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.

Recall that the variance of a binary random variable $X \sim \text{Bernoulli}(p)$ is $p(1-p)$. To get the maximum and minimum, let's calculate the derivative of the variance with respect to p and set it to 0:

$$\frac{d}{dp}p(1-p) = 0 \Rightarrow 1 - 2p = 0 \Rightarrow p = 0.5$$

Note that when $p = 0.5$, $Var(X) = 0.25$, which is clearly the maximum. Additionally, since $p \in (0,1)$, then $p(1-p) \leq 0$, which means that the minimum possible for the variation is 0, what can be obtained when $p = 0 \vee p = 1$.

Therefore, the variance is the largest ($Var(X) = 0.25$) when $p = 0.25$, and the variance is the smallest ($Var(X) = 0$) when $p = 0 \vee p = 1$.

Exercise 2

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

Similarities? They are both discrete distributions, they are both related to the success of a event with probability p, and both uses binomials in their probability formulas.

Differences? Each of the trails in binomial are independent, while in hypergeometric they aren't (as the draws are without replacement). The formula for each is also differnet:

For binomial:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

For hypergeometric:

$$\mathbb{P}(Y = y) = \frac{\binom{m}{y}\binom{N-m}{n-y}}{\binom{N}{n}}$$

Additionally, they both have different expectation and variance formulas.

Exercise 3

How are exponential and Poisson random variables related?

Poisson counts the number of events that occur in a fixed time interval or space. Exponential measures the time between consecutive events with a Poisson distribution. Additionally, they are both part of the same Poisson process (let's say rate is λ), in the sense that the number of events in a given time (or space) is modeled by $X \sim Poisson(\lambda)$ while the time until the first event is modeled by $T \sim Exp(\lambda)$.

Exercise 4

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

Similarities? Both are related to the first success/event. In specific, geometric measures the number of trials until the first success while the exponential one measures the time until the first event. Both are related to the Poisson process Both does not depend on time on how much time has passed, that is

$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t)$$

Differences? Geometric is discrete and Exponential is continuous, because Geometric measures number of trials and Exponential measures time. The pdf of both are different

Geometric:

$$\mathbb{P}(X = k) = (1-p)^{k-1} p$$

Exponential:

$$f(t) = \lambda e^{-\lambda t}$$

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?

Since we are selecting X different board members, once we select some, we can't selec them again. That is, this involves drawing without replacement.

With that in mind, the hypergeometric distribution would be the best one, where $N = 35$ population, $K = 10$ successes (i.e. math department), $n = 5$ number of board members selected. In summary

$$\mathbb{P}(X = x) = \frac{\binom{10}{x}\binom{25}{5-x}}{\binom{35}{5}}$$

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?

Since we are truing to look the count in a limited space, we can use Poisson distribution to model the data. Poisson models the number of events in a fixed space. Since the data is independent and occurr at an approximately constant rate λ, then Poisson is ideal.

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?

Since this model is looking for an event with a binary probability (to be classified as low birthweight or not), then we can use the Bernoulli distribution. Now, if we are trying to model the sample of newborns that are classified as birthweight, then we would have to use the Binomial distribution.

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?

Since we are looking to count the number of mating in a specific time interval (a year), and the predictor variable, age, is continuous, we could use Poisson distribution to model this situation, and we can assume that the events would be independent from each other.

Exercise 9

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

Suppose that passengers at RDU airport are randomly selected at TSA for extra screening over time (like me!). If we assume that those selections happen at an approximate constant rate, we might be able to get the total time since the airport opening in the morning until 10 passengers have been selected for extra screening. Since the waiting time can be represented with exponential distribution, then the total time can be modelled with Gamma distribution.

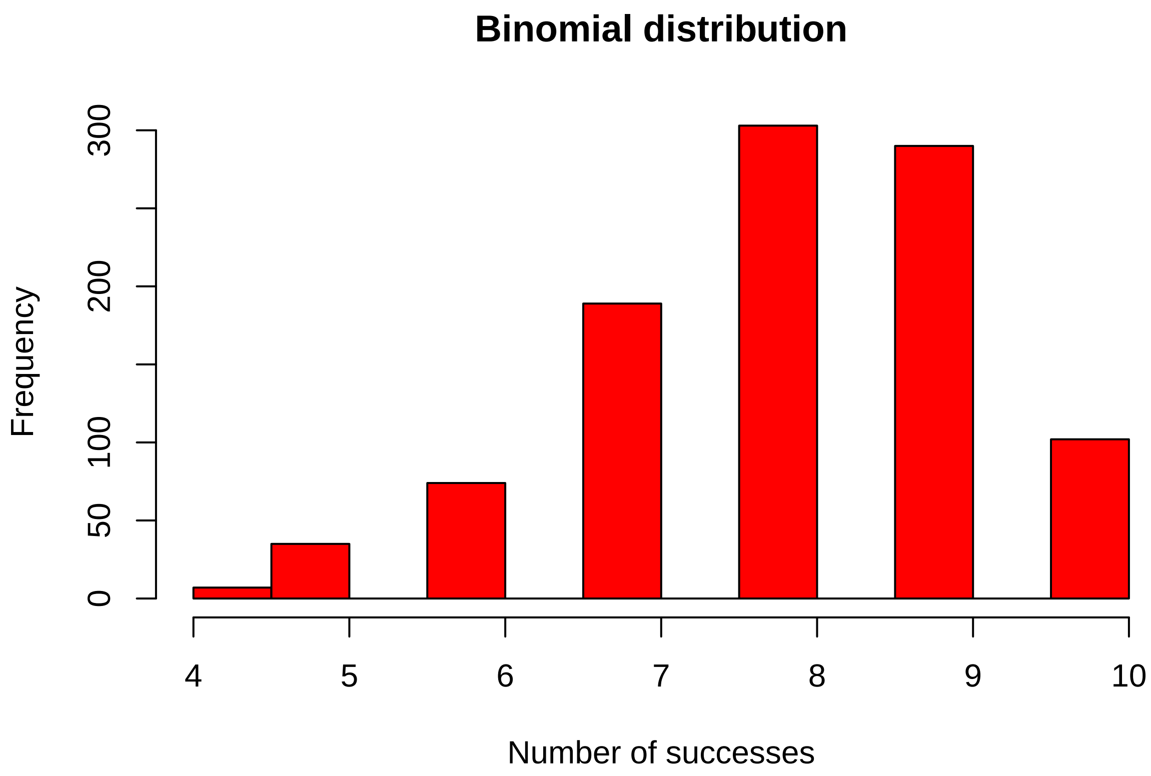
Exercise 10

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:

```r
set.seed(310) #reproducibility

binom = rbinom(1000, 10, 0.8)
hist(binom,
     main = "Binomial distribution",
     xlab = "Number of successes",
     col = "red")
```



a. Draw p_i from the beta distribution with $\alpha = 4$ and $\beta = 1$.

```r
p_i <- rbeta(1000, shape1 = 4, shape2 = 1)
```

b. Generate an observation y_i from a binomial distribution with $n = 10$ and $p = p_i$.

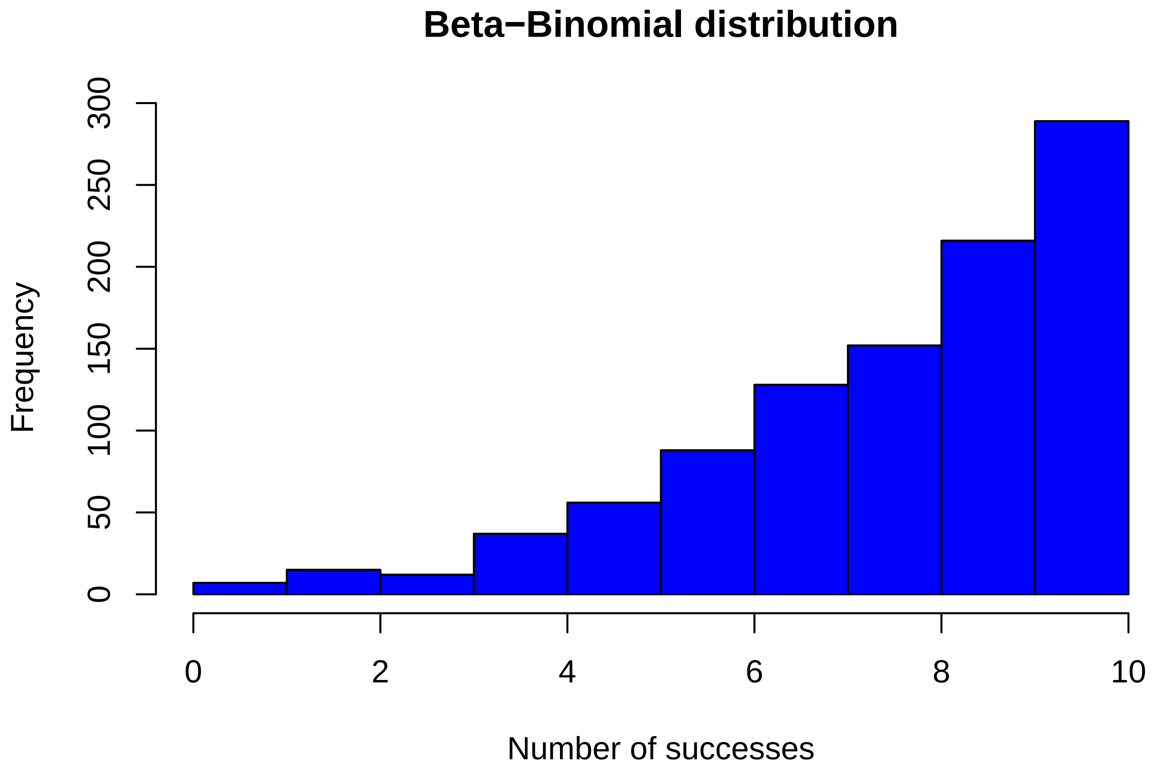
```r
y_i <- rbinom(1000, 10, p_i)
```

c. Repeat (a) and (b) 1,000 times ($i = 1, \ldots, 1000$).

Already did this by using $n = 1,000$ on each `rbinom`, `rbeta` function calls.

d. Plot a histogram of these beta-binomial observations.

```
hist(y_i,
     main = "Beta-Binomial distribution",
     xlab = "Number of successes",
     col = "blue")
```



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

They both take values in the same set $\{0, 1, \ldots, 10\}$.

In terms of shape, the plain binomial is roughly symmetric (a bit left-skewed due to the relatively high p value), and most of the mass is concentrated around 8-9 successes. For the beta-binomial, it is more spread out and has a more pronounced left-skewed distribution. This distribution has heavier tails towards the low and high counts.

In terms of means, we have that for the binomial distribution, this is around $\mathbb{E}[X] = np = 10 \times 0.8 = 8$, while the mean for the beta-binomial would be $\mathbb{E}[X] = \frac{n\alpha}{\alpha+\beta} = \frac{10\times4}{4+1} = 8$. Thus, both means are the same. We can confirm this by calculating

```
mean(binom)
```

```
## [1] 8.024
```

```
mean(y_i)
```

```
## [1] 7.973
```

Finally, in terms of the variability, we know that for the binomial distribution this will be $Var(X) = np(1-p) = 10 \times 0.8 \times 0.2 = 1.6$ while for the beta-binomial it'll have a variation of $Var(X) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{10\times4\times1\times15}{(5)^2\times6} = \frac{600}{150} = 4$ which is considerably largers than the variation for the binomial distribution. We can also confirm this by calculating

```
var(binom)
```

```
## [1] 1.611035
```

```
var(y_i)
```

```
## [1] 4.164435
```

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.