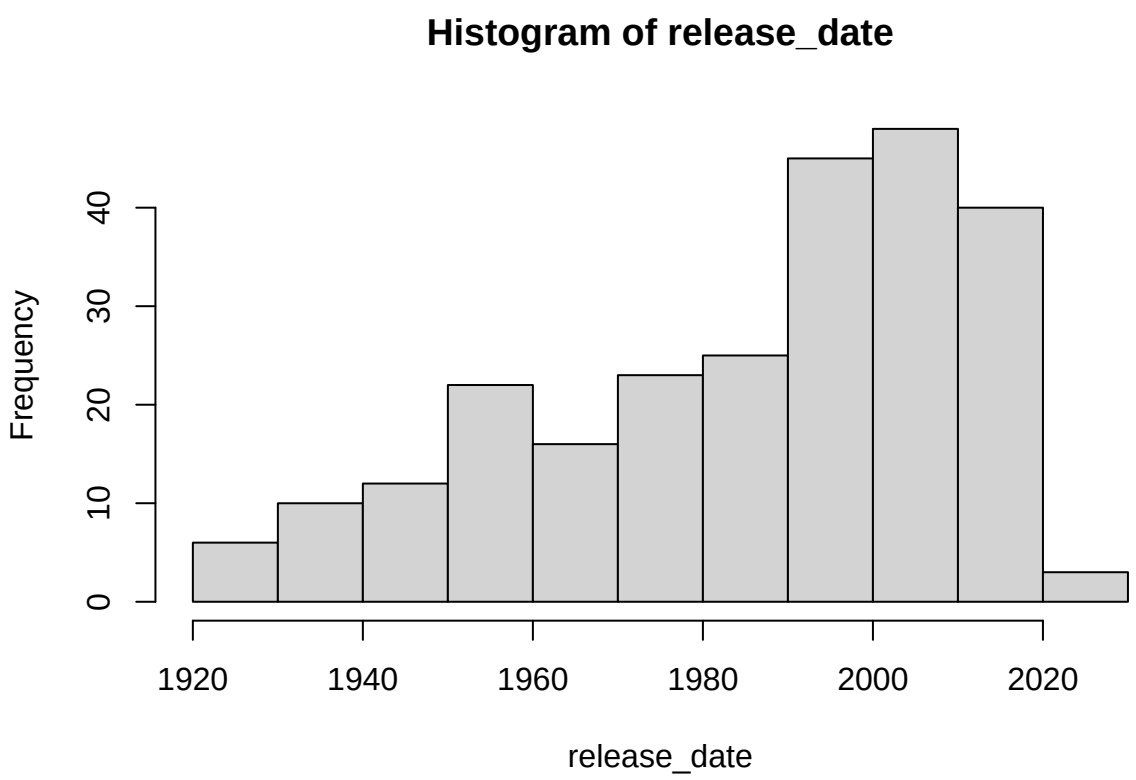
```
newBasket1 <- scenario(1)
newBasket2 <- scenario(2)


prob_ <- function(a, b1, b2){

    # pick white from A, pick white from B of Scenario 1
    total_prob_p1 <- pick_white_prob(a) * pick_white_prob(b1)

    #pick red from A, pick white from new B
    total_prob_p2 <- pick_red_prob(a) * pick_white_prob(b2)

    probability <- total_prob_p1 + total_prob_p2

    return(probability)

}

prob_(bas_A, newBasket1, newBasket2)
```

```
##    w
## 0.55
```

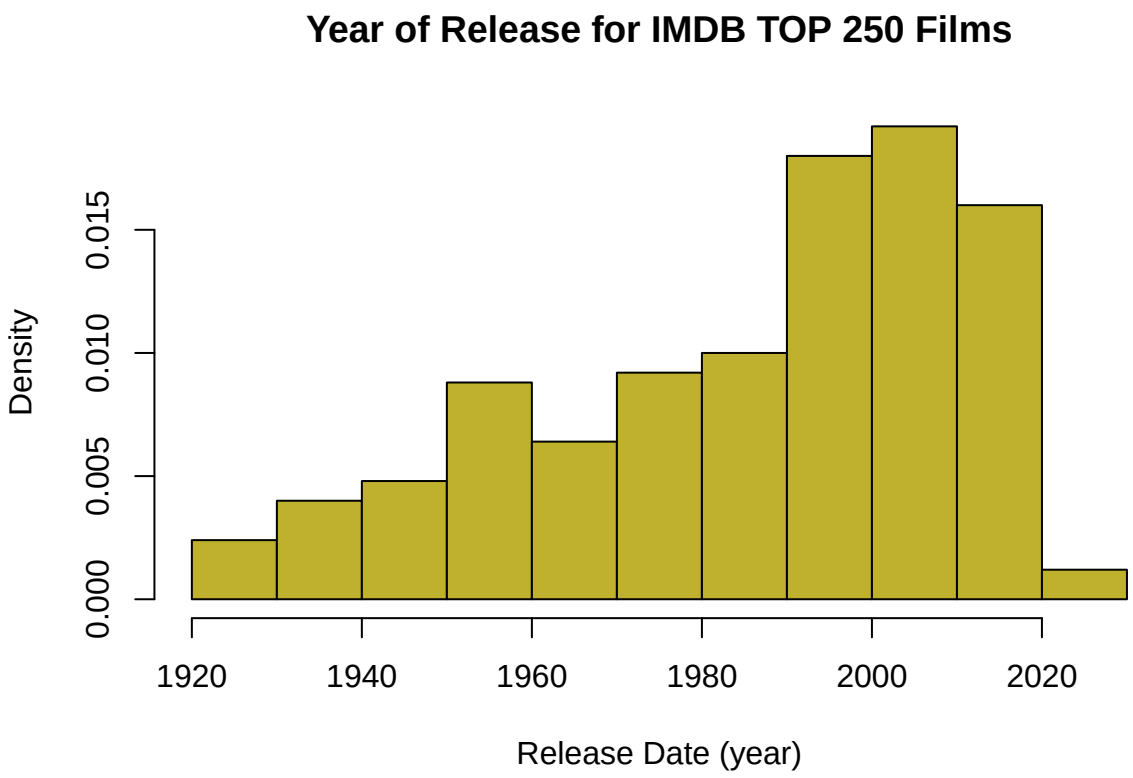
**add . . . Histograms. . .**

Let's see the distribution of films in the Top 250 based on year of release.

```
data <- read.csv('movies.csv')
release_date <- data$year
hist(release_date)
```




```
hist(release_date,
main="Year of Release for IMDB TOP 250 Films",
xlab="Release Date (year)",
col="#bfb12e",
freq=FALSE)
```




Save your Histogram in a file:

```
# Give the chart file a name.
png(file = "session1_histogram.png")

hist(release_date,
main="Year of Release for IMDB TOP 250 Films",
xlab="Release Date (year)",
col="#bfb12e")

# Save the file.
dev.off()
```

```
## pdf
##   2
```

**1. . . . Joint Probability Distributions. . .**

Random Variable $X$ has values in the sample space $S$. Random Variable $Y$ has values in the sample space $T$. Random Variable $(X, Y)$ has values in the sample space $S * T$.

The probability distribution of $(X, Y)$ is the Joint Probability Distribution. This means: $P(X = x, Y = y)$ : Which is probability of $X$ taking some value $x$, and $Y$ taking some value $y$.

Marginal Distribution is the probability of $X$ taking the value of $x$ regardless of the value $Y$ has taken. The mathematical formulation is:

$$P(X = x) = \sum_y P(X = x, Y = y)$$

» Example. The joint probability table:

| Values | Y = 0 | Y = 5 | Y = 10 | Y = 15 |
|---|---|---|---|---|
| X = 0 | .02 | .06 | .02 | .10 |
| X = 5 | .04 | .15 | .20 | .10 |
| X = 10 | .01 | .15 | .14 | .01 |

Calculate the joint probabilities and marginal probabilities for the RVs.

```
#create this table in R

joint_table <- matrix(c(.02,.04,.01,.06,.15,.15,.02,.20,.14,.10,.10,.01), ncol=4)

joint_table
```

```
##      [,1] [,2] [,3] [,4]
## [1,] 0.02 0.06 0.02 0.10
## [2,] 0.04 0.15 0.20 0.10
## [3,] 0.01 0.15 0.14 0.01
```

```
#calculate joint probability of X and Y
#it's just a table lookup!
# P(X = 5, Y = 10) is row = 2 and column = 3 --- = 0.20

p_x5_y10 <- joint_table[2,3]
print(paste('joint probability of X = 5, Y = 10 is : ', p_x5_y10))
```

```
## [1] "joint probability of X = 5, Y = 10 is :  0.2"
```

```
#calculate marginal probability of X (regardless of y's values)
#in apply: 1 means row-wise, 2 means column-wise
marginal_x <- apply(joint_table, 1, sum)

print(paste('marginal probabilities of X: ', marginal_x))
```

```
## [1] "marginal probabilities of X:  0.2"  "marginal probabilities of X:  0.49"
## [3] "marginal probabilities of X:  0.31"
```

```
marginal_y <- apply(joint_table, 2, sum)

print(paste('marginal probailities of Y: ', marginal_y))
```

```
## [1] "marginal probailities of Y:  0.07" "marginal probailities of Y:  0.36"
## [3] "marginal probailities of Y:  0.36" "marginal probailities of Y:  0.21"
```

Some resources for further reading: https://bayesball.github.io/BOOK/joint-probability-distributions.html

**2. . . . Means and Variances of Linear Combinations of Random Variables. . .**

**Theorem.** Suppose $X_1$,$X_2$, . . . , $X_n$ are $n$ independent random variables with means $\mu_1$, $\mu_2$, . . . $\mu_n$ and variances $\sigma_1^2$, $\sigma_2^2$, . . . , $\sigma_n^2$. Then the mean and variance of the linear combination $Y = \sum_i a_i * X_i$ where $a_i$'s are real constants are:

$$\mu_Y = \sum_i a_i * \mu_i$$

and

$$\sigma_Y^2 = \sum_i a_i^2 * \sigma_i^2$$

Example 2.

$$Y = X^2 - 2X + 6$$

and $X$ has the following probability distribution function:

| x | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| f | 1/6 | 1/2 | 0 | 1/3 |

calculate E(Y).

```
mean_X2 <- 0*0*(1/6) + 1*1*(1/2) + 2*2*(0) + 3*3*(1/3)
mean_X2
```

```
## [1] 3.5
```

```
mean_2X <- 0*(1/6) + 1*(1/2) + 2*(0) + 3*(1/3)
mean_2X
```

```
## [1] 1.5
```

```
mean_6 <- 6
mean_6
```

```
## [1] 6
```

```
e_y <- mean_X2 - 2 * mean_2X + mean_6
e_y
```

```
## [1] 6.5
```

Some good resources for further reading: https://www.statlect.com/probability-distributions/normal-distribution-linear-combinations

*HOMEWORK*: Calculate Variance on your own.

**ADDITIONAL . . . LLN. . .** The "weak" law of large numbers states:

Let $X_1$, $X_2$, . . . , $X_n$ be i.i.d random variables with a finite expected value of $E(X_i) = \mu$. Then for any $\epsilon > 0$:

$$\lim_{n \to \infty} P(|\overline{x} - \mu| \geq \epsilon) = 0$$

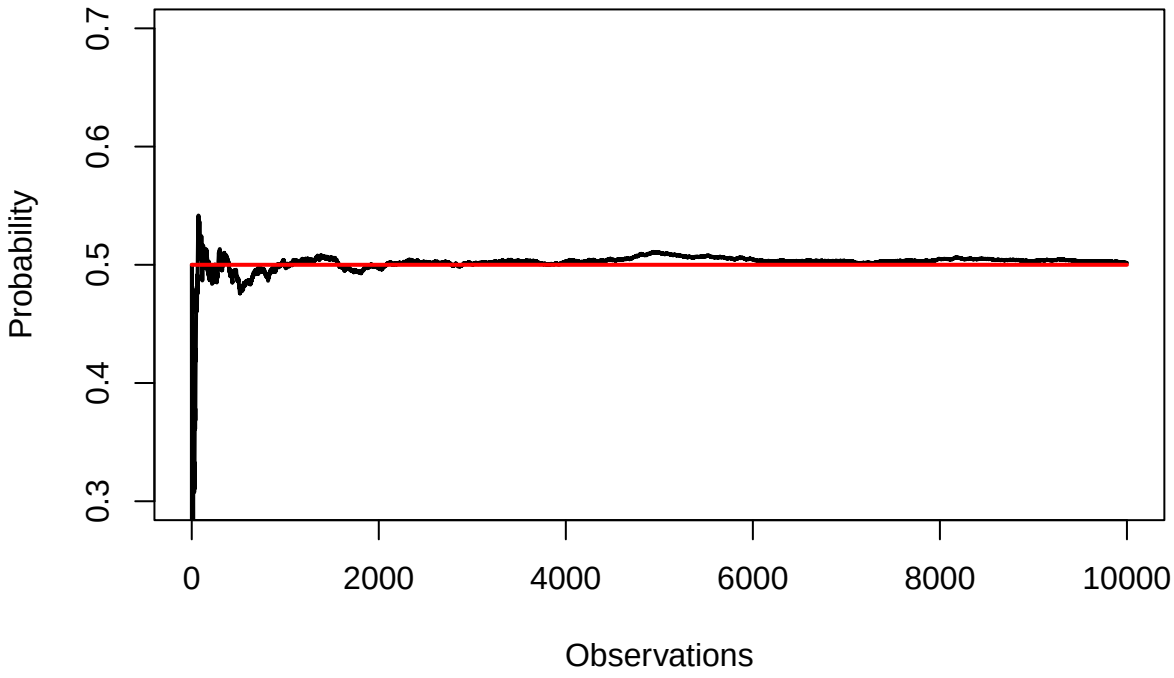
What does this mean? It means that there is a very close (only about $\epsilon$ away from the actual mean) mean value that you can find, and the more times you repeat the sampling and experiments the more likely you are to get it exactly right.

The probability of that difference between actual mean and estimated mean growing farther and farther away is going to 0 (the chances of that happening are becoming more and more unlikely) as you increase the number of n's, which is how many times you sample or repeat an experiement.

```
# source: https://rpubs.com/dolinger_nscc/lolnums

N <- 10000
o <- 10
set.seed(1)

x <- sample(0:1, N, replace = T)
s <- cumsum(x)

r.avg <- s / (1:N)
r.stats <- round(cbind(x, s, r.avg), 3)[1:o, ]
print(r.stats)
```

```
##       x s r.avg
##  [1,] 0 0 0.000
##  [2,] 1 1 0.500
##  [3,] 0 1 0.333
##  [4,] 0 1 0.250
##  [5,] 1 2 0.400
##  [6,] 0 2 0.333
##  [7,] 0 2 0.286
##  [8,] 0 2 0.250
##  [9,] 1 3 0.333
## [10,] 1 4 0.400
```

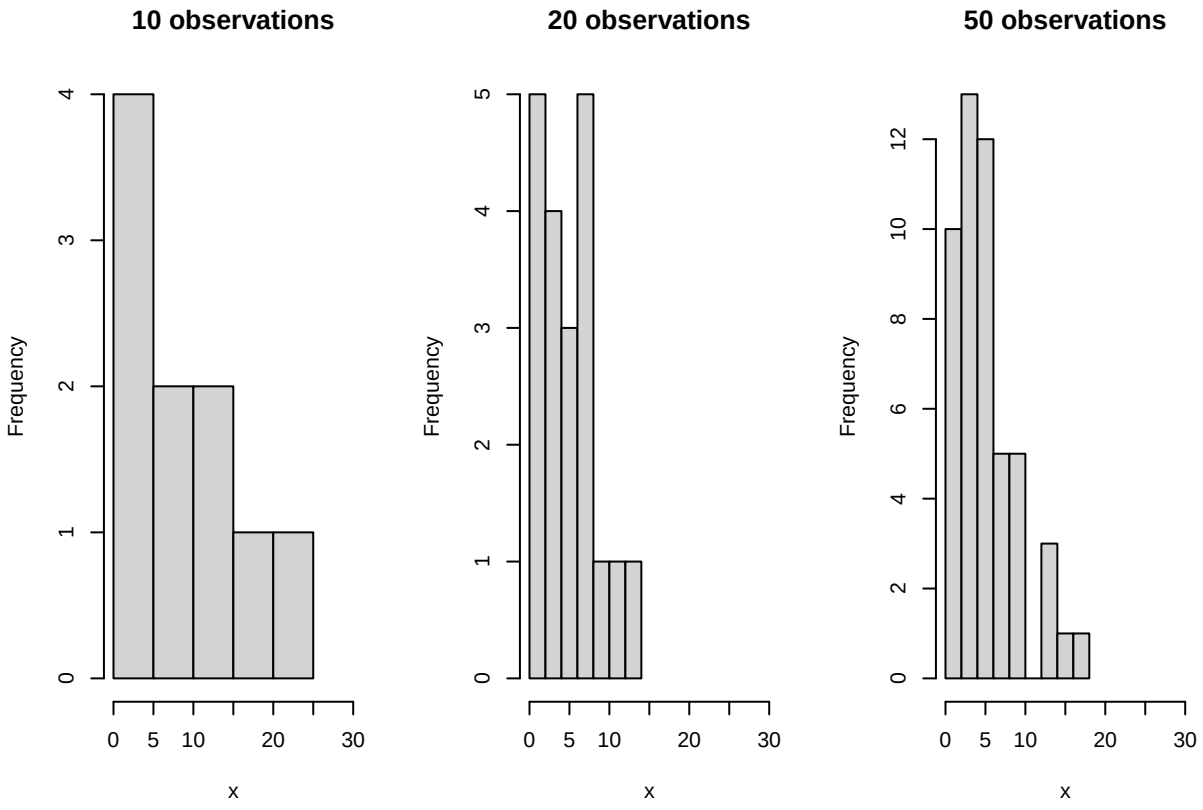
```
options(scipen = 10)
plot(r.avg, ylim=c(.30, .70), type = "l", xlab = "Observations"
     ,ylab = "Probability", lwd = 2)
lines(c(0,N), c(.50,.50),col="red", lwd = 2)
```




Coin flips example, further reading: https://uw-statistics.github.io/Stat311Tutorial/limit-theorems.html

. . . **CLT**. . . https://onlinestatbook.com/2/sampling_distributions/SampDist_v1.html https://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm

```
set.seed(12)

e10 <- rexp(n = 10, rate = 0.2)
e20 <- rexp(n = 20, rate = 0.2)
e50 <- rexp(n = 50, rate = 0.2)

par(mfrow = c(1,3))

hist(e10, xlim = c(0,30), xlab = 'x', main = '10 observations')
hist(e20, xlim = c(0,30), xlab = 'x', main = '20 observations')
hist(e50, xlim = c(0,30), xlab = 'x', main = '50 observations')
```




```
set.seed(100)

d <- runif(n=1000, min=1.2, max=6)
print(paste('dataset mean is: ', mean(d)))
```

```
## [1] "dataset mean is:  3.68679200683571"
```

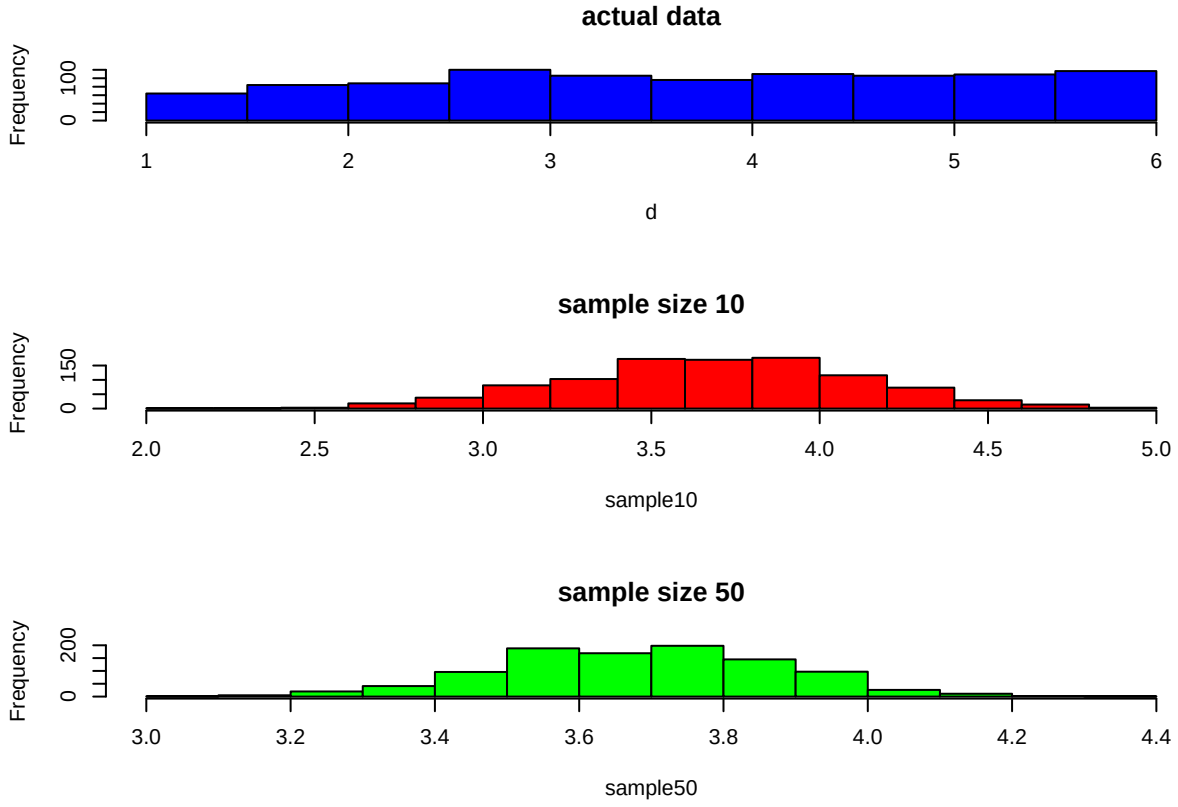
```
sample10 <- c() #empty sample
n <- 1000
for (i in 1:n) {
    sample10[i] <- mean(sample(d, 10, replace = T))
}

print(paste('10 sample mean is: ', mean(sample10)))
```

```
## [1] "10 sample mean is:  3.68431978301737"
```

```
sample50 <- c()
for (i in 1:n){
    sample50[i] <- mean(sample(d, 50, replace = T))
}

print(paste('50 sample mean is: ', mean(sample50)))
```

```
## [1] "50 sample mean is:  3.68384094880584"
```

```
par(mfrow = c(3,1))
hist(d, col = 'blue', main = 'actual data')
hist(sample10, col = 'red', main = 'sample size 10')
hist(sample50, col = 'green', main = 'sample size 50')
```




. . . **Simulation**. . . Resource: 1. https://rstudio-pubs-static.s3.amazonaws.com/301283_8ba77a4c9d8d4a2db3d07372b7b22c8
html

2. https://web.stanford.edu/class/bios221/labs/simulation/Lab_3_simulation.html