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. **Solution:**

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.

   Solution: The discrete distribution we have chosen is the Bernoulli distribution. The Bernoulli distribution is a discrete probability distribution for a Bernoulli trial - a probabilistic experiment that can have one of two outcomes, success (x = 1) and failure (x = 0), and in which the probability fo success is p. Often p is called the Bernoulli probability parameter (Forbes et al., 2011). In Bernoulli distribution, the random variable X can have only one of two values: 0 or 1. This means that the support of our discrete random variable is the set $\{0, 1\}$. This distribution can be summarized as follows:

$$\begin{aligned} p &\in (0,1) && \textbf{[Parameter]} \\ \mathcal{X} &= \{x : x \in \{0,1\}\} && \textbf{[Support]} \\ f_X(x \mid p) &= p^x(1-p)^{1-x} I(x \in \{0,1\}) && \textbf{[PMF]} \\ F_X(x \mid p) &= P(X \leq \lfloor x \rfloor) \\ &= [(1-p) I(\lfloor x \rfloor = 0)] + I(\lfloor x \rfloor \geq 1) && \textbf{[CDF]} \end{aligned}$$

   Simply, the Bernoulli distribution can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes-no question (Vukadin et al., 2021). This distribution is named after the 17th century Swiss mathematician Jacob Bernoulli, because he was the one who explicitly defined the concept of Bernoulli trial (in his book *Ars Conjectandi*) described above. Bernoulli distribution serves as a building block for discrete distributions which model Bernoulli trials, such as Binomial distribution and geometric distribution (Cthaeh, 2020). Logistic regression, a widely used classification model that models a binary outcome, also takes advantage of the Bernoulli distribution (Kleinbaum and Klein, 2010).

   Since the Bernoulli distribution is not cataloged by R, we have to introduce it into our calculations with the following functions:

```r
# Bernoulli PMF
dbern<-function(x,prob){
  if(prob<0 | prob>1){
    errormsg <- "This function is only valid for success probabilities between 0 and 1."
    stop(errormsg)
  }
  indicator <- rep(0, length(x))
  indicator[x==0] <- 1 # indicator should be one if x=0
  indicator[x==1] <- 1 # indicator should be one if x=1
  fx <- (prob^x * (1-prob)^(1-x)) * indicator # PMF formula
  return(fx)
}
# Bernoulli CDF
pbern<-function(q, prob){
  if(prob<0 | prob>1){
    errormsg<-"This function is only valid for success probabilities between 0 and 1."
    stop(errormsg)
  }
  indicator1 <- rep(1, length(q))
  indicator1[q != 0] <- 0 #indicator should be zero if x!=0
  indicator2 <- rep(1, length(q))
  indicator2[q < 1] <- 0 #indicator should be zero if x<1
  Fx <- (1-prob) * indicator1 + indicator2
  return(Fx)
}
```

   The R packages Tidyverse (Wickham et al., 2019) and Patchwork (Pedersen, 2020) will be used in all of the following plots in question 3.

(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.

Solution: For a PMF to be valid, it has to fulfill the following statements:

1. $0 \leq f_X(x) \leq 1$ for all $x \in \mathbb{R}$
2. $\sum_{-\infty}^{\infty} f_X = \sum_{\mathcal{X}} f_X = 1$

By definition, the Bernoulli distribution satisfies statement 1, since the support is 0,1 and the probability parameter $p \in (0, 1)$, so the term for PMF (see above) cannot have a value smaller than 0 or greater than 1.

We can also show that the statement 2 is true for the Bernoulli distribution:

$$\begin{aligned}\sum_{\mathcal{X}} f_X = \sum_{x=0}^{1} f_X(x) &= \sum_{x=0}^{1} p^x(1-p)^{1-x} \\ &= p^0(1-p)^{1-0} + p^1(1-p)^{1-1} \\ &= (1)(1-p) + p(1) \\ &= 1 - p + p \\ &= 1\end{aligned}$$

Therefore, we have a valid PMF for our distribution.

(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.

Solution: In the Bernoulli distribution, where p is the probability parameter and $q = 1 - p$, when p>q, there are more "successes" (that is when X is 1) than failure (when X is 0), so the median must also be a success. Similar, when q>p, there are more "failures" than "successes", hence resulting in the median being 0. When $p = q$, there is an equal probability of X being 0 or 1, thus the median is ambiguous. Therefore, the median for when $p = 0.4$ is zero and the median for when $p = 0.6$ is 1. Further inquiry **(QUOTE: Wikipedia)** suggests that this really is the case for the median of the Bernoulli distribution:

$$Median = \begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$$

(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?

Solution: We can plot the PMF of the Bernoulli distribution for different parameters.Note that the initially chosen parameters were $p = 0.4$ and $p = 0.6$. Let us also look at the PMF when the parameter values are 0.5, and 0.8.

```
library(tidyverse)
library(patchwork)
plotbernPMF <- function(prob){ # Pass in the success probability
  ggdat <- data.frame(x = (-1:2),
                      f = dbern(x = (-1:2), prob = prob),
                      F = pbern(q = (-1:2), prob = prob))
  ## Plot PMF
  PMF <- ggplot(data = ggdat, aes(x = x)) +
    geom_linerange(aes(ymax = f), ymin = 0) +
    geom_hline(yintercept = 0) +
    theme_bw() +
    ylim(0, 1) +
    xlab("X") +
    ylab(bquote(f[x](x))) +
    ggtitle("Bernoulli Distribution",subtitle = paste("p =", prob))

  return(PMF)
}

plotbernPMF(0.4) + plotbernPMF(0.5) + plotbernPMF(0.6) + plotbernPMF(0.8)
```

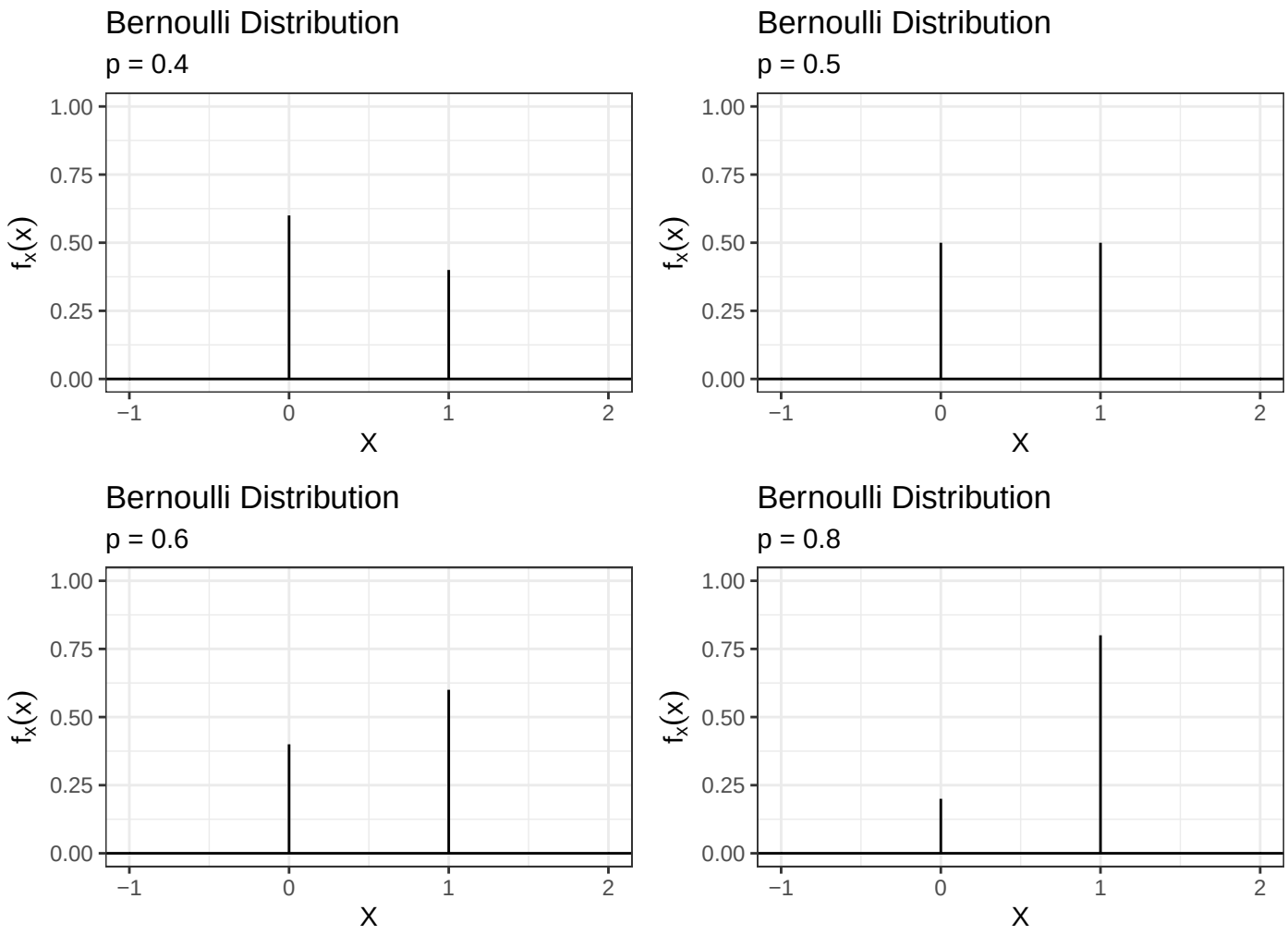

Figure 1: The CDF of the Bernoulli distribution for different probability parameters p

We observe the same characteristic of the Bernoulli distribution in Figure 1 as we discussed about its median above. If the $p < 0.5$ then there are more failures than successes, indicated by the taller PMF at 0. On the other hand, if If the $p > 0.5$ then there are more successes than failures, which is reflected in the PMF plot being taller at 1. The plot of the PMF when $p = 0.5$ suggests that there is an equal chance for success as there is for failure.

(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.

Solution: We can plot the CDF of the Bernoulli distribution for the different parameters 0.4, 0.5, 0.6, and 0.8 below:

```
plotbernCDF <- function(prob){ # Pass in the success probability
  ggdat <- data.frame(x = (-1:2),
                      f = dbern(x = (-1:2), prob = prob),
                      F = pbern(q = (-1:2), prob = prob))
  ggdat.openpoints <- data.frame(x = ggdat$x,
                                 y = pbern(ggdat$x-1, prob = prob))
  ggdat.closedpoints <- data.frame(x = ggdat$x,
                                   y = pbern(ggdat$x,prob=prob))
  CDF<-ggplot(data = ggdat, aes(x = x, y = F)) +
    geom_step()+
    geom_point(data = ggdat.openpoints, aes(x = x, y = y), shape = 1) +
    geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
    geom_hline(yintercept = 0.5, linetype="dotted", color="red")+
```

```
    theme_bw()+
    xlab("X")+
    ylab(bquote(F[x](x)))+
    ggtitle("Bernoulli CDF",subtitle=(paste("p =", prob)))
  return(CDF)
}
plotbernCDF(0.4) + plotbernCDF(0.5) + plotbernCDF(0.6) + plotbernCDF(0.8)
```

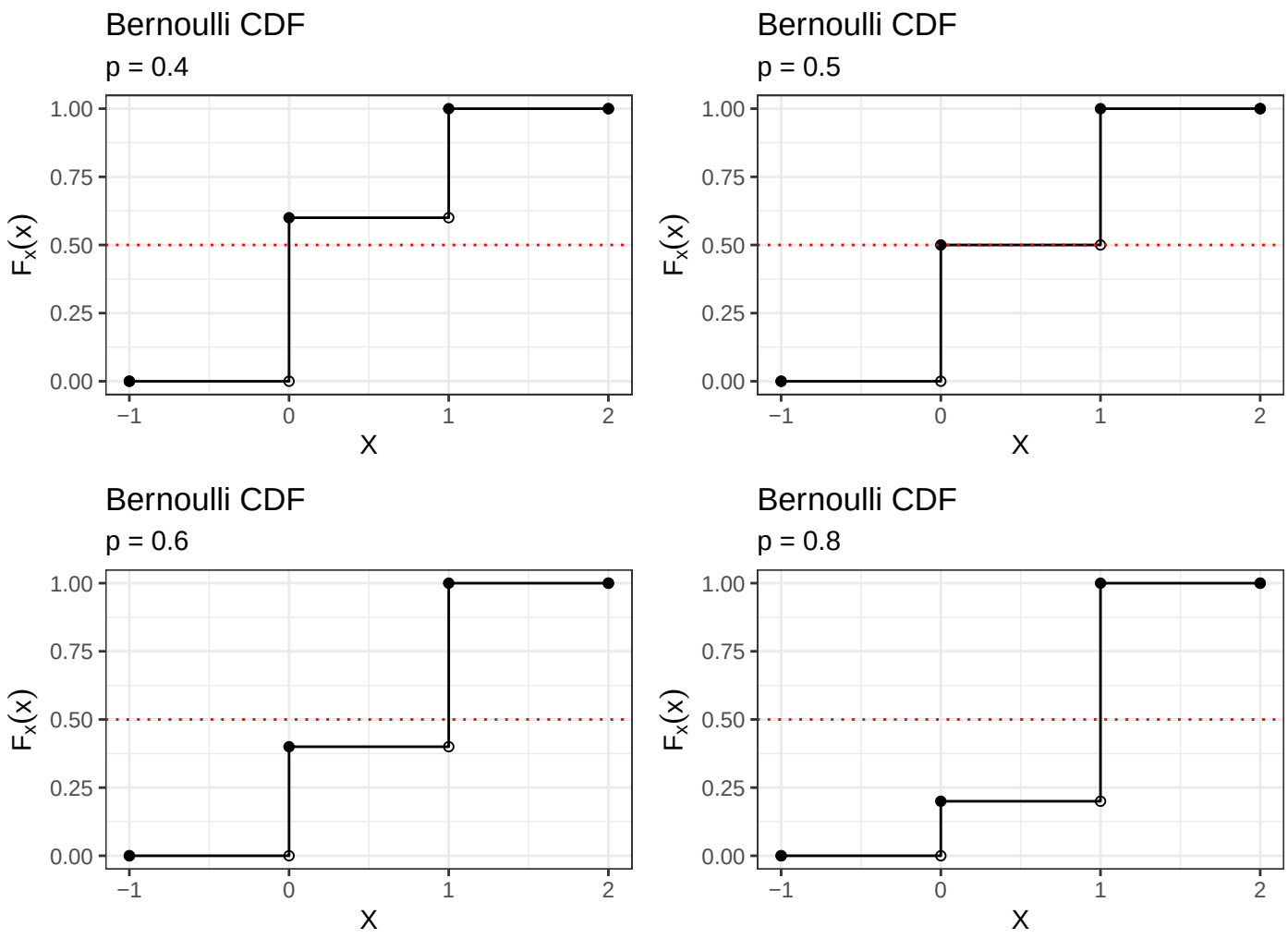

Figure 2: The CDF of the Bernoulli distribution for different probability parameters p

In the Figure 2, we can see that with increasing p, the most of the "area" under the CDF is moving towards 1, which correctly indicates that the number of successes increase as the value of p increases. Furthermore, we can see that in all the CDF plots, the functions always add up to 1. The CDF are also monotonically increasing towards the right. All of these above characteristics together indicate that our CDF is valid. Notice that in all of the plots of Figure 2, we have drawn a horizontal line at 1/2. This is another way we can determine the median of the Bernoulli distribution using the CDF. Where the horizontal line intersects the CDF indicates where the median is for the given value of p. These results agree with our results from problem 3(c).

(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.

Solution: The two parameters we are working with are p = 0.4 and p = 0.6

```r
# Using sampling with replacement to generate arrays of 0 and 1

library(tidyverse)
library(patchwork)

plot_df <- list( x1 = bern_sample(10, 0.4),
                 x2 = bern_sample(25, 0.4),
                 x3 = bern_sample(100, 0.4),
                 x4 = bern_sample(1000, 0.4),
                 y1 = bern_sample(10, 0.6),
                 y2 = bern_sample(25, 0.6),
                 y3 = bern_sample(100, 0.6),
                 y4 = bern_sample(1000, 0.6))


buildingPlot2 <- function(source, prob, sample_size){
  df <- data.frame(value=source) #turning values from the list into a df
  colnames(df) <- "value" #changing the name of the column
  df_PMF <- data.frame(x = (-1:2),
                       f = dbern(x = (-1:2), prob = prob))
  answer<-ggplot(df, aes(value))+
     geom_histogram(data = df, aes(y=..density..), binwidth=1,
                 color="black")+
  geom_linerange(data=df_PMF, aes(x=x, ymax = f), ymin = 0, size=2, color="red")+
  theme_bw() +
  ggtitle("Bernoulli Distribution",subtitle = paste("Sample Size =", sample size, ", Prob =", prob))
  answer
}

x1 <- buildingPlot2(plot_df[1], 0.4, 10)
x2 <- buildingPlot2(plot_df[2], 0.4, 25)
x3 <- buildingPlot2(plot_df[3], 0.4, 100)
x4 <- buildingPlot2(plot_df[4], 0.4, 1000)

y1 <- buildingPlot2(plot_df[5], 0.6, 10)
y2 <- buildingPlot2(plot_df[6], 0.6, 25)
y3 <- buildingPlot2(plot_df[7], 0.6, 100)
y4 <- buildingPlot2(plot_df[8], 0.6, 1000)


(x1+y1)/(x2+y2)/(x3+y3)/(x4+y4)
```

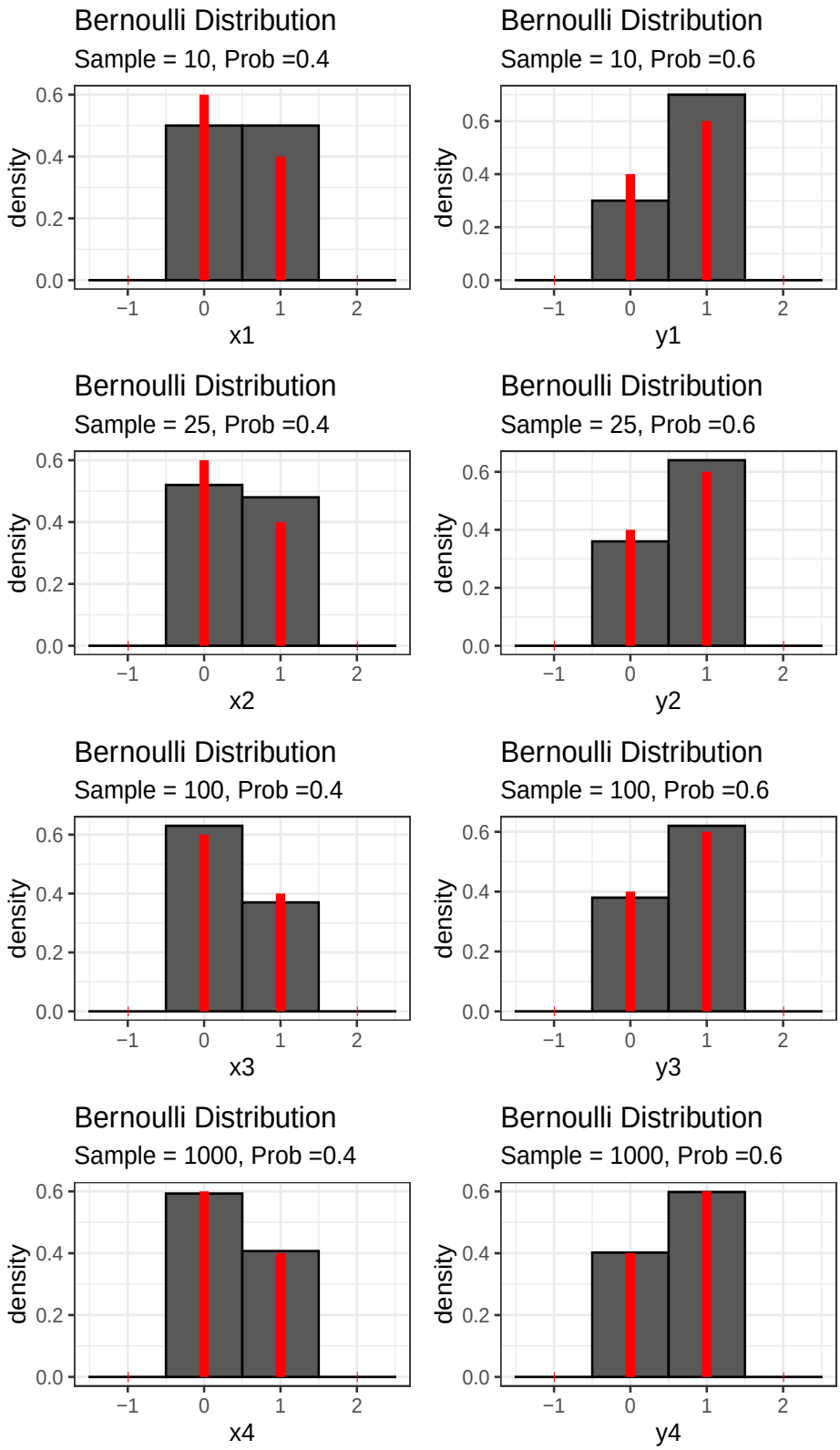

Figure 3: (Top row) Samples of 10, 25, 100, 1000 for the Bernoulli distribution with p = 0.4. (Bottom row) Samples of 10, 25, 100, 1000 for the Bernoulli distribution with p = 0.6. In each plot the red lines indicate the true PMF function of the corresponding Bernoulli distribution. Notice that the y-axis indicates the prportional frequency in the histograms

For each of the cases of p (see the first and the second row of plots in Figure 3), the more we sample, the better our histograms agree with the the true PMF of the Bernoulli distribution. These histograms above represent only one variation of the random sampling done on the Bernoulli distribution. While the histograms for n=1000 does not stray

much from the PMF as we take newere samples, the smaller samples (10, 25) vary significantly from the PMF. From sample to sample, the results of the histograms vary (in terms of agreement with the true PMF). The larger the size of the sample, the more consistent is its agreement with the true PMF.

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

Cthaeh, T. (2020). The bernoulli distribution: Intuitive understanding.

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical Distributions*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Kleinbaum, D. G. and Klein, M. (2010). Introduction to Logistic Regression. In Kleinbaum, D. G. and Klein, M., editors, *Logistic Regression: A Self-Learning Text*, Statistics for Biology and Health, pages 1–39. Springer, New York, NY.

Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. R package version 1.1.1.

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

Vukadin, O., Lee, S., Williams, C., and Khim, J. (2021). Bernoulli distribution | Brilliant.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.