STA310 HW2

Olivia Fu

2025-01-24

```r
library(knitr)
library(ggplot2)
```

Exercise 1

The variance of a binary random variable is the smallest when p = 0 or 1.

The variance of a binary random variable is the largest when p = 0.5.

$$E[X] = 1 \times p + 0 \times (1 - p) = p$$
$$E[X^2] = 1^2 \times p + 0^2 \times (1 - p) = p$$
$$Var(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$$

Considering that p is bounded within [0,1], the minimum variance is 0, which occurs when p=0 or p=1. To determine the maximum variance, we can take the first derivative of the variance equation and set it to 0. Solving this gives p=0.5, which results in the maximum variance, because we know that the quadratic function is concave downward. We can also confirm by observing that the second derivative is negative.

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

Intuitively, when p=0 or p=1, the random variable X becomes deterministic, always taking the same value, which results in a variance of 0. Conversely, when P=0.5, X has an equal probability of taking the values 0 or 1, leading to the maximum variation.

Exercise 2

Similarities between hypergeometric and binomial random variables

- Both are discrete random variables that describe the number of successes in a given set of trials.

Differences between hypergeometric and binomial random variables

- For binomial random variable, each trial is independent from each other, and the probability of success (p) remaines constant across all trials. However, for hypergeometric random variable, the trials are dependent. The probability of success is dynamic, depending on previous selections.

- Binomial random variable describes sampling with replacement, while hypergeometric random variable involves sampling without replacement (e.g., drawing cards without replacing them).

Exercise 3

$$\text{Poisson} : P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad \text{for } y = 0, 1, \dots, \infty$$

$$\text{Exponential} : f(y) = \lambda e^{-\lambda y}, \quad \text{for } y > 0$$

Both Poisson and exponential random variables are related to the Poisson process. Poisson random variable is a discrete RV that describes **the number of events** that occur per unit of time in a Poisson process with rate λ. Exponential random variable is a continuous RV that describes the **wait time** until the first event (or time until the next event) occurs in a Poisson process with rate λ.

Exercise 4

$$\text{Geometric}: P(Y = y) = (1 - p)^y \cdot p, \quad \text{for } y = 0, 1, \dots, \infty$$
$$\text{Exponential}: f(y) = \lambda e^{-\lambda y}, \quad \text{for } y > 0$$

Geometric random variable represents the number of independent, identical Bernoulli trials until the first success. Exponential random variable represents the waiting time until the first event in a Poisson process.

Similarities between geometric and exponential random variables

- Both describe the waiting time (or number of trials) until the occurrence of the first event or success. Both focus on modeling the time or trials until the first target event. The exponential distribution is the continuous-time counterpart of the geometric distribution.

- Both decay as y increases.

- Both are memoryless.

Differences between geometric and exponential random variables

- Geometric random variable is discrete, while exponential random variable is continuous.

- A geometric random variable is associated with Bernoulli process, whereas an exponential random variable is related to Poisson process.

Exercise 5

Hypergeometric distribution could be helpful to model the probability of electing X board members from the math department.

This is because the election process involves selection without replacement from a small, finite population of 35 applicants. The group of interest consists of 10 applicants from the math department. Therefore, the hypergeometric distribution is appropriate for modeling the number of board members from the math department (out of a total of 5 board members), selected from a population of 35 applicants, where 10 are from the math department.

In this case, we can model the probability with $P(X = x) = \frac{\binom{10}{x}\binom{25}{5-x}}{\binom{35}{5}}$

Exercise 6

Poisson distribution might be useful to model particulate in count per cubic foot.

Firstly, Poisson distribution's support matches the data, as particulate counts are non-negative integers. Additionally, the Poisson process is often used to model the number of events occurring per unit of time or space — in this case, the particulate counts per cubic foot of air.

Exercise 7

Bernoulli distribution might be useful to model if a newborn has low birthweight.

Researchers classify low birthweight as below 2500g, making the response variable binary: 1 for low birthweight and 0 for not low birthweight. The Bernoulli distribution is appropriate for modeling variables with only two possible outcomes. Furthermore, each newborn represents a single trial, which is naturally suited to the Bernoulli distribution. Then, researchers can investigate whether socioeconomic status and parental stability are predictive of low birthweight.

Exercise 8

Poisson distribution would be useful to model the number of matings in a given year for these elephants.

Firstly, Poisson distribution's support matches the data, as the numbers of matings are non-negative integers. Additionally, the Poisson process is often used to model the number of events occurring in a fixed time or space. In this context, we have the number of matings in a given year (a fixed time interval), making Poisson distribution a good fit.

Exercise 9

The Gamma distribution models the total waiting time until r events occur in a Poisson process. For example, suppose customers arrive at a pizza restaurant at a constant average rate of λ (e.g., 10 customers per hour). If we want to model the total time it takes for the 20th customer to arrive (e.g., the probability that it takes fewer than 2 hours for the restaurant to serve 20 customers), the Gamma distribution is good fit.

Exercise 10

```r
set.seed(310)
plain_binom <- rbinom(n = 1000, size = 10, prob = 0.8)

plain_binom_df <- data.frame(count = plain_binom)
ggplot(plain_binom_df, aes(x = count)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  scale_x_continuous(breaks = 0:10) +
  labs(title = "Histogram of Plain Binomial Observations",
       subtitle = "n=10 and p=0.8",
       x = "Number of Successes",
       y = "Frequency") +
  theme_minimal()
```
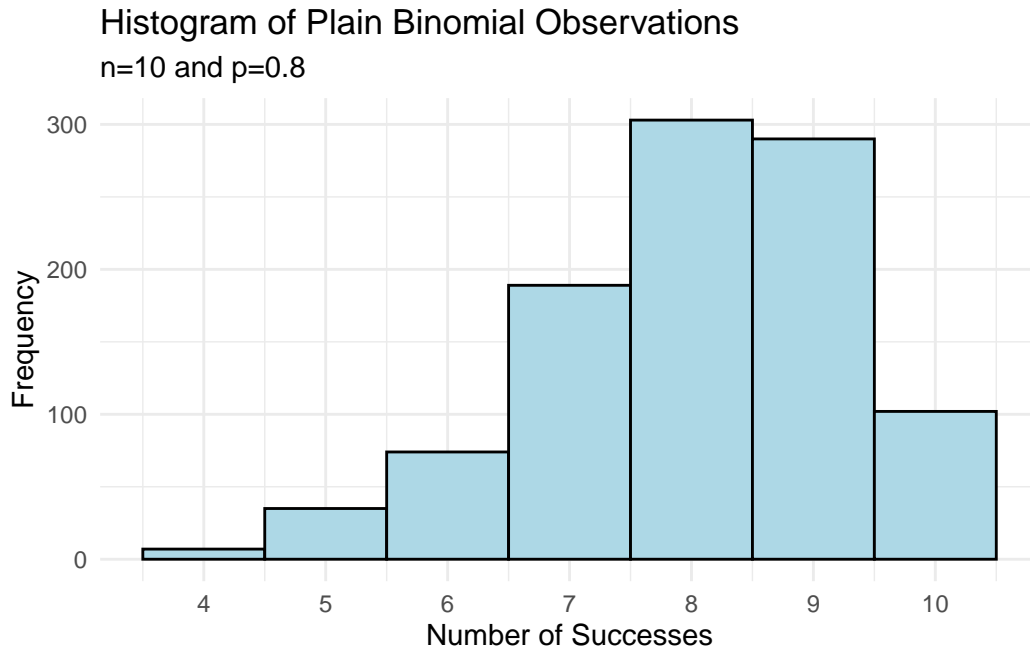


```r
set.seed(310)
beta_binom <- numeric(1000)
for (i in 1:1000) {
  p <- rbeta(1, shape1 = 4, shape2 = 1)
  beta_binom[i] <- rbinom(1, size = 10, prob = p)
}
```

```
beta_binom_df <- data.frame(count = beta_binom)

ggplot(beta_binom_df, aes(x = count)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  scale_x_continuous(breaks = 0:10) +
  labs(title = "Histogram of Beta-Binomial Observations",
       x = "Number of Successes",
       y = "Frequency") +
  theme_minimal()
```

Histogram of Beta–Binomial Observations
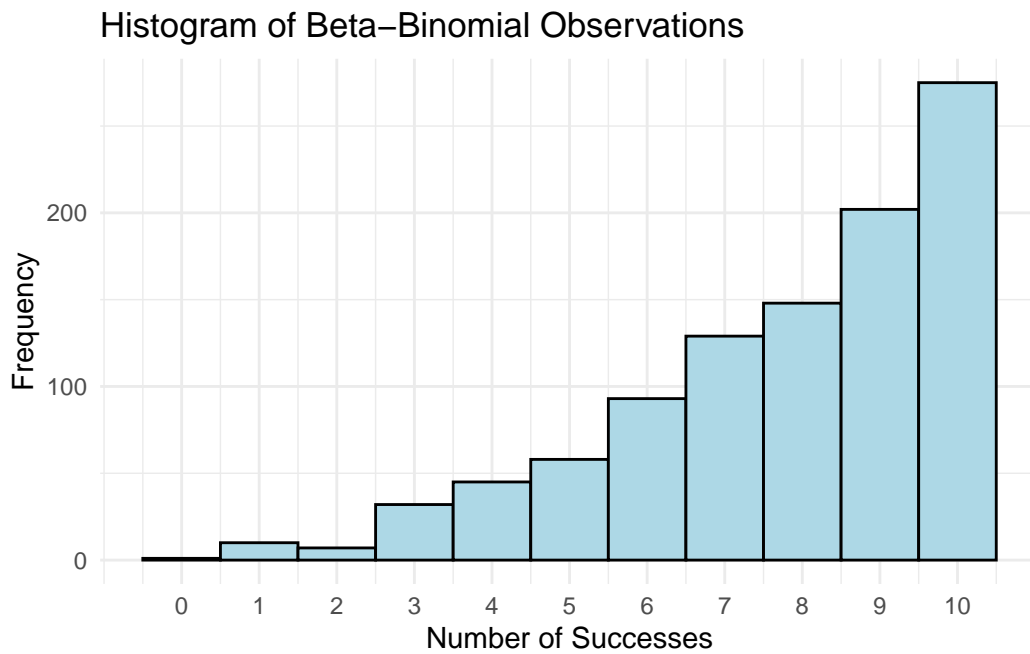

```
plain_stats <- c(
  mean = mean(plain_binom),
  sd = sd(plain_binom),
  min = min(plain_binom),
  max = max(plain_binom)
)

beta_stats <- c(
  mean = mean(beta_binom),
  sd = sd(beta_binom),
  min = min(beta_binom),
  max = max(beta_binom)
```

```
)

data.frame(
  Statistic = c("Mean", "Standard Deviation", "Minimum", "Maximum"),
  Plain_Binomial = plain_stats,
  Beta_Binomial = beta_stats
) |>
  kable(row.names = FALSE)
```

Statistic	Plain_Binomial	Beta_Binomial
Mean	8.024000	7.803000
Standard Deviation	1.269265	2.148676
Minimum	4.000000	0.000000
Maximum	10.000000	10.000000

The histogram of the plain binomial distribution is left-skewed and centered around 8, reflecting the fixed probability of success (p=0.8). In contrast, the histogram of the beta-binomial distribution is more spread out, with frequencies gradually increasing from 0 to 10. This shape aligns with the density function of the Beta(4, 1) distribution, which gradually increases from 0 to 1.

The mean of the plain binomial distribution is 8.024, which closely matches the theoretical expectation of $10 \times 0.8 = 8$. The mean of the beta-binomial distribution is slightly lower at 7.803. The two means are very similar because the expected value of p in the Beta distribution is also $\frac{4}{4+1} = 0.8$. However, the means still differ due to the additional variability in generating p from the Beta distribution.

The standard deviation of the beta-binomial distribution (2.149) is higher than that of the plain binomial distribution (1.269), as sampling p from a Beta distribution introduces an extra layer of variability beyond the fixed p used in the plain binomial distribution.

The range of possible values for the plain binomial distribution is from 4 to 10, which is consistent with p=0.8 and outcomes centered around the mean. On the other hand, the beta-binomial distribution has a broader range of values from 0 to 10, reflecting the increased variability in p, which allows for a wider spread of possible outcomes.