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.

$$
\begin{aligned}
E(x) &= \mu && \textbf{[Mean]} \\
var(X) &= \sigma^2 && \textbf{[Variance]} \\
skew(X) &= 0 && \textbf{[Skewness]} \\
kurt(X) &= 0 && \textbf{[Kurtosis]}
\end{aligned}
$$

   The population skewness of the PDF is 0, which indicates that the Normal distribution is symmetric (with line of symmetry at the mean). The population excess kurtosis of the PDF is 0, which indicates that the Normal distribution is mesokurtic.

   (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.

```r
library(e1071)
library(tidyverse)
library(patchwork)
obs <- c(10, 25, 100, 1000)
```

```r
s1 <- data.frame(x = rnorm(n = obs[1], mean = 0, sd = 1))
s1_stats <- s1 %>%
  summarize(Mean = mean(x), SD = sd(x), Skewness = skewness(x), "Excess Kurtosis" = kurtosis(x)
s1_stats
```

```
##         Mean        SD  Skewness Excess Kurtosis
## 1 -0.2891175 0.7894614 0.8377559      -0.1875298
```

Interpretation:
Center: The sample mean of a normal distribution is the balancing point of the distribution. In this case, this will be equal to around -

```
s1_stats$Mean
```

```
## [1] -0.2891175
```

Spread: The average distance from the sample mean is -

```
s1_stats$SD
```

```
## [1] 0.7894614
```

Skewness: The skewness of a dataset describes the symmetry of a distribution.

```
s1_stats$Skewness
```

```
## [1] 0.8377559
```

Since skewness ¿ 0, there are more observations for low values and fewer with high values. Kurtosis describes how the peaked data is in relation to the Gaussian distribution (the bell-shaped curve).

```
s1_stats$`Excess Kurtosis`
```

```
## [1] -0.1875298
```

Since the excess kurtosis ¡ 0, it indicates that the data is platykurtic.

```
s2 <- data.frame(x = rnorm(n = obs[2], mean = 0, sd = 1))
s2_stats <- s2 %>% summarize(Mean = mean(x), SD = sd(x), Skewness = skewness(x), "Excess Kurtos
s2_stats
```

```
##         Mean       SD   Skewness Excess Kurtosis
## 1 -0.1022959 1.157837 -0.4381694      -0.6596139
```

Interpretation:
Center: The sample mean of a normal distribution is the balancing point of the distribution. In this case, this will be equal to around -

```
s2_stats$Mean
```

```
## [1] -0.1022959
```

Spread: The average distance from the sample mean is -

```
s2_stats$SD
```

```
## [1] 1.157837
```

Skewness: The skewness of a dataset describes the symmetry of a distribution.

```
s2_stats$Skewness
```

```
## [1] -0.4381694
```

Since skewness is slightly greater than 0, there are more observations for low values and fewer with high values. However, note that the skewness for this sample is closer to 0 than s1.
Kurtosis describes how the peaked data is in relation to the Gaussian distribution (the bell-shaped curve).

```
s2_stats$`Excess Kurtosis`
```

```
## [1] -0.6596139
```

Since the excess kurtosis ¡ 0, it indicates that the data is platykurtic. However, note that the excess kurtosis for this sample is closer to 0 than s1.

```
s3 <- data.frame(x = rnorm(n = obs[3], mean = 0, sd = 1))
s3_stats <- s3 %>% summarize(Mean = mean(x), SD = sd(x), Skewness = skewness(x), "Excess Kurtos
s3_stats
```

```
##         Mean       SD   Skewness Excess Kurtosis
## 1 -0.1078957 1.001425 -0.1934592     -0.05901574
```

Interpretation:
Center: The sample mean of a normal distribution is the balancing point of the distribution. In this case, this will be equal to around -

```
s3_stats$Mean
```

```
## [1] -0.1078957
```

Spread: The average distance from the sample mean is -

```
s3_stats$SD
```

```
## [1] 1.001425
```

Skewness: The skewness of a dataset describes the symmetry of a distribution.

```
s3_stats$Skewness
```

```
## [1] -0.1934592
```

Since skewness is slightly greater than 0, there are more observations for low values and fewer with high values. However, note that the skewness for this sample is closer to 0 than s2.
Kurtosis describes how the peaked data is in relation to the Gaussian distribution (the bell-shaped curve).

```
s3_stats$`Excess Kurtosis`
```

```
## [1] -0.05901574
```

Since the excess kurtosis ¡ 0, it indicates that the data is platykurtic. However, note that the excess kurtosis for this sample is closer to 0 than s2.

```
s4 <- data.frame(x = rnorm(n = obs[4], mean = 0, sd = 1))
s4_stats <- s4 %>% summarize(Mean = mean(x), SD = sd(x), Skewness = skewness(x), "Excess Kurtos
```

Interpretation:
Center: The sample mean of a normal distribution is the balancing point of the distribution. In this case, this will be equal to around -

```
s4_stats$Mean
```

```
## [1] -0.0180137
```

Spread: The average distance from the sample mean is -

```
s4_stats$SD
```

```
## [1] 0.9640693
```

Skewness: The skewness of a dataset describes the symmetry of a distribution.

```
s4_stats$Skewness
```

```
## [1] 0.0620886
```

Since skewness is almost 0 (slightly greater), the distribution is almost symmetric.
Kurtosis describes how the peaked data is in relation to the Gaussian distribution (the bell-shaped curve).

```
s4_stats$`Excess Kurtosis`
```

```
## [1] 0.1202453
```

Since the excess kurtosis is almost 0 (slightly lesser), it indicates that the data is almost mesokurtic.
As the sample size increases, the distribution moves towards becoming more Normal - the mean tends to equal 0, the SD tends to equal 1, and both the skewness and excess kurtosis tend to equal 0.

(c) Generate a random sample of size $n = 10$ for your two sets of parameter(s).

```
dat1 <- data.frame(x = rnorm(n = obs[1], mean = 0, sd = 1))
```

Calculate the method of moments estimator(s) and maximum likelihood estimator(s).

```
library(nleqslv)
#########################################################
# MOM estimator
#########################################################
norm.mom<-function(par, data){
  mu <- par[1]
  sigma <- par[2]

  EX1 <- mu               # Expected value of a normal distribution
  EX2 <- sigma            # Variance of a normal distribution

  xbar1 <- mean(data)
  xbar2 <- mean(data^2)

  c(EX1-xbar1, EX2-xbar2)
}

# Entering the starting guess, the function(s) we want to solve for c(0, 0),
# and the dataframe a arguments of the non-linear equation solver
mom1 <- nleqslv(x = c(0,1),fn = norm.mom, data = dat1$x)

#########################################################
# MLE
#########################################################
norm.ll<-function(par, data, neg=T){
  mu <- par[1]
  sigma <- par[2]
  ll <- sum(dnorm(x=data, mean=mu, sd=sigma, log = T))
```

```
  ifelse(neg, -ll, ll)
  # Since the optim() function minimizes, we use neg because on multiplying by negative,     i
}

mle1 <- optim(par = c(0,1), fn = norm.ll, data=dat1$x)
```

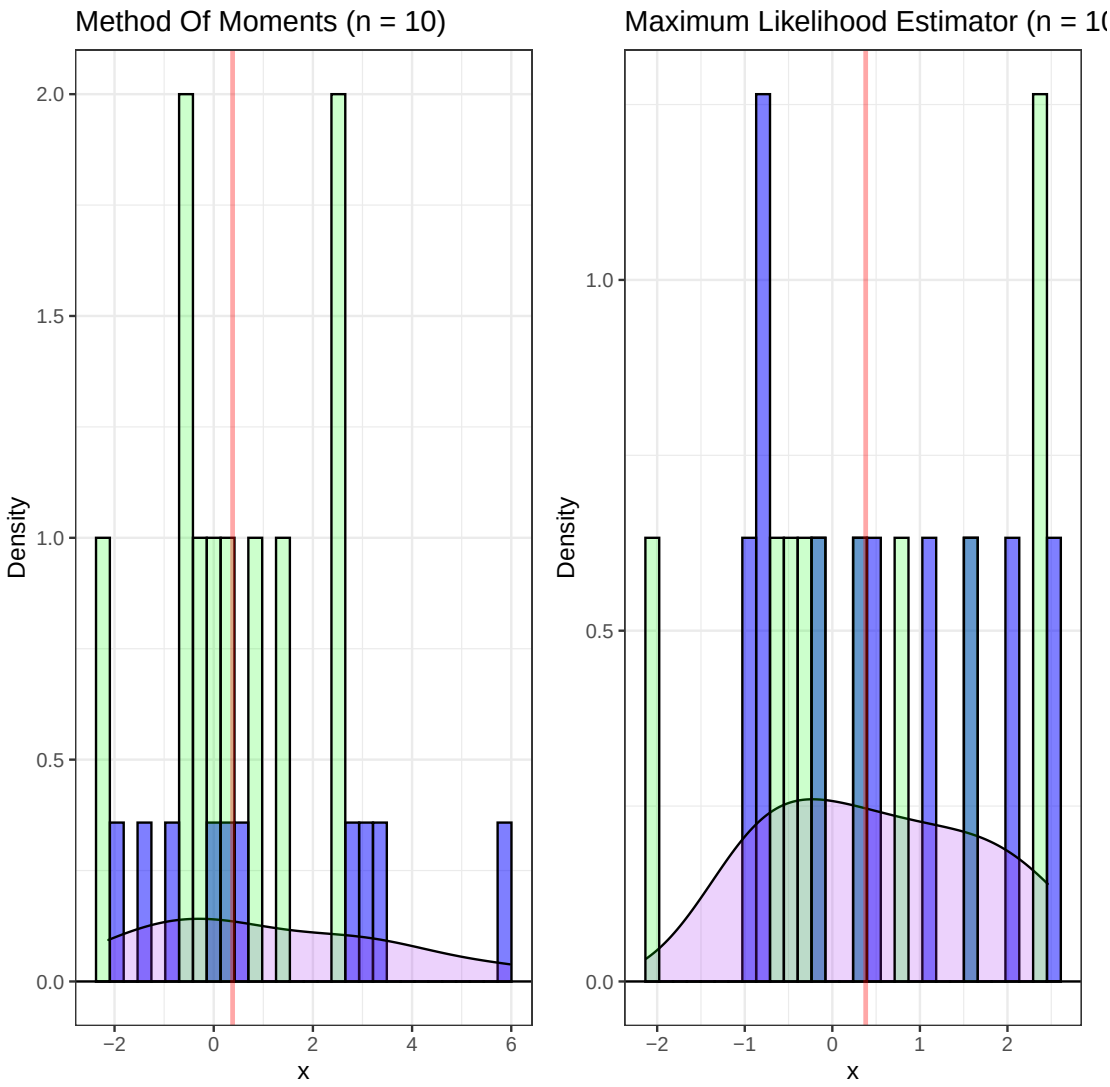
In a 1×2 grid, plot a histogram of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.

```
mom_dat1 <- data.frame(x = rnorm(n = obs[1], mean = mom1$x[1], sd = mom1$x[2]))
mom1_p <- ggplot(data=mom_dat1, aes(x=x))+
  geom_histogram(color="black", fill = "blue", alpha = 0.5, aes(y=..density..))+
  geom_hline(yintercept=0)+ geom_density(alpha=.2, fill="purple") +
  theme_bw()+ xlab("x")+ ylab("Density")+
  ggtitle("Method Of Moments (n = 10)") +
  geom_histogram(aes(x = dat1$x), color="black", fill = "green", alpha = 0.2) +
  geom_vline(xintercept = mean(dat1$x), alpha=0.35, color="red",size=1)

mle_dat1 <- data.frame(x = rnorm(n = obs[1], mean = mle1$par[1], sd = mle1$par[2]))
mle1_p <- ggplot(data=mle_dat1, aes(x=x))+ geom_histogram(color="black", fill = "blue", alpha =
  geom_hline(yintercept=0)+ geom_density(alpha=.2, fill="purple") +
  theme_bw()+ xlab("x")+ ylab("Density")+
  ggtitle("Maximum Likelihood Estimator (n = 10)")  +
  geom_histogram(aes(x = dat1$x, y=..density..), color="black", fill = "green", alpha = 0.2) +
  geom_vline(xintercept = mean(dat1$x), alpha=0.35, color="red",size=1)

mom1_p + mle1_p
```

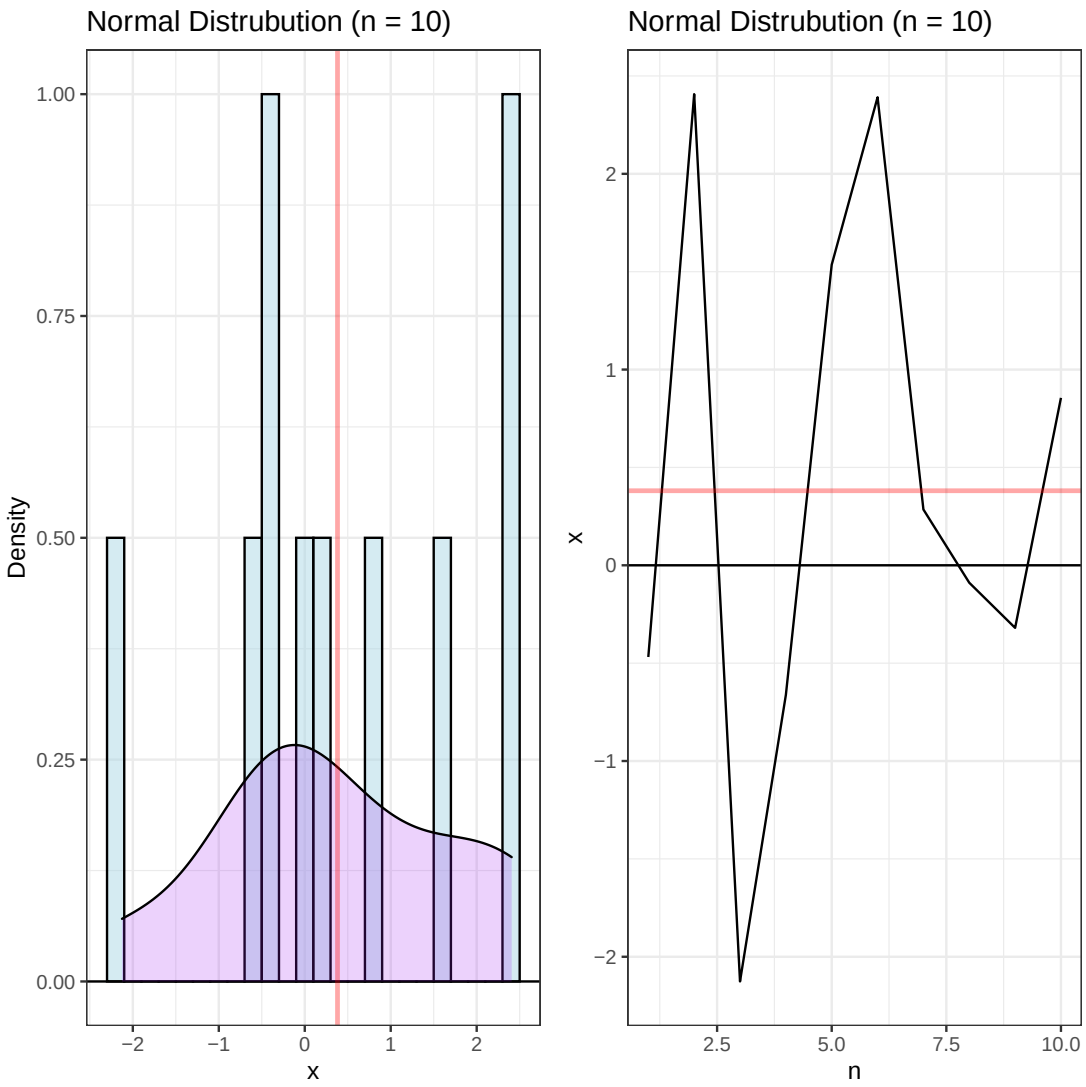
```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```




```
dat1_h <- ggplot(data=dat1, aes(x=x))+ geom_histogram(color="black", fill = "lightblue", alpha

dat1<-dat1 %>% mutate(n = seq(1,obs[1],1))
dat1_p <- ggplot(data=dat1, aes(x=n, y=x))+ geom_line()+ geom_hline(yintercept=0)+ geom_hline(

dat1_h + dat1_p
```



(d) Generate a random sample of size $n = 25$ for your two sets of parameter(s).

```
dat2 <- data.frame(x = rnorm(n = obs[2], mean = 0, sd = 1))
```

Calculate the method of moments estimator(s) and maximum likelihood estimator(s).

```
mom2 <- nleqslv(x = c(0,1),fn = norm.mom, data = dat2$x)
mle2 <- optim(par = c(0,1), fn = norm.ll, data=dat2$x)
```

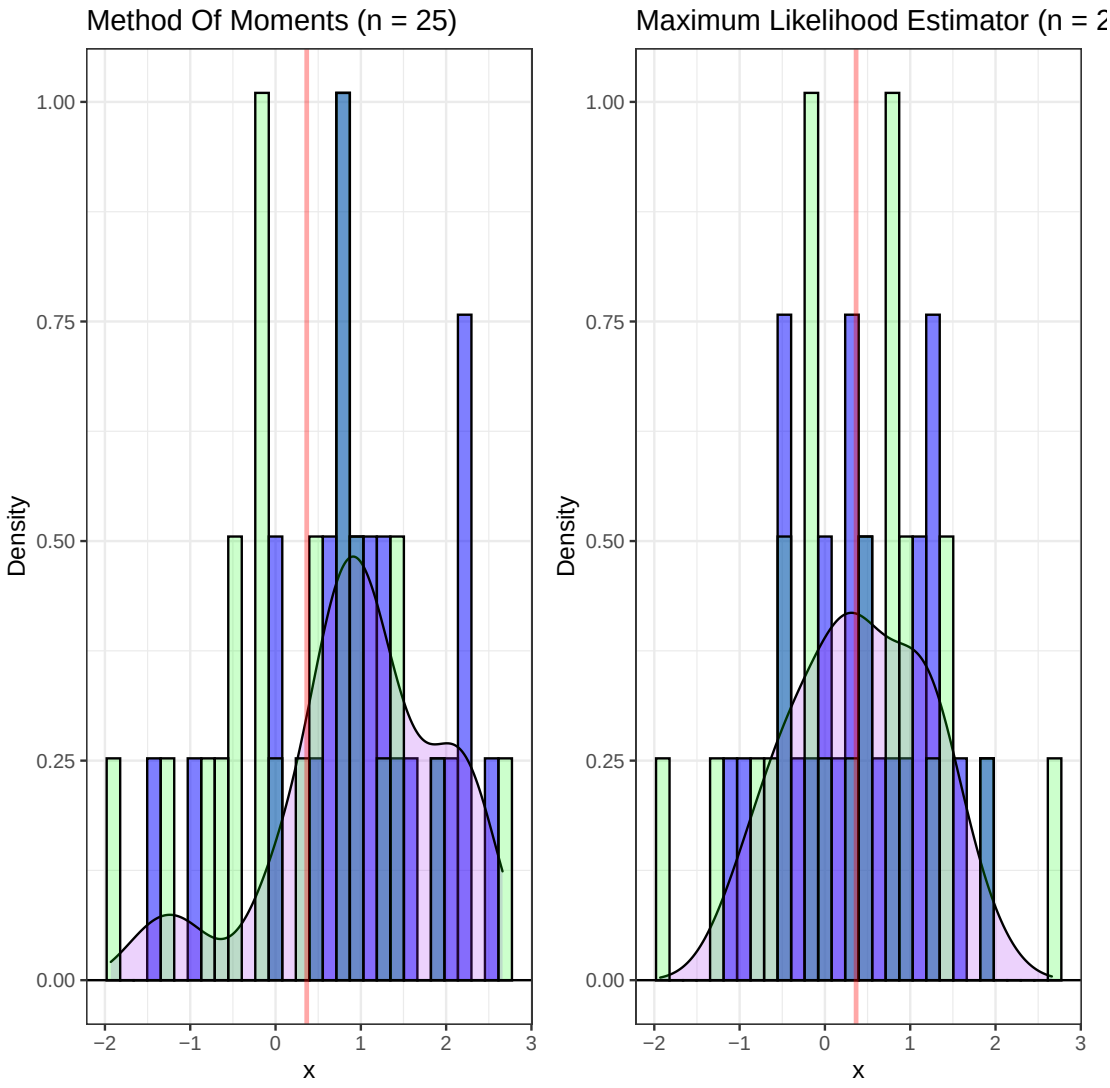
In a 1×2 grid, plot a histogram of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.

```
mom_dat2 <- data.frame(x = rnorm(n = obs[2], mean = mom2$x[1], sd = mom2$x[2]))
mom2_p <- ggplot(data=mom_dat2, aes(x=x))+ geom_histogram(color="black", fill = "blue", alpha =

mle_dat2 <- data.frame(x = rnorm(n = obs[2], mean = mle2$par[1], sd = mle2$par[2]))
mle2_p <- ggplot(data=mle_dat2, aes(x=x))+ geom_histogram(color="black", fill = "blue", alpha =
```

```
mom2_p + mle2_p
```

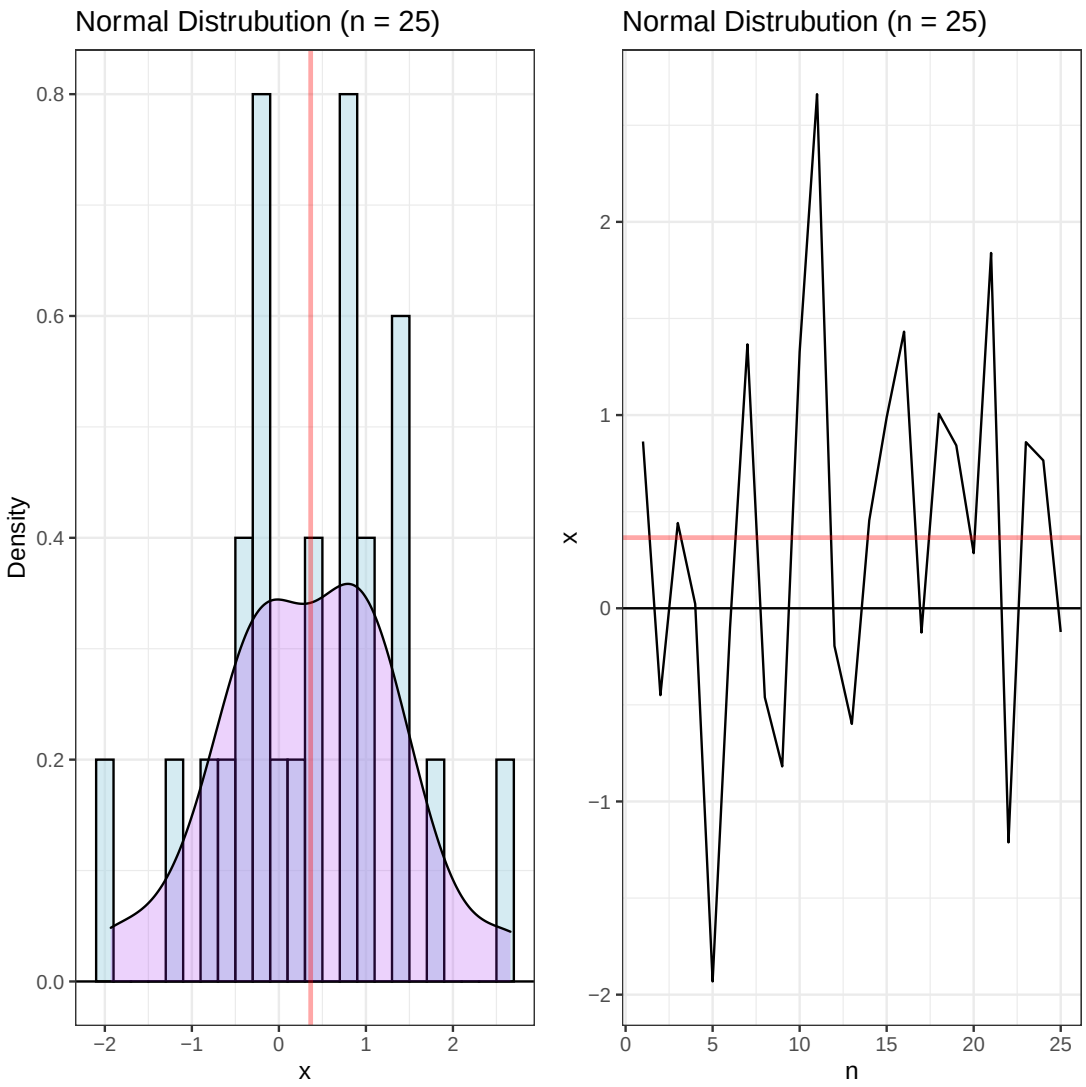
```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```




```
dat2_h <- ggplot(data=dat2, aes(x=x))+ geom_histogram(color="black", fill = "lightblue", alpha

dat2 <-dat2 %>% mutate(n = seq(1,obs[2],1))
dat2_p <- ggplot(data=dat2, aes(x=n, y=x))+ geom_line()+ geom_hline(yintercept=0)+ geom_hline(y

dat2_h + dat2_p
```



(e) Generate a random sample of size $n = 100$ for your two sets of parameter(s).

```
dat3 <- data.frame(x = rnorm(n = obs[3], mean = 0, sd = 1))
```

Calculate the method of moments estimator(s) and maximum likelihood estimator(s).

```
mom3 <- nleqslv(x = c(0,1),fn = norm.mom, data = dat3$x)
mle3 <- optim(par = c(0,1), fn = norm.ll, data=dat3$x)
```

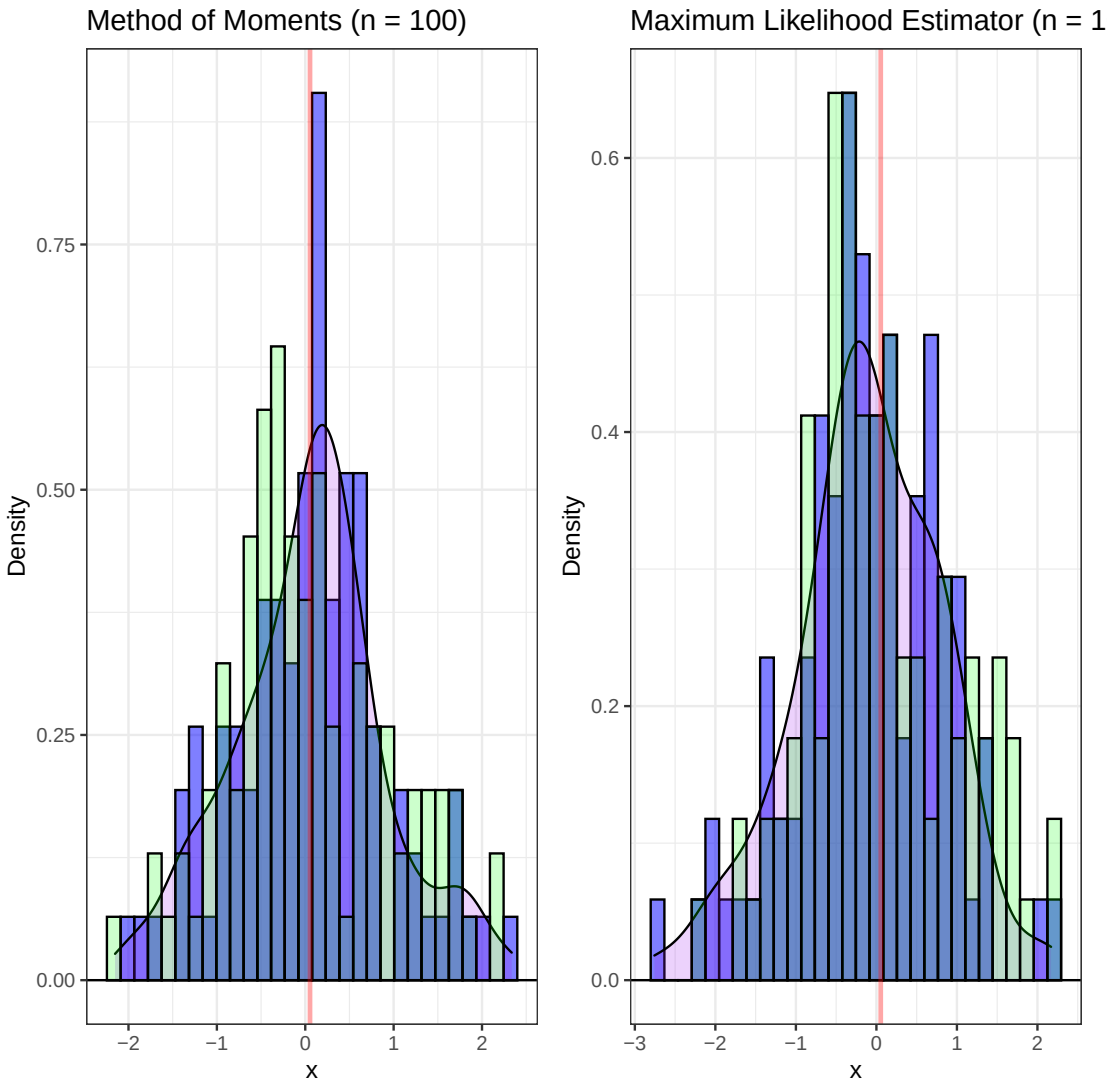
In a 1×2 grid, plot a histogram of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.

```
mom_dat3 <- data.frame(x = rnorm(n = obs[3], mean = mom3$x[1], sd = mom3$x[2]))
mom3_p <- ggplot(data=mom_dat3, aes(x=x))+ geom_histogram(color="black", fill = "blue", alpha =

mle_dat3 <- data.frame(x = rnorm(n = obs[3], mean = mle3$par[1], sd = mle3$par[2]))
mle3_p <- ggplot(data=mle_dat3, aes(x=x))+ geom_histogram(color="black", fill = "blue", alpha =
```

```
mom3_p + mle3_p
```

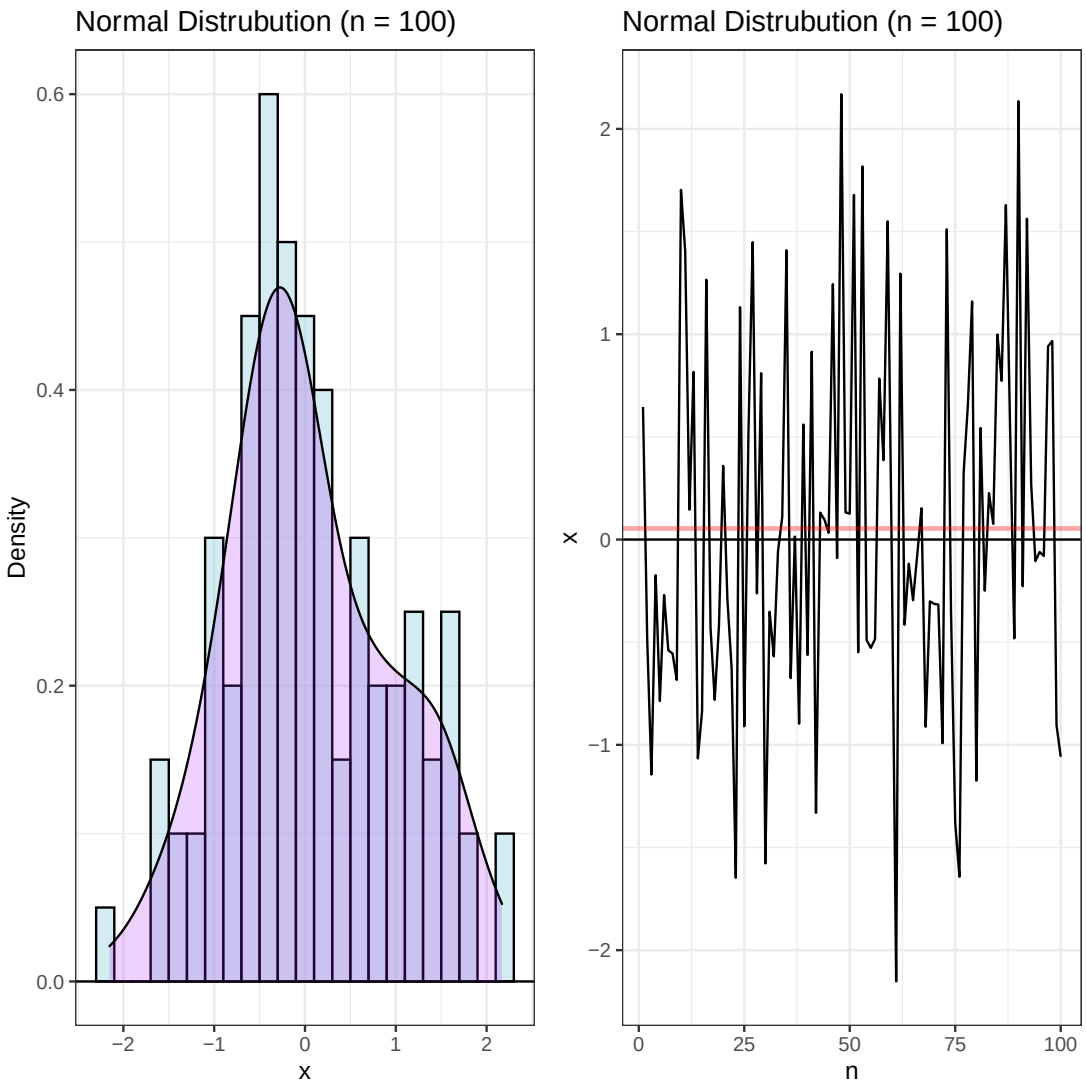
```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```




```
dat3_h <- ggplot(data=dat3, aes(x=x))+ geom_histogram(color="black", fill = "lightblue", alpha

dat3 <-dat3 %>% mutate(n = seq(1,obs[3],1))
dat3_p <- ggplot(data=dat3, aes(x=n, y=x))+ geom_line()+ geom_hline(yintercept=0)+ geom_hline(y

dat3_h + dat3_p
```



(f) Generate a random sample of size $n = 100$ for your two sets of parameter(s).

```
dat4 <- data.frame(x = rnorm(n = obs[4], mean = 0, sd = 1))
```

Calculate the method of moments estimator(s) and maximum likelihood estimator(s).

```
mom4 <- nleqslv(x = c(0,1),fn = norm.mom, data = dat4$x)
mle4 <- optim(par = c(0,1), fn = norm.ll, data=dat4$x)
```

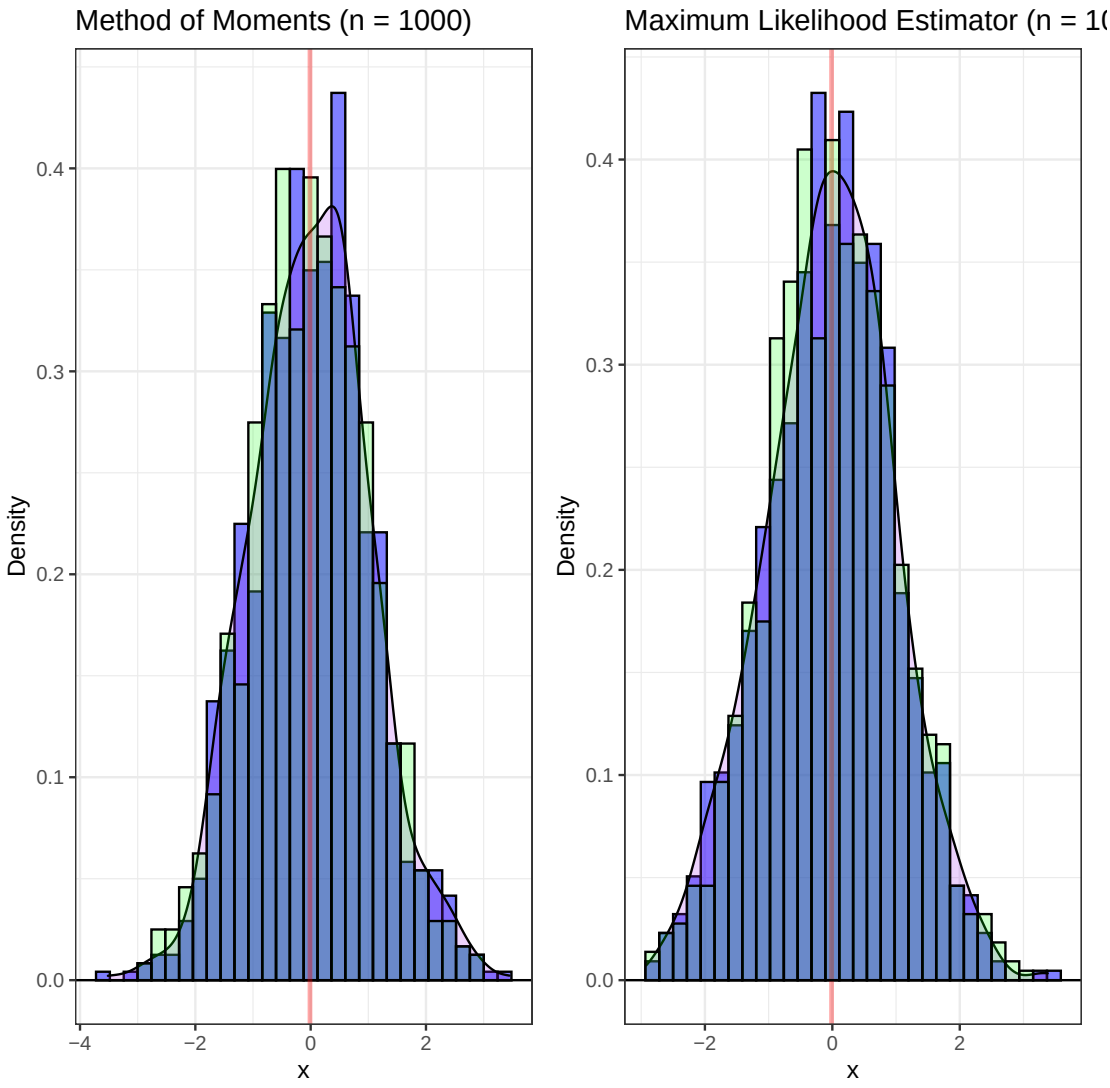
In a 1×2 grid, plot a histogram of each set of data with (1) the method of moments estimated distribution, (2) the maximum likelihood estimated distribution, and superimpose the true distribution in both.

```
mom_dat4 <- data.frame(x = rnorm(n = obs[4], mean = mom4$x[1], sd = mom4$x[2]))
mom4_p <- ggplot(data=mom_dat4, aes(x=x))+ geom_histogram(color="black", fill = "blue", alpha =

mle_dat4 <- data.frame(x = rnorm(n = obs[4], mean = mle4$par[1], sd = mle4$par[2]))
mle4_p <- ggplot(data=mle_dat4, aes(x=x))+ geom_histogram(color="black", fill = "blue", alpha =
```

```
mom4_p + mle4_p
```

```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```

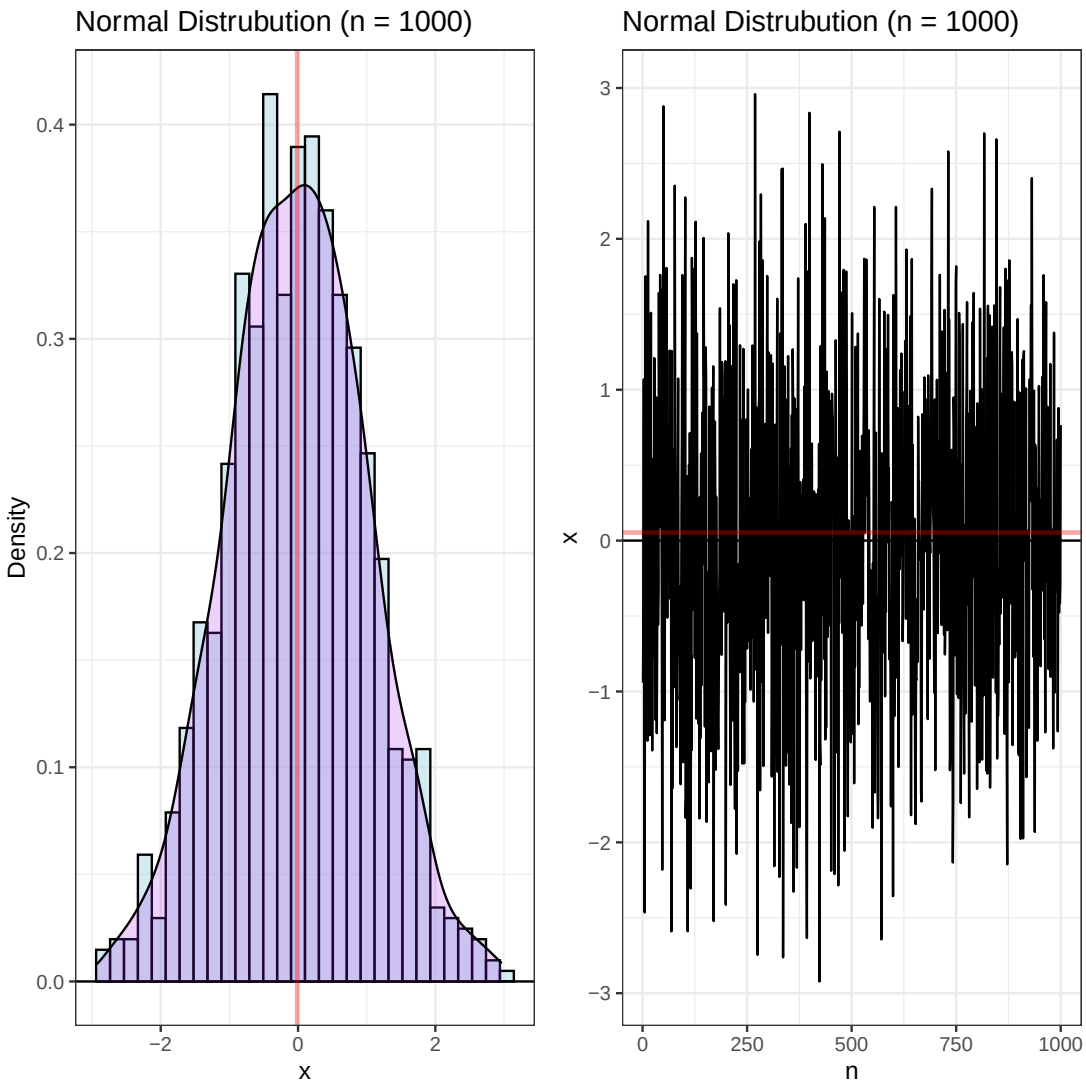



```
dat4_h <- ggplot(data=dat4, aes(x=x))+ geom_histogram(color="black", fill = "lightblue", alpha

dat4 <-dat4 %>% mutate(n = seq(1,obs[4],1))
dat4_p <- ggplot(data=dat4, aes(x=n, y=x))+ geom_line()+ geom_hline(yintercept=0)+ geom_hline(y

dat4_h + dat4_p
```

```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



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

Parts (c)-(f) illustrate the weak law of large numbers. As the sample size n increases, we see that both our computed estimators (MOM and MLE) tend to overlap more with their corresponding probability distributions - there is very less overlap when n = 10; however, at n = 1000, both the histograms are essentially superimposed on top of one another. Thus, as the sample size increases, our estimators do a better job at estimating the population statistics of the Normal distribution. We also notice that as the sample size increases, the sample mean gets closer to 0.

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.

   (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 (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.

   (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.

   (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.

   (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.

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

References

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