MSIS 5503 – Statistics for Data Science – Fall 2021 - Assignment 2

1. You buy a lottery ticket to a lottery that costs \$10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one \$500 prize, two \$100 prizes, and four \$25 prizes.

    a. Define the random variable.
    b. Show its probability distribution in the form of a table.
    c. Find your expected gain or loss.
    d. What is the standard deviation of you gain or loss?
    e. What type of skewness does the probability distribution represent?

Solution:

a. The random variable X = Gain/loss in Dollars from purchasing one ticket
b.

X = -10	X = 15	X = 90	X = 490
P(X) = 0.93	P(X) = 0.04	P(X) = 0.02	P(X) = 0.01

c. E(X) = (-10)*(0.93) + 15*0.04 + 90*0.02 + 490*0.01 = -9.3 + 0.6 + 1.8 + 4.90 = -2 (Loss of \$2)
d. Var(X) = $0.93(-10 - (-2))^2 + (15 - (-2))^2(0.04) + (90 - (-2))^2(0.02) + (490 - (-2))^2(0.01) = 2661$
   Therefore, Standard Deviation = SQRT(2816.31) = \$51.58
e. The distribution is right-skewed because it has a long right-tail.

2. Consider the experiment of tossing two dice. Your random variable is **D**, the **square** of difference of the numbers showing on the faces of the two dice.

    a. Show its probability distribution in the form of a table.
    b. Find the mean, median and mode of the distribution.
    c. Find the variance of **D**.
    d. What type of skewness does the probability distribution represent?
    e. Is the Chebychev inequality satisfied for c=1 and c=2? Show the calculations.

a. Show its probability distribution in the form of a table.

D	0	1	4	9	16	25	Sum
P(D)	0.1667	0.2778	0.2222	0.1667	0.1111	0.0556	1.0000
D*P(D)	0.0000	0.2778	0.8889	1.5000	1.7778	1.3889	5.8333
((D-E(D)^2)*P(D)	5.671296	6.489198	0.746914	1.671296	11.48457	20.40895	46.4722

b. Find the mean, median and mode of the distribution.
   Mean = E(D) = $\sum_{c \in A} cP(W = c)$ = 70/36 = 5.8333
   There is no exact median. We could say that median = 4, because the probability on either side of 2 is the same (16/36).
   Mode = 1, because 1 occurs most frequently(10/36)

c. Find the variance of D.
   Variance V(D) = ((D-E(D)^2)*P(D) = 46.4722

d. What type of skewness does the probability distribution represent?
   This probability distribution is right-skewed, with a long right-tail. The median will be less than the mean of 5.833.

e. Is the Chebychev inequality satisfied for c=1 and c=2? Show the calculations.

$P\ (x\ > \mu + c\sigma) + P\ (x < \mu - c\sigma) < \frac{1}{c^2}$ is the Chebychev inequality.
Here σ = Standard deviation = $\sqrt{V(D)}$ = 6.817

For c=1:

$P\ (D\ > 5.8333 + 6.817) + P\ (D < 5.8333 - 6.817) < 1$
$P\ (D\ > 12.65) + P\ (D < -0.984) < 1$
$0.1666 + 0 < 1$
Hence, the Chebychev inequality is satisfied.

For c=2:

$P\ (D\ > 5.8333 + 2 * 6.817) + P\ (D < 5.8333 - 2 * 6.817) < 0.25$
$P\ (D\ > 19.4673) + P\ (D < -7.8007) < 0.25$
$0.0556 + 0 < 0.25$
Hence, the Chebychev inequality is satisfied.

3. Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.
   a. In words, define the random variable X.
      X is the number of students who attend their graduation.
   b. List the values that X may take on.
      X=0,1,2,…,22
   c. Give the distribution of X. *X ~ Binomial(n=22,p=0.85)*
   d. How many are expected to attend their graduation?
      *Expected value of Binomial = np=22*0.85 = 18.7 students.*
   e. Find the probability that 17 or 18 attend.
      *P(X=17)= (22!/17!*5!) *$(0.85)^{17}(0.15)^{5}$ = 0.126;*
      *P(X= 18) = (22!/18!*4!) *$(0.85)^{18}(0.15)^{4}$ = 0.198*
      The probability that 17 or 18 will attend will be =0.324 because (X=17) and (X=18) are mutually exclusive.
   f. Find the probability that at most 15 will attend (practice with calculator).
      The probability that at most 15 will attend = P(X <= 15) = 1 – P(X>=16) = P(16) +P(17) +…+P(22) = 0.03684.
   g. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.
      *P(X=22) = (22!/22!*0!) *$(0.85)^{22}(0.15)^{0}$ = 0.028; Yes, because of the low probability it would be a surprise if all 22 attended.*

4. For the previous problem, do the following using **R**. ***Copy and paste the R commands <u>from the console</u>, along with the output.***
   a. Find the probability that at least 16 will attend the graduation.
   b. There is a 30% chance that at least ________ will attend the graduation.
   c. There is a 10% chance that at most _________ will attend the graduation.
   d. Show the table of probabilities and cumulative probabilities.
   e. The empirical mean ______ is and the empirical standard deviation is _______.
   f. Show the histogram of the pmf and cdf.

   a.

```
> p16 <- 1 - pbinom(15, 22, 0.85, lower.tail = TRUE, log.p = FALSE)
> print(paste("Probability that at least 16 will attend is",round(p16, 4), sep = " "))
[1] "Probability that at least 16 will attend is 0.9632"
```

b. We need x such that P(X > x) = 0.30. That means, we need the 70th percentile (quantile) of the distribution.
Answer:
`qbinom(0.70, 22, 0.85, lower.tail = TRUE, log = FALSE)` with answer = 20

c. We need x such that P(X <= x) = 0.10. That means, we need the 10th percentile (quantile) of the distribution.
`qbinom(0.10, 22, 0.85, lower.tail = TRUE, log = FALSE)` with answer = 17

d.

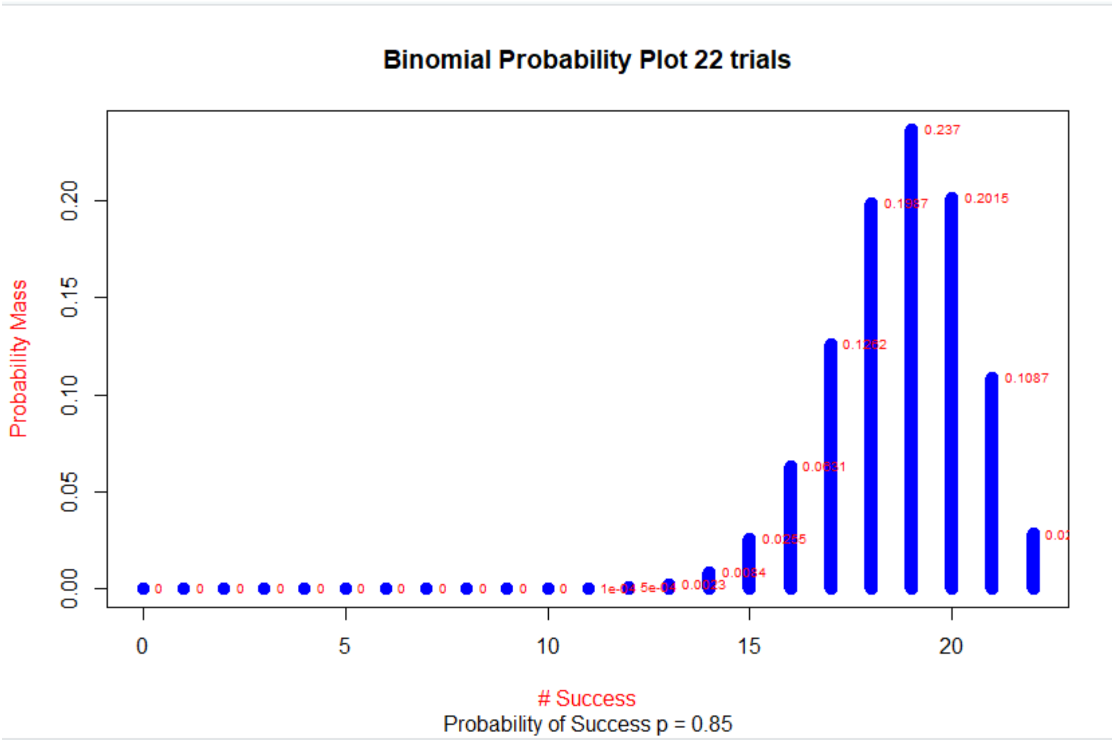
```
> #table of probabilities
> for (i in 0:nsize) {
+   result[i+1] <- dbinom(i, nsize, prob, log = FALSE)
+   xPx[i+1] <- i*result[i+1]
+   x2Px[i+1] <- i*xPx[i+1]
+   cum_result[i+1] <- pbinom(i, nsize, prob, log = FALSE)
+   print(paste("X = ",i, "probability = ",round(result[i+1],4),
+               "cumumlative probability = ", round(cum_result[i+1],4), sep = " "))
+ }
[1] "X =  0 probability =  0 cumumlative probability =  0"
[1] "X =  1 probability =  0 cumumlative probability =  0"
[1] "X =  2 probability =  0 cumumlative probability =  0"
[1] "X =  3 probability =  0 cumumlative probability =  0"
[1] "X =  4 probability =  0 cumumlative probability =  0"
[1] "X =  5 probability =  0 cumumlative probability =  0"
[1] "X =  6 probability =  0 cumumlative probability =  0"
[1] "X =  7 probability =  0 cumumlative probability =  0"
[1] "X =  8 probability =  0 cumumlative probability =  0"
[1] "X =  9 probability =  0 cumumlative probability =  0"
[1] "X =  10 probability =  0 cumumlative probability =  0"
[1] "X =  11 probability =  1e-04 cumumlative probability =  1e-04"
[1] "X =  12 probability =  5e-04 cumumlative probability =  7e-04"
[1] "X =  13 probability =  0.0023 cumumlative probability =  0.003"
[1] "X =  14 probability =  0.0084 cumumlative probability =  0.0114"
[1] "X =  15 probability =  0.0255 cumumlative probability =  0.0368"
[1] "X =  16 probability =  0.0631 cumumlative probability =  0.0999"
[1] "X =  17 probability =  0.1262 cumumlative probability =  0.2262"
[1] "X =  18 probability =  0.1987 cumumlative probability =  0.4248"
[1] "X =  19 probability =  0.237 cumumlative probability =  0.6618"
[1] "X =  20 probability =  0.2015 cumumlative probability =  0.8633"
[1] "X =  21 probability =  0.1087 cumumlative probability =  0.972"
[1] "X =  22 probability =  0.028 cumumlative probability =  1"
```

e.

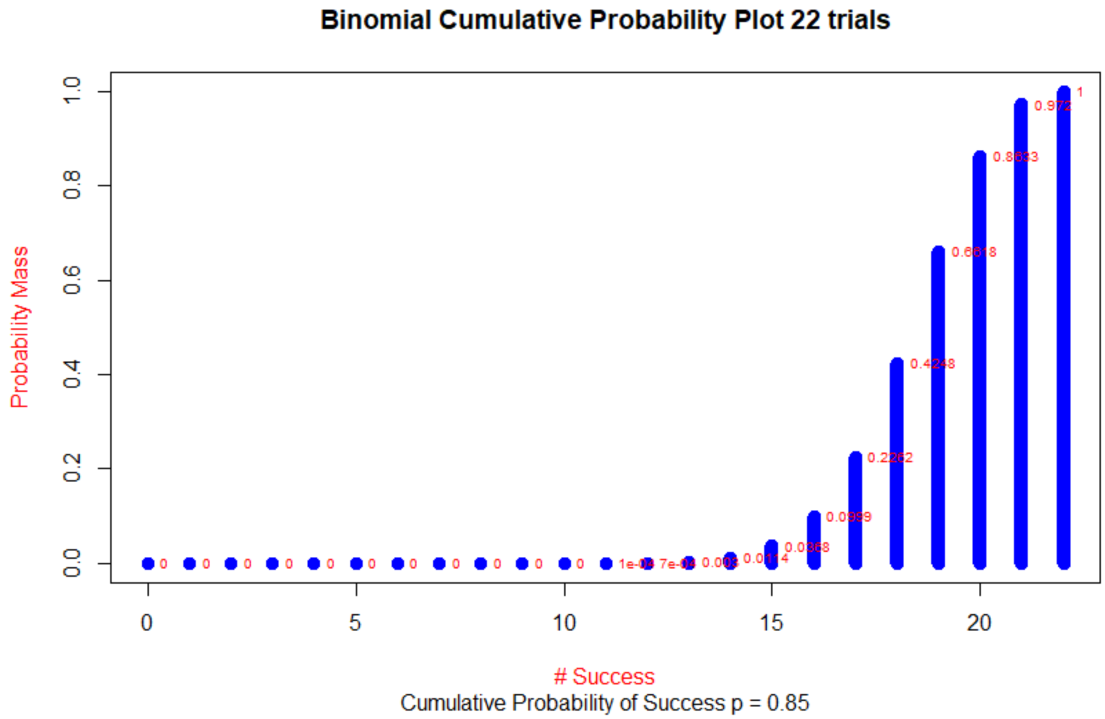
```
> # Checking the mean = np and variance = np(1-p)
>
> # Mean = sum of xPx
> Exp_val = sum(xPx)
> print(paste("The Expected Value is",round(Exp_val, 4), sep = " "))
[1] "The Expected Value is 18.7"
>
> # Var = sum of x2Px - (sum(xPx)^2)
> varian = sum(x2Px) - Exp_val*Exp_val
> print(paste("The standard deviation is",round(sqrt(varian), 4), sep = " "))
[1] "The standard deviation is 1.6748"
```

f.

```
> # Plot the probabilities
> len_result <- length(result)
> indx = len_result - 1
> plot((0:indx),result[1:len_result], type = "h", main = "Binomial Probability Plot 22 trials",
+      sub = "Probability of Success p = 0.85", xlab = "# Success", ylab = "Probability Mass",
+      col  = "blue", col.lab ="red", lwd=10)
> text((0:indx), result[1:len_result], round(result[1:len_result], 4), cex=0.6, pos=4, col="red")
>
```




```
> # Plot the cumulative probabilities
> len_result <- length(result)
> indx = len_result - 1
> plot((0:indx),cum_result[1:len_result], type = "h", main = "Binomial Cumulative Probability Plot 22 trials",
+      sub = "Cumulative Probability of Success p = 0.85", xlab = "# Success", ylab = "Probability Mass",
+      col  = "blue", col.lab ="red", lwd=10)
> text((0:indx), cum_result[1:len_result], round(cum_result[1:len_result], 4), cex=0.6, pos=4, col="red")
> |
```



Binomial Cumulative Probability Plot 22 trials
Probability Mass
Success
Cumulative Probability of Success p = 0.85