Exploring Distributions - STA 310: Homework 2

Drew Davison

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 equation for variance of a binary random variable is equal to (p)(1-p).

Thus, the variance is smallest when p is equal to 0 or 1. At these extremes, there is no variance.

$(0)(1-0) = 0$ AND $(1)(1-1) = 0$.

The variance is largest when p is equal to 0.5.

$(0.5)(1-0.5) = .25$

The function for variance, (p)(1-p), is a parabola with roots at p=0 and p=1, with the vertex at p=0.5.

Exercise 2

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

The random variables are different in many ways.

Firstly, hypergeometric random variables are sampled without replacement, while binomial random variables are sampled with replacement.

In hypergeometric cases, the population size N and the number of successes K in the population are explicitly included in the equations to solve the probability. In the binomial case, the population is assumed to be infinite or large enough that the probability of success remains the same in all trials.

The probability of success changes after each trial in hypergeometric cases, while the probability of success is fixed in binomial cases.

The parameters of a hypergeometric random variable are population size, number of successes in the population, and number of trials. The parameters of a binomial random variable are number of trials and probability of success.

However, on the other hand, there are many similarities between the two random variables.

Both distributions are discrete probability distributions. They both have a range of successes from 0 to n, the number of trials. Lastly, both distributions rely on a binary success or failure structure.

Exercise 3

How are exponential and Poisson random variables related?

The exponential and Poisson distributions are two distributions in a Poisson process. The exponential random variable models the time between events, while the Poisson random variable models the number of events in a fixed interval. Both are driven by the same rate parameter λ. An exponential distribution is the continuous random variable related to the Poisson discrete random variable through the Poisson proccess.

Exercise 4

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

Geometric and exponential random variables are similar because they describe the time or number of trials until the first success in a random process. Both distributions are memoryless, which implies that the probability of waiting longer is independent of how much time or how many trials have passed.

The difference between the two distributions is that geometric distributions are discrete, while exponential distributions are continuous. While related, the probability functions for each distribution are different.

Geometric: $\mathrm{P(X = k)} = (1-p)^{k-1}\ p$

Exponential: $f_T(t) = \lambda\ \mathrm{e}^{-\lambda t}$

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 could be helpful to model this scenario. Because selecting board members occurs without replacement, we are limited to distributions that model scenarios without replacement. The population size is 35, the number of successes in the population is 10, and the sample size is 5. These are the parameters needed for a hypergeometric distribution.

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?

The Poisson distribution could be useful to model this scenario. This distribution is measured in counts, which fits the type of data modeled by a Poisson distribution. The Poisson distribution is used to model the number of events occurring in a fixed interval of time, space, or volume. Additionally, it is reasonable to assume that the distribution of events is independent.

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?

The Bernoulli distribution could be useful to model if a newborn has low birthweight. The response variable is binary (1 for low birthweight, and 0 for not low birthweight), and the Bernoulli distribution is used to model a trial with two possible outcomes. The probability of low birthweight represents the "success" probability.

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?

The Poisson distribution would be useful because the scenario above deals with count data (the number of matings) occurring within a fixed period (one year).

Exercise 9

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

Three friends are out bird-watching. On average they see three cardinals per hour, and their goal is to see seven cardinals. What is the probability that they take less than 2 hours to reach their goal?

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?

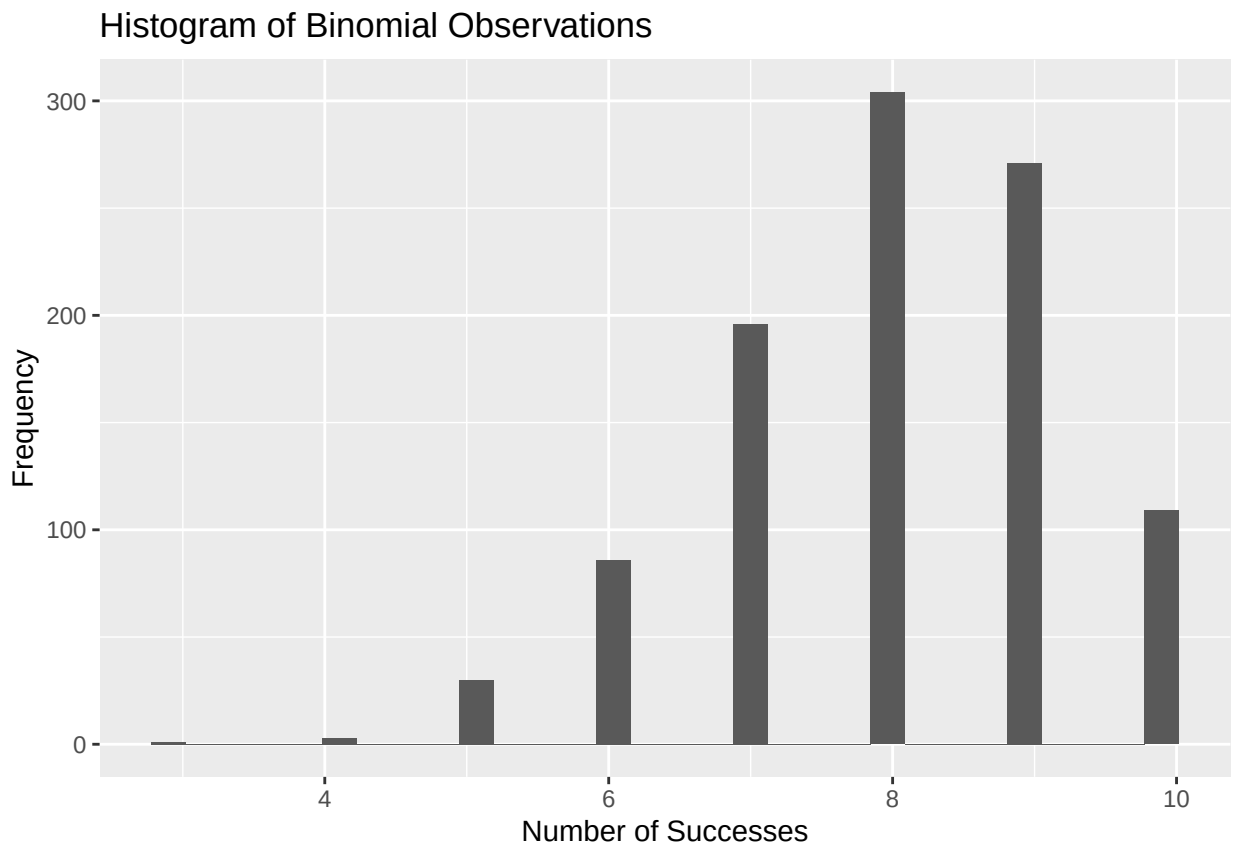
```r
library(tidyverse)
library(tidymodels)
library(skimr)

set.seed(123)

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

binom_df <- data.frame(observation = binom_obs, type = "Binomial")

ggplot(data = binom_df, aes(x = observation)) +
    geom_histogram() +
      labs(x = "Number of Successes",
         y = "Frequency",
         title = "Histogram of Binomial Observations")
```

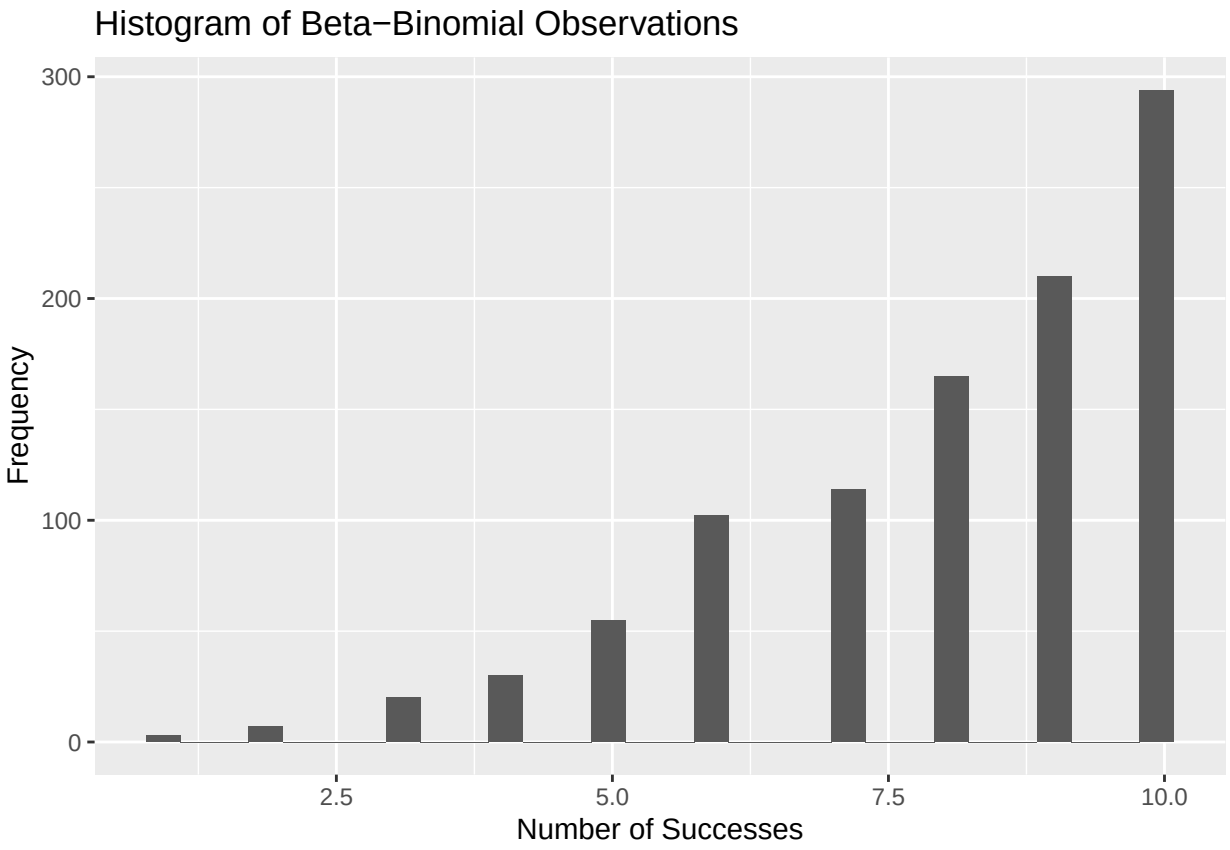



```r
beta_binom_obs <- numeric(1000)

for(i in 1:1000) {
  pi <- rbeta(1, shape1 = 4, shape2 = 1)
  beta_binom_obs[i] <- rbinom(1, size = 10, prob = pi)
}

beta_binom_df <- data.frame(observation = beta_binom_obs,
                            type = "Beta-Binomial")
```

```
ggplot(data = beta_binom_df, aes(x = observation)) +
   geom_histogram() +
     labs(x = "Number of Successes",
        y = "Frequency",
        title = "Histogram of Beta-Binomial Observations")
```




```
combined_df <- rbind(binom_df, beta_binom_df)

summary_stats <- combined_df %>%
 group_by(type) %>%
 summarise(
   mean_value = mean(observation),
   sd_value = sd(observation)
 )
print(summary_stats)
```

```
## # A tibble: 2 x 3
##   type          mean_value sd_value
##   <chr>              <dbl>    <dbl>
## 1 Beta-Binomial       8.03     1.95
## 2 Binomial            8.01     1.26
```

Both distributions are left-skewed, though the beta-binomial distribution is more left skewed than the binomial distribution. The means for the distributions are very similar, separated by.018, though there is a

roughly .684 difference in their standard deviations. The Beta-Binomial distribution has a greater standard deviation. The possible values within the Beta-Binomial distribution has a greater range (1 success to 10 successes), compared to the Binomial distribution, (3 successes to 10 successes).

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.