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.

   Solution:
   Our group has chosen to use a normal distribution for the `Q1` and `Q2`!
   To start with, I will define what a normal distribution is. A continuous variable X has a normal distribution — with mean μ and σ^2 — X N(μ, σ^2), if it has the following properties:

$$\mu \in \mathbb{R}; \sigma \in \mathbb{R}^+ \qquad (\textbf{Parameters})$$

$$X = \{x : x \in \mathbb{R}\} \qquad (\textbf{Support})$$

$$f_X(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (\textbf{PDF})$$

$$F_X(x|\mu,\sigma) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (\textbf{CDF})$$

   Description: Normal distribution is one of the most important distributions in the world of statistics, and it's used in many fields. The graph of this probability function is a symmetric, bell-shaped curved. Symmetry of this function implies that its mean is going to be equal to its median value. We are going to prove it later on.
   History: One of the first major great discoveries in the field of statistics happened in 1713, as Jacob Bernoulli published his work on proving the Weak Law of Large Numbers. According to Patel and Read (1996), initially normal distribution appeared in 1733 as an approximation to the probability for sums of binominally distributed quantities to lie between two values.
   Today: According to Ahsanullah et al. (2014), normal distribution plays an important role in many applied problems in biology, economics, engineering, genetics, hydrology, mechanics, medicine, number theory, statistics, physics, psychology and so on.

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

   Solution:
   Here's the equation of PDF for normal distribution!

$$f_X(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (\textbf{PDF})$$

   Now, let's prove that it's valid!
   Since normal (or gaussian — whatever you prefer!) distribution is a continous probability distribution, it implies that the area under the curve is equal to 1 (or 100%!). This statement takes its roots from the second Kolmogorov axiom that states that the entirety of sample space is equal to one.
   Therefore, let's use CDF to prove that our PDF formula is going to return one! Since the support

for normal distribution contain all rational numbers, our integral is going to be from negative infinity to positive infinity.

$$F_X(x|\mu,\sigma) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{\textbf{CDF}}$$

I am going to use `integrate()` function in order to compute this equation.

```
mean<-1 #mu
sd<-3 #sigma
pdf <- function(x){
  (1/(sd*sqrt(2*pi)))*exp((-(x-mean)^2)/(2*sd^2)) #our PDF
}
integrate(pdf, -Inf, Inf) #CDF for total area under the curve

## 1 with absolute error < 2.1e-07
```

(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.

Solution:
As we have established in part `(a)` of this problem, one of the key features of the normal distribution is the fact that it's symmetrical. Let's set out to prove it through the direct proof!
Let `m` be median!

$$P(X \leq m) = P(X \geq m) = \frac{1}{2} \tag{\textbf{Definition of the median}}$$

Let normal distribution be symmetric. If it's symmetric, then the statement $\mu = m$ holds true. Therefore, the following equation should also hold true:

$$\int_{-\infty}^{\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \int_{\mu}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{2} \tag{\textbf{Assumption}}$$

Let's check it through R!

```
        mean<-1 #mu
        sd<-3 #sigma
        func <- function(x){
          (1/(sd*sqrt(2*pi)))*exp((-(x-mean)^2)/(2*sd^2))
        }
        firstPart <- integrate(func, -Inf, mean)
        secondPart <- integrate(func, mean, Inf)

        (firstPart$value)

## [1] 0.5

        (firstPart$value==secondPart$value)

## [1] TRUE
```

It would appear that $\mu = m$! Therefore, mean is equal to median within the normal distribution. Therefore, since $\mu_1 = 1$ and $\mu_2 = -2$, the medians are 1 and -2 respectively!

(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?

Solution:
Let's take various values and plot different PDFs! I am going to use the ggplot2 (Wickham, 2016) library for it!

```
library(ggplot2)
plot.df <- data.frame(
  x=seq(-15, 15, 0.001),
  f1=dnorm(x=seq(-15, 15, 0.001), mean=1, sd=1),
  f2=dnorm(x=seq(-15, 15, 0.001), mean=-2, sd=4),
  f3=dnorm(x=seq(-15, 15, 0.001), mean=0, sd=5),
  f4=dnorm(x=seq(-15, 15, 0.001), mean=3, sd=3)
)

ggplot(plot.df, aes(x=x))+
  geom_line(aes(y=f1, color="m=1, sd=1"))+
  geom_line(aes(y=f2, color="m=-2, sd=4"))+
  geom_line(aes(y=f3, color="m=0, sd=5"))+
  geom_line(aes(y=f4, color="m=3, sd=3"))+
  theme_bw()
```

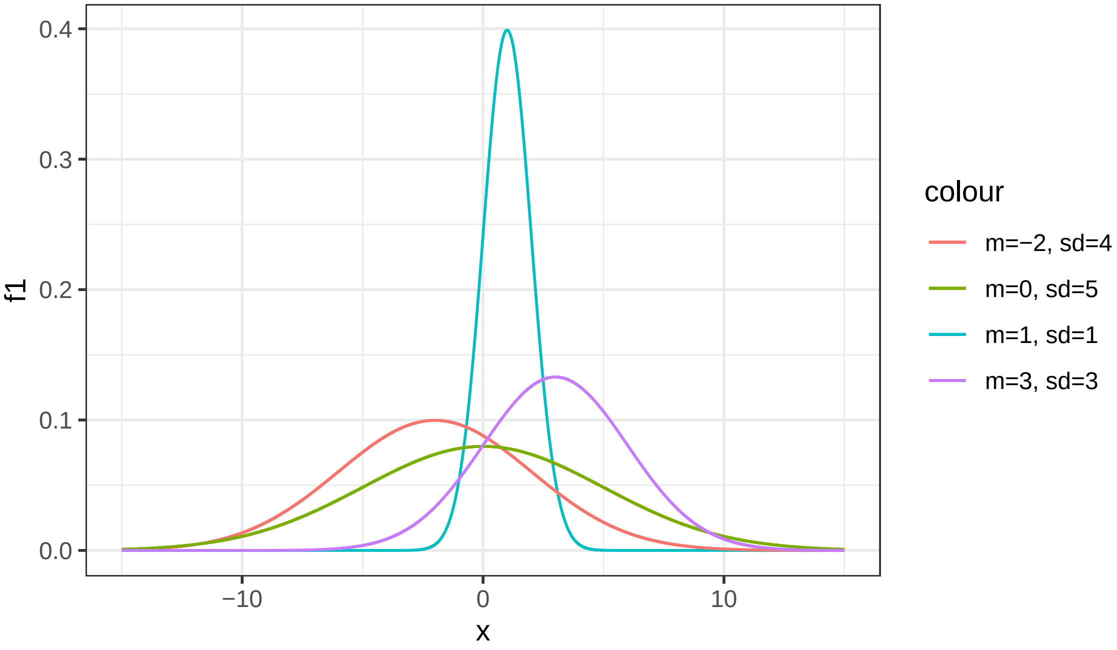

Figure 1: Gaussian PDF with various sets of parameters

As we can see on Figure 1, changing our parameters does indeed change the graph we have. Changing the μ transforms graph to the right (if the mean is positive) and to the left (if the mean is negative). Changing σ changes the kurtosis of the graph. Thus, distributions with lower value of σ show higher peaks.

(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.

Solution:

```
plot.df <- data.frame(
  x=seq(-15, 15, 0.001),
  f1=pnorm(q=seq(-15, 15, 0.001), mean=1, sd=1),
  f2=pnorm(q=seq(-15, 15, 0.001), mean=-2, sd=4),
  f3=pnorm(q=seq(-15, 15, 0.001), mean=0, sd=5),
  f4=pnorm(q=seq(-15, 15, 0.001), mean=3, sd=3)
)

ggplot(plot.df, aes(x=x))+
  geom_line(aes(y=f1, color="m=1, sd=1"))+
  geom_line(aes(y=f2, color="m=-2, sd=4"))+
  geom_line(aes(y=f3, color="m=0, sd=5"))+
  geom_line(aes(y=f4, color="m=3, sd=3"))+
  theme_bw()
```

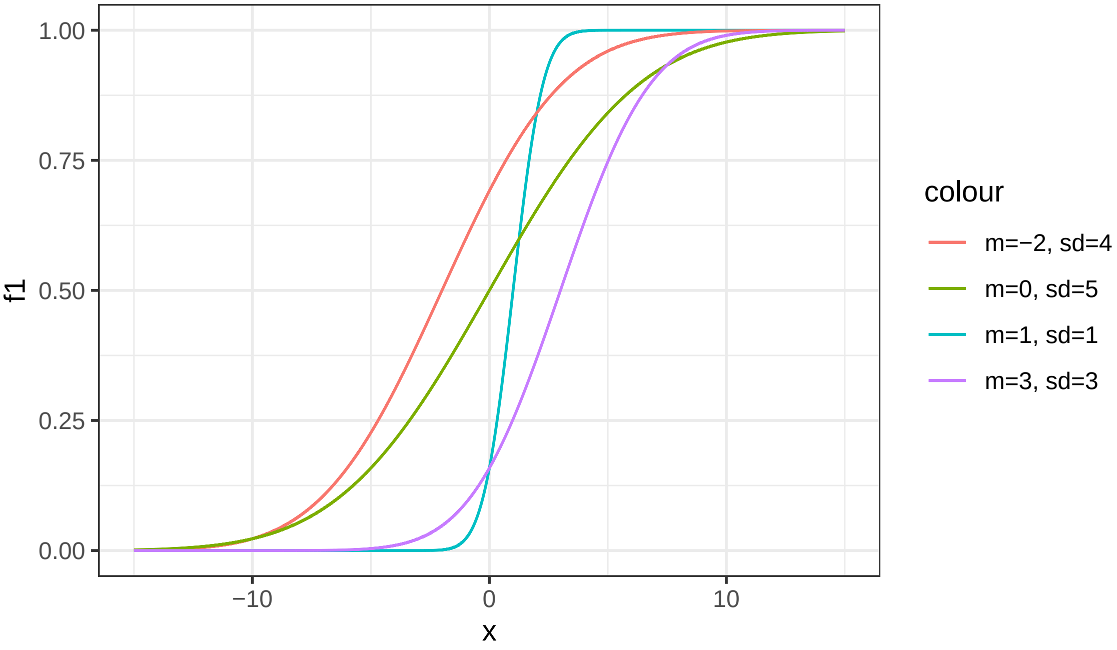

Figure 2: Gaussian CDF with various sets of parameters

We see that Figure 2 is a graph of CDF function with different sets of parameters. This function shows the probability that a probability that a random variable X will take a value less than or equal to x. We can see that our PDF is valid, for when we went over all values of X, the probability is 1.00 or 100%. Changing the value of μ changes the position of a graph (moves it to the right or to the left), while the change in σ increases the slope of the function: the higher the value of σ, the faster the function will go over all the values and reach 100%.

(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.

Solution:

```r
library(patchwork)
library(tidyverse)

sample.df <- list(x1=rnorm(10, mean=1, sd=1),
                       x2=rnorm(25, mean=1, sd=1),
                       x3=rnorm(100, mean=1, sd=1),
                       x4=rnorm(1000, mean=1, sd=1),
                       y1=rnorm(10, mean=-2, sd=4),
                       y2=rnorm(25, mean=-2, sd=4),
                       y3=rnorm(100, mean=-2, sd=4),
                       y4=rnorm(1000, mean=-2, sd=4))

buildingPlot <- function(source, mu, sigma){
  df <- data.frame(value=source) #turning values from the list into a df
  colnames(df) <- "value" #changing the name of the column

  answer<-ggplot(df, aes(value))+
    geom_histogram(aes(y=..density..), bins=10,
                 color="black")+ #building a histogram
    geom_function(fun=dnorm, args = list(mean = mu, sd = sigma),
        color="red")+
    theme_bw() +#superimposing the function
    labs(x="Value", y="Density")
  answer
}

x1<-buildingPlot(sample.df[1], 1, 1)+labs(title="Sample=10",
                                          subtitle = "mean=1, sd=1")
x2<-buildingPlot(sample.df[2], 1, 1)+labs(title="Sample=25",
                                          subtitle = "mean=1, sd=1")
x3<-buildingPlot(sample.df[3], 1, 1)+labs(title="Sample=100",
                                          subtitle = "mean=1, sd=1")
x4<-buildingPlot(sample.df[4], 1, 1)+labs(title="Sample=1000",
                                          subtitle = "mean=1, sd=1")

y1<-buildingPlot(sample.df[5], -2, 4)+labs(title="Sample=10",
                                          subtitle = "mean=-2, sd=4")
y2<-buildingPlot(sample.df[6], -2, 4)+labs(title="Sample=25",
                                          subtitle = "mean=-2, sd=4")
y3<-buildingPlot(sample.df[7], -2, 4)+labs(title="Sample=100",
                                          subtitle = "mean=-2, sd=4")
y4<-buildingPlot(sample.df[8], -2, 4)+labs(title="Sample=1000",
                                          subtitle = "mean=-2, sd=4")
#(x1|x2|x3|x4)/(y1|y2|y3|y4)
(x1+y1)/(x2+y2)/(x3+y3)/(x4+y4)
```

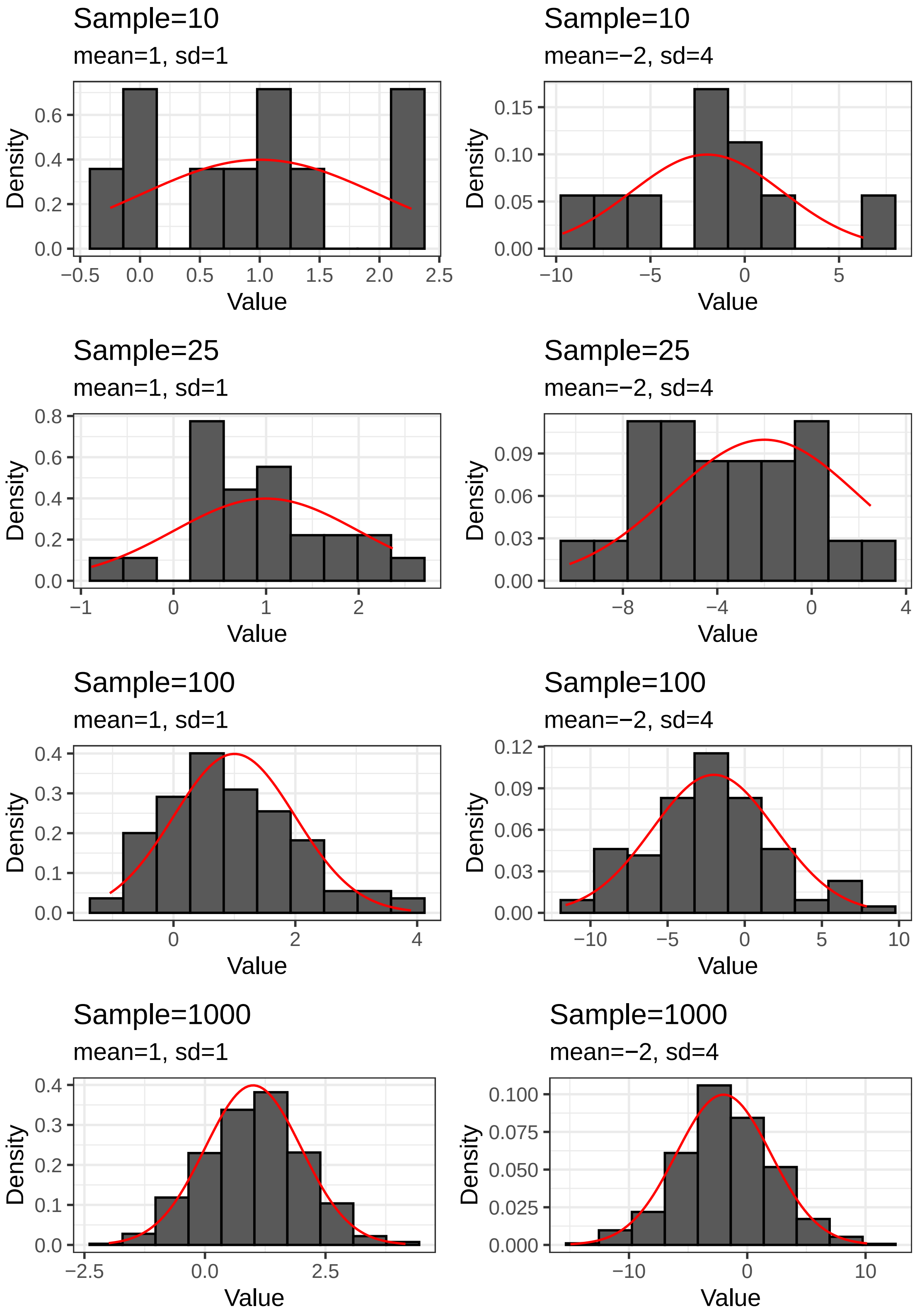

Figure 3: Histograms of each set of data with superimposed true density function

As we can see, the bigger sample we have, the closer the resulting histogram is to the real density function that we built on our set of parameters. We can notice that it's particularly true for $n > 30$.

i. Our PDF is such a good fit for the data because we are using `rnorm()` function that randomly generates numbers for the normal distribution.

ii. A good fit can also be explained via Central Limit Theorem: as our sample enlarges, the distribution of our random variable follows a normal distribution.

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.

   (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

Ahsanullah, M., Kibria, B. G., and Shakil, M. (2014). Normal distribution. In *Normal and Student st Distributions and Their Applications*, pages 7–50. Springer.

Patel, J. K. and Read, C. B. (1996). *Handbook of the normal distribution*, volume 150. CRC Press.

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

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.