Exploring Distributions

STA 310: Homework 2

Instructions

- Write all narrative using full sentences. Write all interpretations and conclusions in the context of the data.
- Be sure all analysis code is displayed in the rendered pdf.
- If you are fitting a model, display the model output in a neatly formatted table. (The `tidy` and `kable` functions can help!)
- If you are creating a plot, use clear and informative labels and titles.
- Render and back up your work reguarly, such as using Github.
- When you're done, we should be able to render the final version of the Rmd document to fully reproduce your pdf.
- Upload your pdf to Gradescope. Upload your Rmd, pdf (and any data) to Canvas.

These exercises come from BMLR or are adapted from BMLR, Chapter 3.

Exercises

Exercise 1

At what value of p is the variance of a binary random variable smallest? When is the variance the largest? Back up your answer empirically or mathematically.

The SD of a binary random variable is $\sqrt{p(1-p)}$, so the variance is simply $p(1-p)$. We must then take the first derivative to find the local maximum and minimum. We then get 1 - 2p. This first derivative = 0 when p = 0.5, giving us the critical point of (0.5, 0.25). Testing the two endpoints of the domain of p, 0 and 1, gives us the points (0,0) and (1,0). Therefore, p = 0.5 is the value of p for which the variance is largest, and 0 and 1 are the values of p for which the variance is smallest.

Exercise 2

How are hypergeometric and binomial random variables different? How are they similar?

Both of these random variables represent the number of successes in a fixed number of trials given n, the number of trials and p, the probability of a success. The only difference is that binomial random variables assume replacement, meaning that the probability of success is constant for each trial, but hypergeometric assumes no replacement, meaning that the probability of success depends on the outcome of previous trials (e.g. drawing hearts out of a deck of cards without replacing each drawn card).

Exercise 3

How are exponential and Poisson random variables related?

A Poisson random variable counts the number of events per unit of time/space for which the number of events depends upon the interval of time/space. For example, a poisson distribution might count the number of snowflakes that fall on a roof in 1 hour. A exponential distribution, in this example, might predict the time it takes for the first snowflake to fall on the roof. Exponential distributions give us predictions for the time before or between events.

Exercise 4

How are geometric and exponential random variables similar? How are they different?

Geometric and exponential random variables are similar in that they both model the time/number of trials until a event occurs. They also both have the "memoryless" property, meaning that the probabilities they generate are independent from past trials/number of time elapsed. They are different because geometric distributions map onto discrete values (number of trials), while exponential distributions map onto continuous values (time or space).

Exercise 5

A university's college of sciences is electing a new board of 5 members. There are 35 applicants, 10 of which come from the math department. What distribution could be helpful to model the probability of electing X board members from the math department?

A hypergeometric distribution would help with this probability. This distribution would give us the probability for each possible value of X (1-5) and would take into consideration the fact that this process is done without replacement.

Exercise 6

Chapter 1 asked you to consider a scenario where *"The Minnesota Pollution Control Agency is interested in using traffic volume data to generate predictions of particulate distributions as measured in counts per cubic feet."* What distribution might be useful to model this count per cubic foot? Why?

A Poisson distribution would be useful here because it counts the number of events per unit of time/space for which the number of events depends upon the interval of time/space. In this example, it would count the number of particles per cubic feet and give estimations based on the distribution.

Exercise 7

Chapter 1 also asked you to consider a scenario where *"Researchers are attempting to see if socioeconomic status and parental stability are predictive of low birthweight. They classify a low birthweight as below 2500 g, hence our response is binary: 1 for low birthweight, and 0 when the birthweight is not low."* What distribution might be useful to model if a newborn has low birthweight?

A binary/Bernoulli distrubtion would be useful here since we are trying to model probabilities of a binary variable. In this case, 1 for low birthweight, and 0 for not low birthweight.

Exercise 8

Chapter 1 also asked you to consider a scenario where *"Researchers are interested in how elephant age affects mating patterns among males. In particular, do older elephants have greater mating success, and is there an optimal age for mating among males? Data collected includes, for each elephant, age and number of matings in a given year."* Which distribution would be useful to model the number of matings in a given year for these elephants? Why?

A Poisson distribution would be useful here because it counts the number of events per unit of time/space for which the number of events depends upon the interval of time/space. In this example, it would count the number of matings per year and give estimations based on the distribution.

Exercise 9

Describe a scenario which could be modeled using a gamma distribution.

Suppose I am a birdwatcher whose goal for the day is to record the appearance of 5 birds coming to a feeder. Suppose the number of birds who visit the feeder per house follows a Poisson distribution. I could then use a gamma distribution to model the amount of time it would take to record 5 birds since gamma distributions model wait time before r number of events occur.

Exercise 10

Beta-binomial distribution. We can generate more distributions by mixing two random variables. Beta-binomial random variables are binomial random variables with fixed n whose parameter p follows a beta distribution with fixed parameters α, β. In more detail, we would first draw p_1 from our beta distribution, and then generate our first observation y_1, a random number of successes from a binomial (n, p_1) distribution. Then, we would generate a new p_2 from our beta distribution, and use a binomial distribution with parameters n, p_2 to generate our second observation y_2. We would continue this process until desired.

Note that all of the observations y_i will be integer values from $0, 1, \ldots, n$. With this in mind, use `rbinom()` to simulate 1,000 observations from a plain old vanilla binomial random variable with $n = 10$ and $p = 0.8$. Plot a histogram of these binomial observations. Then, do the following to generate a beta-binomial distribution:

a. Draw p_i from the beta distribution with $\alpha = 4$ and $\beta = 1$.
b. Generate an observation y_i from a binomial distribution with $n = 10$ and $p = p_i$.
c. Repeat (a) and (b) 1,000 times ($i = 1, \ldots, 1000$).
d. Plot a histogram of these beta-binomial observations.

Compare the histograms of the "plain old" binomial and beta-binomial distributions. How do their shapes, standard deviations, means, possible values, etc. compare?

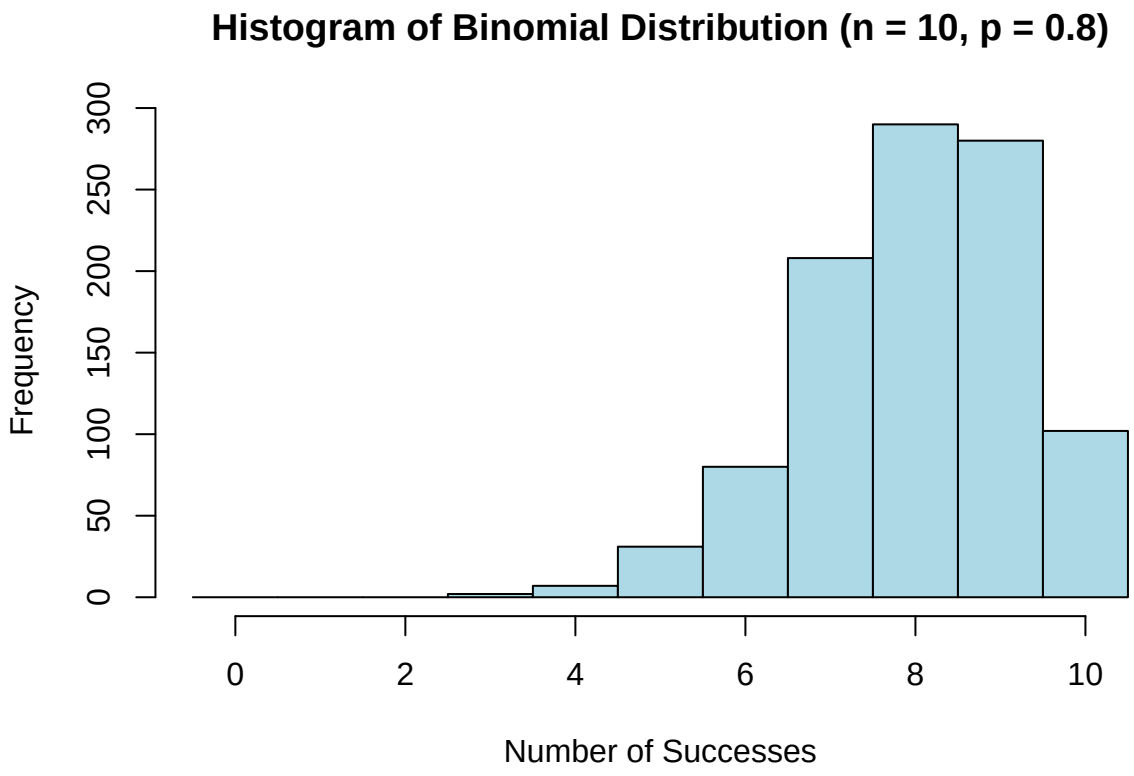
```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

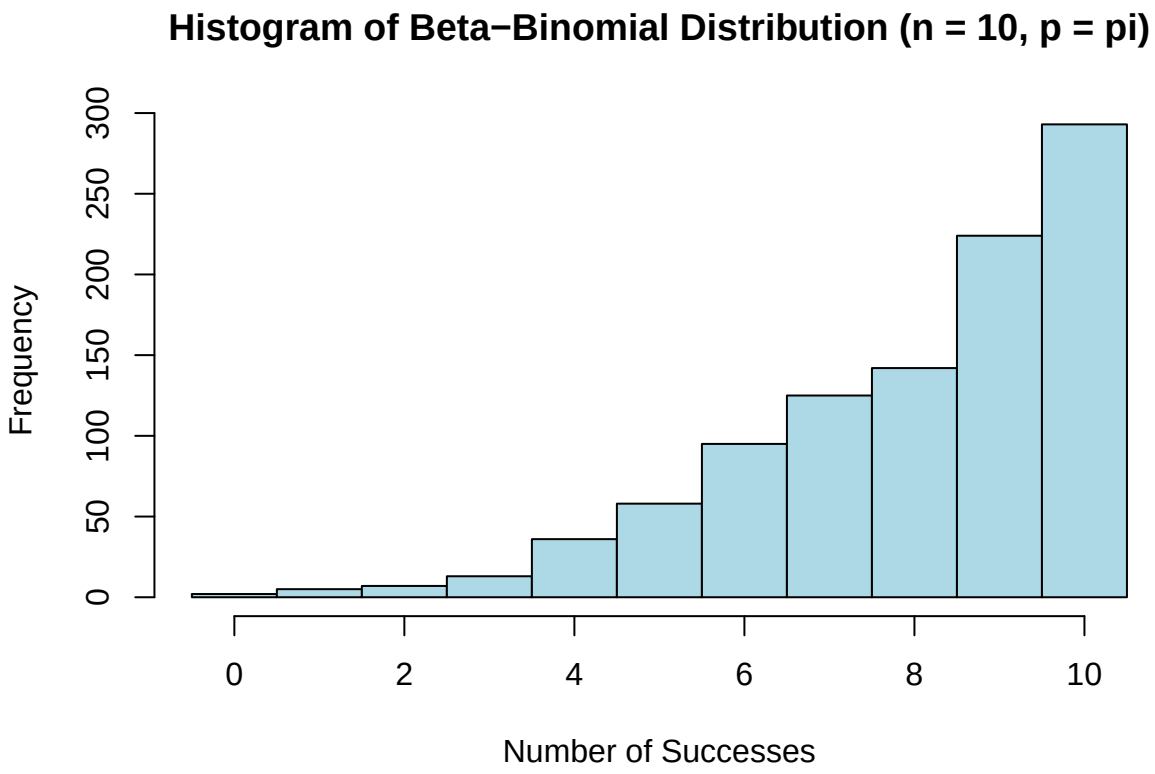
```r
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.2.0 --
## v broom        1.0.6     v rsample      1.2.1
## v dials        1.3.0     v tune         1.2.1
## v infer        1.0.7     v workflows    1.1.4
## v modeldata    1.4.0     v workflowsets 1.1.0
## v parsnip      1.2.1     v yardstick    1.3.1
## v recipes      1.1.0
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use tidymodels_prefer() to resolve common conflicts.
```

```r
library(knitr)

set.seed(3)

binom <- rbinom(n = 1000, size = 10, prob = 0.8)

hist(binom,
     main = "Histogram of Binomial Distribution (n = 10, p = 0.8)",
     xlab = "Number of Successes",
     ylab = "Frequency",
     col = "lightblue",
     border = "black",
     breaks = seq(-0.5, 10.5, by = 1)
     )
```




```r
y <- numeric(length = 1000)
for (i in 1:1000) {
  p_i <- rbeta(1,4,1)
  y[i] = rbinom(1,10,p_i)
}

hist(y,
     main = "Histogram of Beta-Binomial Distribution (n = 10, p = pi)",
     xlab = "Number of Successes",
     ylab = "Frequency",
     col = "lightblue",
     border = "black",
     breaks = seq(-0.5, 10.5, by = 1))
```




```
mean(binom)
```

```
## [1] 7.985
```

```
sd(binom)
```

```
## [1] 1.287802
```

```
mean(y)
```

```
## [1] 8.019
```

```
sd(y)
```

```
## [1] 1.988113
```

Comparing the histograms, the first thing I notice is that they are both skewed left. The binomial distribution peaks around 8, where the beta-binomial peaks at 10, and has a staircase shape. The shapes are also different in that the beta-binomial distribution has a longer left tail, with values represented all from 0-10, where the binomial distribution only stretches to 3.

While the means are very similar, the standard deviation of the beta-binomial is fairly larger, evidenced by the spread of the historgrams. In terms of possible values, they both have possible values 0-10, but of course we can only see the full spread of values in the beta-binomial distribution. This seems to be because for some lower values of p_i taken from the beta distribution, it will be more likely that there would be fewer successes than when the p is fixed at 0.8.

Grading

Total	**33**
Ex 1	5
Ex 2	5
Ex 3	5
Ex 4	5
Ex 5	2
Ex 6	2
Ex 7	2
Ex 8	2
Ex 9	2
Ex 10	5
Workflow & formatting	3

The "Workflow & formatting" grade is to based on the organization of the assignment write up along with the reproducible workflow. This includes having an organized write up with neat and readable headers, code, and narrative, including properly rendered mathematical notation. It also includes having a reproducible Rmd or Quarto document that can be rendered to reproduce the submitted PDF.