MA 354: Data Analysis – Fall 2021 – Due 10/8 at 5p
Homework 2:

Complete the following opportunities to use what we've talked about in class. These questions will be graded for correctness, communication and succinctness. Ensure you show your work and explain your logic in a legible and refined submission.

The starting jobs will be applied in alphabetical order (last name) for question two.

1. **Solver:** provide a solution, if possible, and reasoning for the solution. **Due to group 10/5 or earlier.**
2. **Code Checker:** provides a first check of the solver's worked solutions and ensures they are correct with a solid interpretation.
3. **Checker** checks the solution for completeness, proposes and implements changes if agreed upon by the group. Provides a short paragraph summarizing the discussion of proposals and their reason for acceptance or non-acceptance.
4. **Double Checker** checks the solution for completeness, communication and polish. The Double Checker ensures that the solution is correct and highly polished for submission.

For subsequent questions student roles will move down one position. The rolls change as follows.

1. **Solver** $\Longrightarrow$ **Code Checker**
2. **Code Checker** $\Longrightarrow$ **Checker**
3. **Checker** $\Longrightarrow$ **Double Checker**
4. **Double Checker** $\Longrightarrow$ **Solver**

While students have assigned jobs for each question I encourage students to help each other complete the homework in collaboration.

1. Select a continuous distribution (Not the uniform or exponential). It does not have to be one that we cover in the notes! To explore the PDF of your distribution, specify two sets of parameter(s) for your distribution.

   (a) **History** Discuss what types of random variables are modeled with your distribution. Be sure to include a discussion about the support and ensure to provide the density function, and CDF. This requires some internet research – what's the history of the distribution, why was it created and named? What are some exciting applications of this distribution?

   Cite all of your sources in LaTeX by adding a BibTeX citation to the .bib file. To help, I've cited R (R Core Team, 2021) in parentheses here. R Core Team (2021) provides helpful tools for the rest of the questions below. BibTeX citations are available through Google Scholar by clicking the cite button below the article of interest and selecting the BibTeX option.

   (b) Show that you have a valid PDF. You will find the `integrate()` function in `R` helpful.

   (c) Find the median for your two sets of parameter(s). Conduct some research to find the median based on our PDF to confirm that your numerical approach is correct.

   (d) Graph the PDF for several values of the parameter(s) including the two sets you specified. What does changing the parameter(s) do to the shape of the PDF?

   (e) Graph the CDF for the same values of the parameter(s) as you did in Question 1d. What does changing the parameter(s) do to the shape of the CDF? Comment on the aspects of the CDFs that show that the CDF is valid.

   (f) Generate a random sample of size $n = 10, 25, 100$, and 1000 for your two sets of parameter(s). In a 4×2 grid, plot a histogram of each set of data and superimpose the true density function at the specified parameter values. Interpret the results.

2. Continue with the continuous distribution you selected for Question 1.

   (a) Provide the mean, standard deviation, skewness, and kurtosis of the PDF. Ensure to interpret each.

   (b) Generate a random sample of size $n = 10, 25, 100$, and 1000 for your two sets of parameter(s). Calculate the sample mean, standard deviation, skewness, and kurtosis. Interpret the results.

   (c) Generate a random sample of size $n = 10$ for your two sets of parameter(s). Calculate the method of moments estimator(s) and maximum likelihood estimator(s). In a 1×2 grid, plot a histogram of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.

   (d) Generate a random sample of size $n = 25$ for your two sets of parameter(s). Calculate the method of moments estimator(s) and maximum likelihood estimator(s). In a 1×2 grid, plot a histogram of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.

   (e) Generate a random sample of size $n = 100$ for your two sets of parameter(s). Calculate the method of moments estimator(s) and maximum likelihood estimator(s). In a 1×2 grid, plot a histogram of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.

   (f) Generate a random sample of size $n = 100$ for your two sets of parameter(s). Calculate the method of moments estimator(s) and maximum likelihood estimator(s). In a 1×2 grid, plot a histogram of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.

   (g) Comment on the results of parts (c)-(f).

3. Select a discrete distribution (not the Poisson). It does not have to be one that we cover in the notes! To explore the PMF of your distribution, specify two sets of parameter(s) for your distribution.

   (a) **History** Discuss what types of random variables are modeled with your distribution. Be sure to include a discussion about the support and ensure to provide the mass function, and CDF. This requires some internet research – what's the history of the distribution, why was it created and named? What are some exciting applications of this distribution? Cite all of your sources.

   (b) Show that you have a valid PMF. You can show this approximately by calculating the series in a repeat loop until probability mass evaluations are infinitesimally small.

   (c) Find the median for your two sets of parameter(s). Conduct some research to find the median based on our PMF to confirm that your numerical approach is correct.

   (d) Graph the PMF for several values of the parameter(s) including the two sets you specified. What does changing the parameter(s) do to the shape of the PMF?

   (e) Graph the CDF for the same values of the parameter(s) as you did in Question 3d. What does changing the parameter(s) do to the shape of the CDF? Comment on the aspects of the CDFs that show that the CDF is valid.

   (f) Generate a random sample of size $n = 10, 25, 100$, and 1000 for your two sets of parameter(s). In a 4×2 grid, plot a histogram (with bin size 1) of each set of data and superimpose the true mass function at the specified parameter values. Interpret the results.

4. Continue with the discrete distribution you selected for Question 3.

   (a) Provide the mean, standard deviation, skewness, and kurtosis of the PMF. Ensure to interpret each.
   Solution:
   The binomial distribution helps model discrete sets of data with a certain number of trials with a probability of success. The mean of this distribution is

   $$\textbf{Mean} = np$$

   This makes sense, because the average number of successes is given by the above formula. The standard deviation of this distribution is

   $$\textbf{SD} = np(1-p)$$

   This is just multiplying the probability of failiure with the mean, which then gives the variance. The skewness is

   $$\textbf{Skewness} = \frac{1-2p}{\sqrt{np(1-p)}}$$

   For a small p and small n, the distribution will be skewed right since there are a low number of successes. For a large p and small n, the distribution is skewed left, since there are a high number of successes. For p = 0.5, the skewness is 0, since the distribution becomes symmetric. The kurtosis of the distribution is

   $$\textbf{Kurtosis} = \frac{1-6p(1-p)}{np(1-p)}$$

   The kurtosis tells us the rate of outliers in a given set of data.

   (b) Generate a random sample of size $n = 10, 25, 100$, and 1000 for your two sets of parameter(s). Calculate the sample mean, standard deviation, skewness, and kurtosis. Interpret the results.
   Solution:

```r
library(e1071) #allows skewness and kurtosis
#The two sets of parameters are -
#40 coin tosses with the success being a heads
set.seed(345) #helps with testing again
t1 = 40
prob1 = 0.5
#20 rolls of a D4 with success being a 1 rolled
t2 = 20
prob2 = 0.25
#generating each set of data for each parameter
n.10.1 <- rbinom(10, t1, prob1)
n.10.2 <- rbinom(10, t2, prob2)
n.25.1 <- rbinom(25, t1, prob1)
n.25.2 <- rbinom(25, t2, prob2)
n.100.1 <- rbinom(100, t1, prob1)
n.100.2 <- rbinom(100, t2, prob2)
n.1000.1 <- rbinom(1000, t1, prob1)
n.1000.2 <- rbinom(1000, t2, prob2)
#making a dataframe of the samples
stat1 <- data.frame(n.10.1,
                    n.25.1,
                    n.100.1,
                    n.1000.1,
                    n.10.2,
                    n.25.2,
                    n.100.2,
                    n.1000.2)
#applying the R functions and listing the results
#Mean
apply(stat1, 2, mean)
```

```
##   n.10.1   n.25.1  n.100.1 n.1000.1   n.10.2   n.25.2  n.100.2 n.1000.2
##    19.60    20.52    19.84    19.88     5.90     4.28     4.98     5.01
```

```r
#Standard Deviation
apply(stat1, 2, sd)
```

```
##   n.10.1   n.25.1  n.100.1 n.1000.1   n.10.2   n.25.2  n.100.2 n.1000.2
## 2.108185 3.603135 3.121561 3.203102 2.344247 1.662638 1.828372 1.886653
```

```r
#Skewness
apply(stat1, 2, skewness)
```

```
##      n.10.1      n.25.1     n.100.1    n.1000.1      n.10.2      n.25.2
##  0.27663605 -0.23458002  0.22147552  0.08987464 -0.29900331  0.59578386
##     n.100.2    n.1000.2
##  0.02951240  0.19109951
```

```r
#kurtosis
apply(stat1, 2, kurtosis)
```

```
##      n.10.1      n.25.1     n.100.1    n.1000.1      n.10.2      n.25.2
## -0.70875300 -0.49916581  0.01274940 -0.05466433 -1.40963737  0.94517654
##     n.100.2    n.1000.2
## -0.25228810 -0.04125743
```

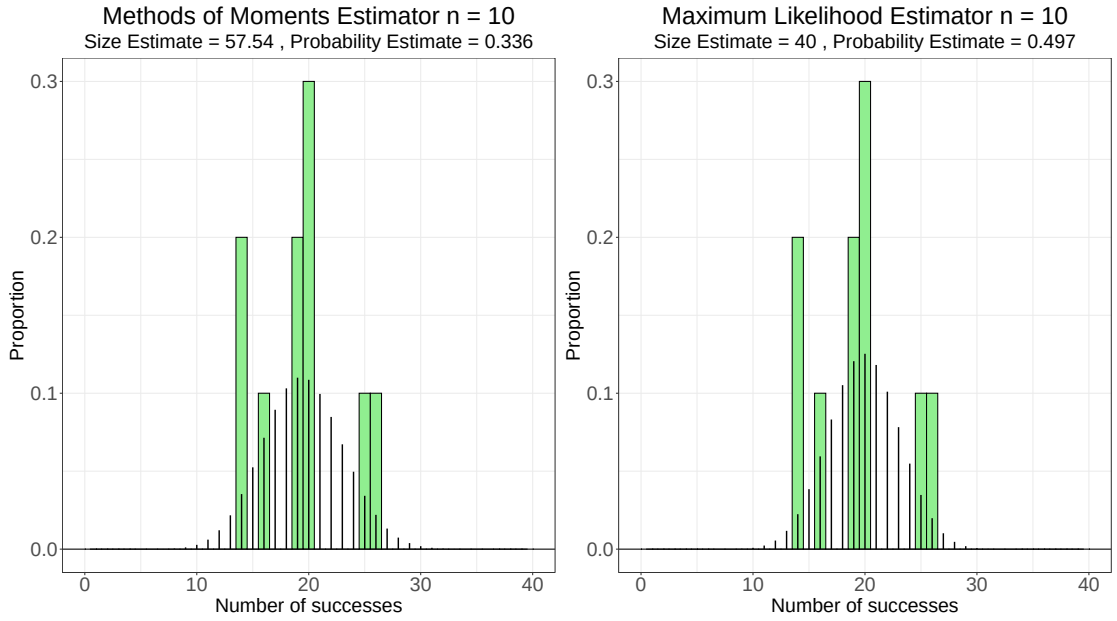
The mean is usually close to the the actual mean regardless of number of samples. It would make sense for the standard deviation to increase with the sample size, since there are more values. The skewness and kurtosis both tend to decreases with a higher number of samples.

(c) Generate a random sample of size $n = 10$ for your two sets of parameter(s). Calculate the method of moments estimator(s) and maximum likelihood estimator(s). In a 1×2 grid, plot a histogram (with bin size 1) of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.
Solution: The functions that are used for parts c through f are listed below.

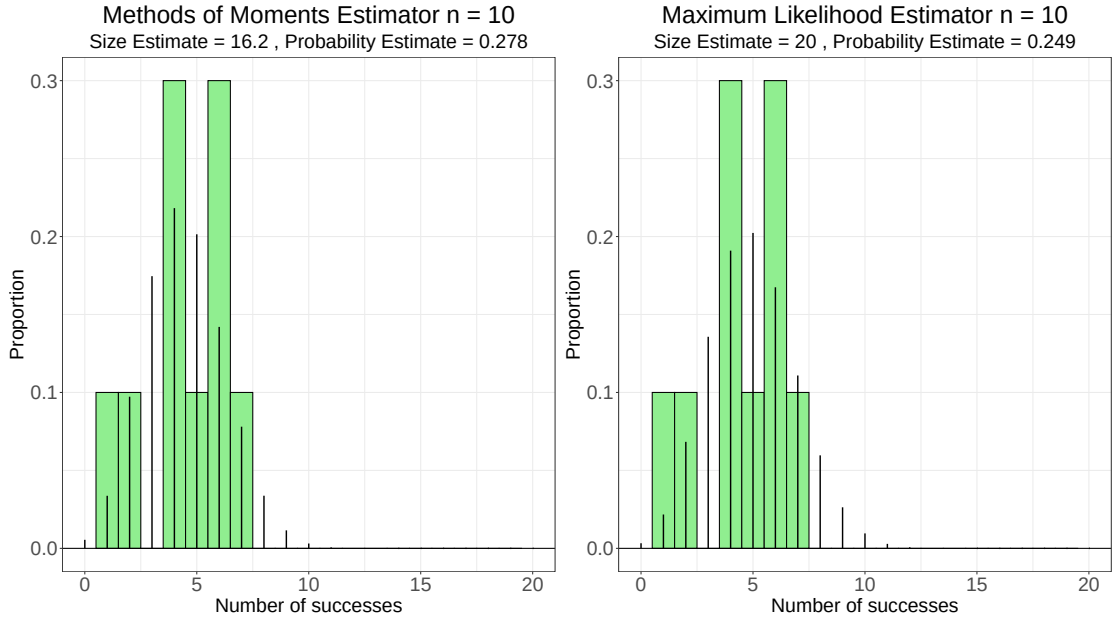
```
set.seed(32423)
library(tidyverse)
library(patchwork)
binom.mom <- function(par, data){#calculates binomial MOM Estimator
  #adapted from Prof. Cipolli's Chapter 7 notes
  n <- par[1] #par[1] has the size n
  p <- par[2] # par[2] has the probability p
  #Here, k = 2 since estimator is highly variable at k>2
  #First two population moments
  EX1 <- n*p
  EX2 <- n*p*(1-p + n*p) #moments found online cite
  eq1 <- EX1 - mean(data) #sample moment 1
  eq2 <- EX2 - mean(data^2) #sample moment 2
  c(eq1, eq2)
}
# calculates Maximum Likelihood Estimator via negative log likelihood
binom.MLE <- function(par, data, neg = T){
  #adapted from Prof. Cipolli's Chapter 7 notes
  n <- par[1]
  p <- par[2]
  #sums up the probability mass function for the sample
  MLE <- sum(dbinom(x = data, size = n, prob = p, log=TRUE))
  ifelse(!neg,MLE,-MLE) #just in case neg is changed
}

library(nleqslv)# used to calculate MOM

find.binom.mom.mle <-function(n, par){
  #function that calculates and plots MOM and MLE
Sample = rbinom(n, size = par[1], prob =par[2])
#creating the sample for this set of parameters
moms<-nleqslv(x = c(par[1], par[2]), #passes in size and probability
fn = binom.mom,
data=Sample)
#this essentially minimizes the difference between the sample and population
#moments. It then returns the corresponding n and p
mles <- optim(fn = binom.MLE,
              par = c(par[1], par[2]),
              data = Sample)
#mles optimizes the MLE for the given parameters
  plot_mom = ggplot() +
    geom_histogram(aes(x = Sample,
                       y = ..density..),
                   binwidth = 1,
                   color = "black",
```

```
                    fill ="lightgreen") +
    geom_hline(yintercept = 0) +
    theme_bw() +
    xlim(0, par[1]) +
    geom_linerange(size = 0.7, aes(x=0:par[1],
                        ymin=0,
                        ymax=dbinom(0:par[1],
                                    size = round(moms$x[1]),
                                    prob = moms$x[2]))) +
    xlab("Number of successes") +
    ylab("Proportion") +
    ggtitle(paste("Methods of Moments Estimator n =", n, sep=" "),
            subtitle = paste("Size Estimate =", round(moms$x[1], 3),
                             ", Probability Estimate =", round(moms$x[2], 3),
                             sep =" ")) +
    theme(plot.title = element_text(hjust = 0.5, size = 25),
          plot.subtitle = element_text(hjust = 0.5, size = 20),
          axis.text=element_text(size=20),
          axis.title.x = element_text(size = 20),
          axis.title.y = element_text(size = 20))


  plot_mle = ggplot() +
    geom_histogram(aes(x = Sample,
                        y = ..density..),
                   binwidth = 1,
                   color = "black",
                   fill ="lightgreen") +
    geom_hline(yintercept = 0) +
    theme_bw() +
    xlim(0, par[1]) +
    geom_linerange(size = 0.7,aes(x=0:par[1],
                        ymin=0,
                        ymax=dbinom(0:par[1],
                                    size = round(mles$par[1]),
                                    prob = mles$par[2]))) +
    xlab("Number of successes") +
    ylab("Proportion") +
    ggtitle(paste("Maximum Likelihood Estimator n =", n, sep=" "),
            subtitle = paste("Size Estimate =", round(mles$par[1], 3),
                             ", Probability Estimate =", round(mles$par[2], 3),
                             sep =" ")) +
    theme(plot.title = element_text(hjust = 0.5, size = 25),
          plot.subtitle = element_text(hjust = 0.5, size = 20),
          axis.text=element_text(size=20),
          axis.title.x = element_text(size = 20),
          axis.title.y = element_text(size = 20))


      plot_mom+plot_mle
}
```

```
find.binom.mom.mle(10, c(t1, prob1))
```

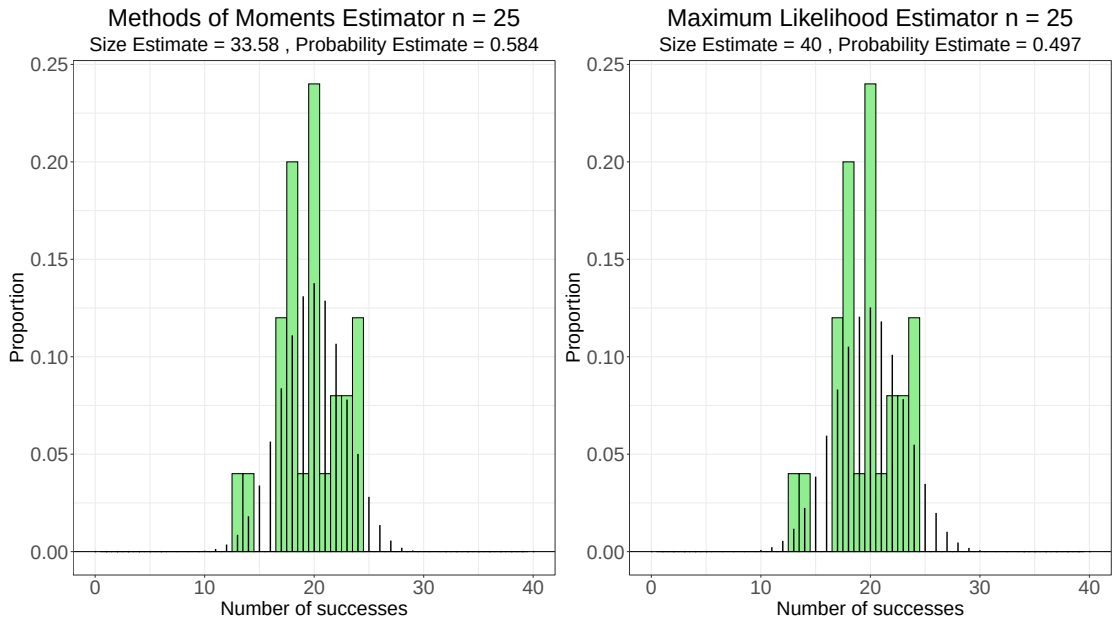



```
find.binom.mom.mle(10, c(t2, prob2))
```

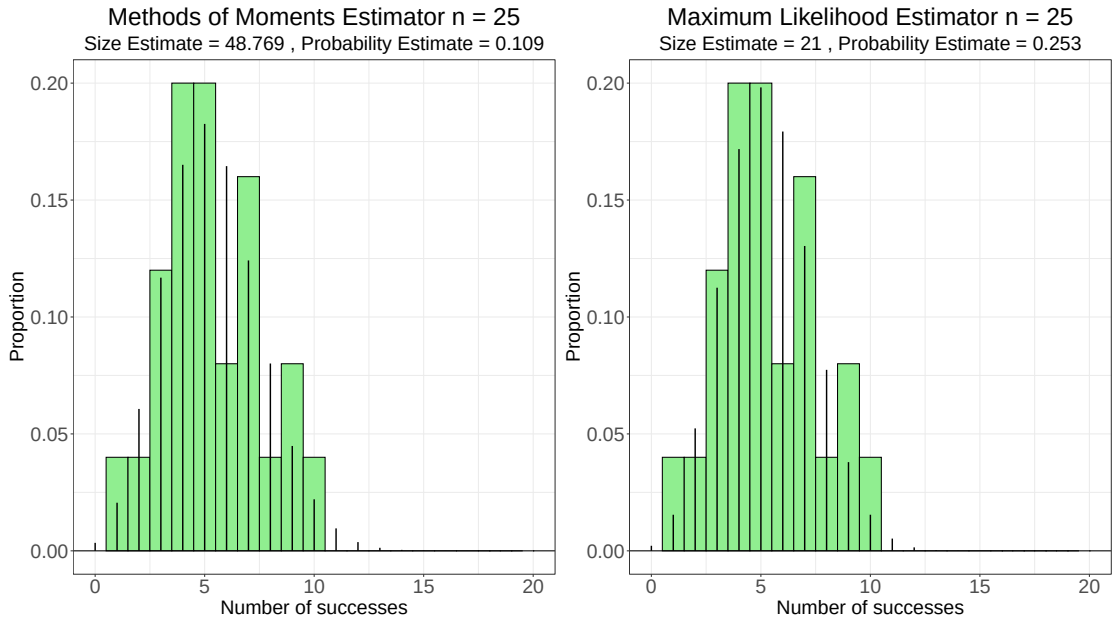


(d) Generate a random sample of size $n = 25$ for your two sets of parameter(s). Calculate the method of moments estimator(s) and maximum likelihood estimator(s). In a 1×2 grid, plot a histogram (with bin size 1) of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.
Solution:

```
find.binom.mom.mle(25, c(t1, prob1))
```

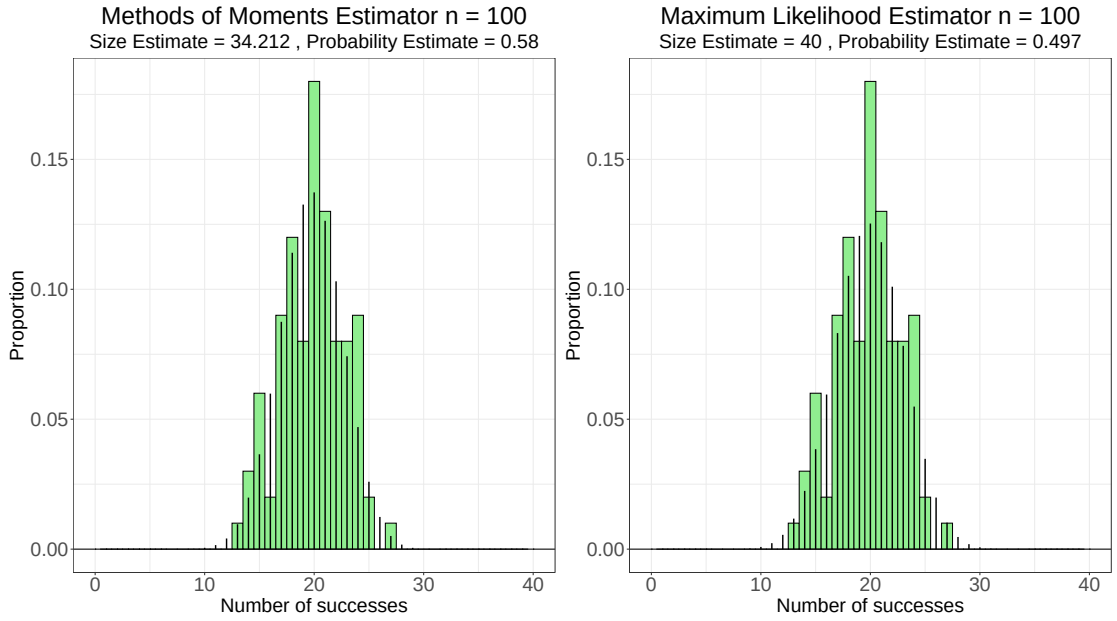



```
find.binom.mom.mle(25, c(t2, prob2))
```

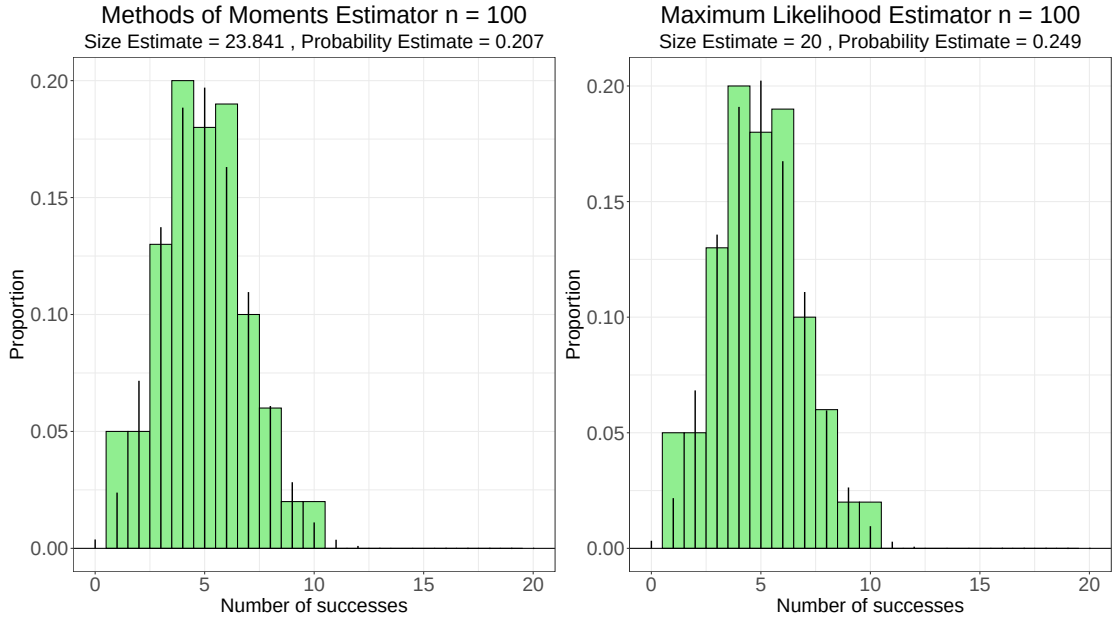


(e) Generate a random sample of size $n = 100$ for your two sets of parameter(s). Calculate the method of moments estimator(s) and maximum likelihood estimator(s). In a 1×2 grid, plot a histogram (with bin size 1) of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.
Solution:

```
find.binom.mom.mle(100, c(t1, prob1))
```

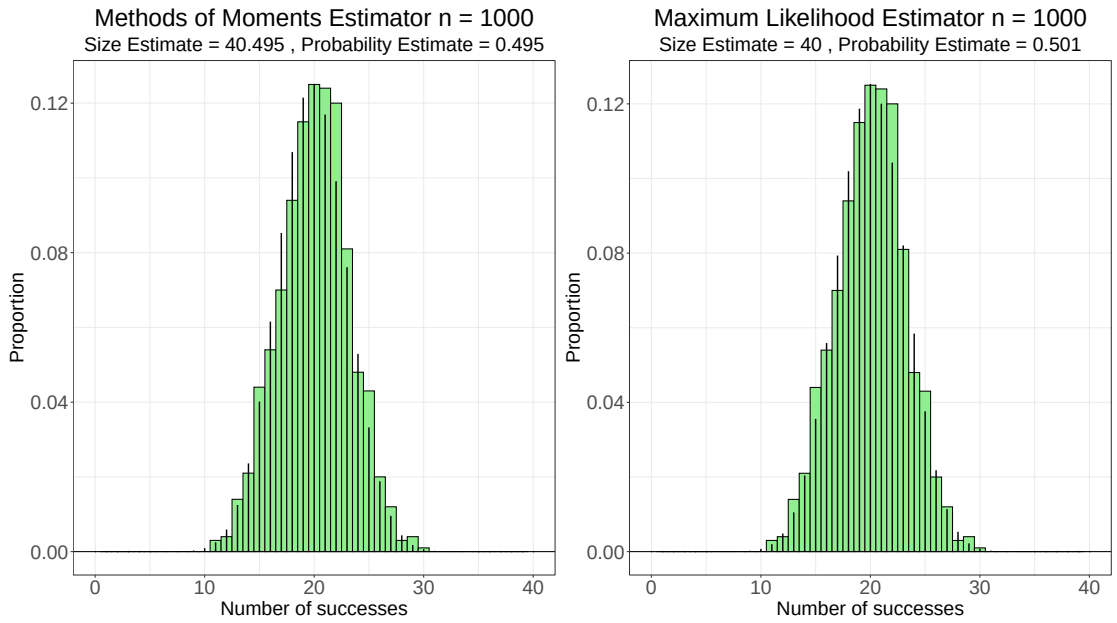



```
find.binom.mom.mle(100, c(t2, prob2))
```

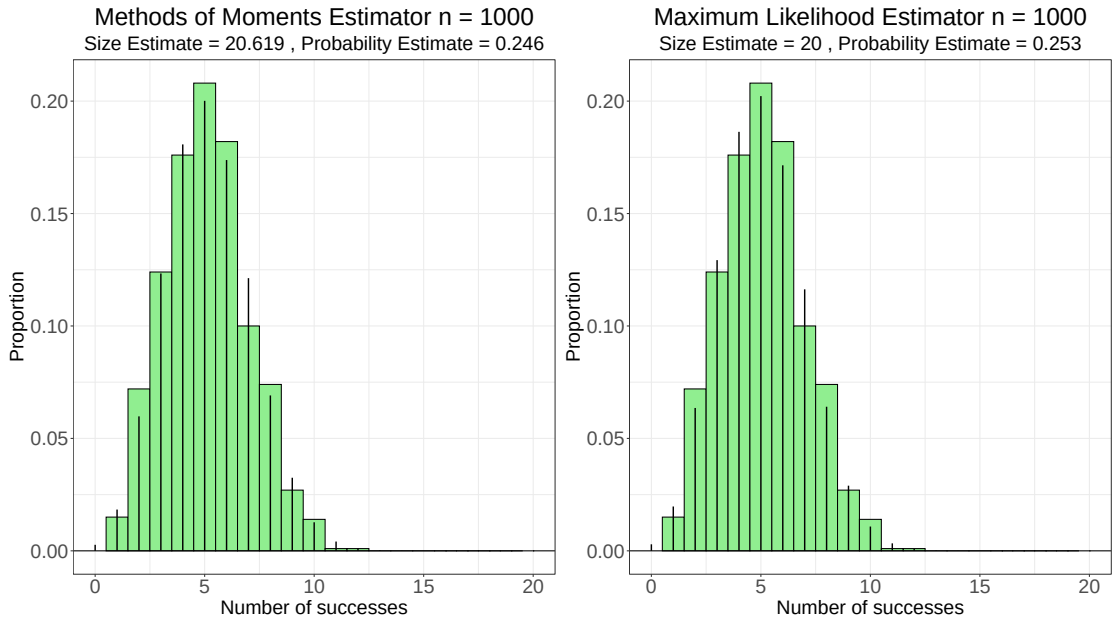


(f) Generate a random sample of size $n = 100$ for your two sets of parameter(s). Calculate the method of moments estimator(s) and maximum likelihood estimator(s). In a 1×2 grid, plot a histogram (with bin size 1) of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.
Solution:

```
find.binom.mom.mle(1000, c(t1, prob1))
```




```
find.binom.mom.mle(1000, c(t2, prob2))
```



(g) Comment on the results of parts (c)-(f).
Solution: As seen in the plots above, the size estimate for the MOM Estimator is very inaccurate for n = 10 and n = 25, but gets increasingly closer as n increases. This corresponds with the Weak Law of Large Numbers. This pattern is also seen for the MOM Estimator of the probability estimate. The MLE estimator is a different since the size estimate provided by it is almost always completely accurate, and the probability estimate is very close even for n = 10. To find out which Estimator is better, we would need to run this for a larger array of samples, since it is completely possible that the MLE is not as good at providing estimate for some other parameters. Overall, these estimates look very close to the data generates, especially at higher n values.

References

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.