M50 Homework 2

Alex Craig

Exercise 1.

(Computing conditional averages): Consider a random variable $Y = (Y_1, Y_2)$ which takes values in the sample space:

$$S = \mathbb{N} \times \mathbb{N} = (i, j), i, j \in \mathbb{N}$$

That is, the sample space consists of all possible pairs of numbers (i, j). Now suppose we have some data:

$$(1, 2), (1, 2), (3, 1), (1, 4), (3, 3), (2, 2), (1, 5)$$

Give you best estimates of the following (either by hand, with Python, or a calculator)

$$E[Y_1], \quad E[Y_1 \mid Y_2 = 2], \quad E[Y_2 \mid Y_1 = 1], \quad E[Y_2 \mid Y_1 > 1]$$

Solution

$$E[Y_1] \approx \frac{1+1+3+1+3+2+1}{7} = \frac{12}{7}$$

$$E[Y_1 \mid Y_2 = 2] \approx \frac{1+1+2}{3} = \frac{4}{3}$$

$$E[Y_2 \mid Y_1 = 1] \approx \frac{2+2+4+5}{4} = \frac{13}{4}$$

$$E[Y_2 \mid Y_1 > 1] \approx \frac{1+3+2}{3} = 2$$

Exercise 2.

(Independence and conditional expectation): Let X and Y be two random variables with sample spaces S_X and S_Y.

Part A

Prove that if X and Y are independent $E[X \mid Y = y] = E[X]$ and $E[Y \mid X = x] = E[Y]$ for all $x \in S_X$ and $y \in S_Y$.

Solution

We know that conditional expectation is defined as:

$$E[X \mid Y = y] = \sum_{x \in S_X} x \cdot \mathbb{P}(X = x \mid Y = y)$$

Keep in mind that conditional probability is defined as:

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

For independent random variables, we know that:

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

Therefore, we may rewrite the conditional probability as:

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)}{\mathbb{P}(Y = y)} = \mathbb{P}(X = x)$$

Therefore, we may rewrite the conditional expectation as:

$$E[X \mid Y = y] = \sum_{x \in S_X} x \cdot \mathbb{P}(X = x \mid Y = y) = \sum_{x \in S_X} x \cdot \mathbb{P}(X = x) = E[X]$$

$$\Rightarrow E[X \mid Y = y] = E[X]$$

Part B

Prove the tower property of expectation that is stated in the class notes.

Solution

Part C

Prove that if X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$.

Hint: Use this formula for variance: $Var(X) = E[X^2] - E[X]^2$.

Solution

$$Var(X + Y) = E[(X + Y)^2] - E[X + Y]^2 = E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2$$

Using the linearity of expectation, we may rewrite the above as:

$$= E[X^2] + 2E[XY] + E[Y^2] - (E[X]^2 + 2E[X]E[Y] + E[Y]^2)$$

If X and Y are independent, then $E[XY] = E[X]E[Y]$. Therefore, we may rewrite the above as:

$$= E[X^2] + 2E[X]E[Y] + E[Y^2] - (E[X]^2 + 2E[X]E[Y] + E[Y]^2)$$

Exercise 3.

(Aspects of the binomial distribution): Suppose Y_1 and Y_2 are two independent binomial distributions:

$$Y_1 \sim Bin(n_1, p_1), \quad Y_2 \sim Bin(n_2, p_2)$$

with $n_1, n_2 \in \mathbb{N}$ and $p_1, p_2 \in (0, 1)$

Part A

If $p = p_1 = p_2$, what is the distribution of $Y_1 + Y_2$?

Solution

Given that $p = p_1 = p_2$, we may define the probability mass function of Y_k as:

$$\mathbb{P}(Y_k = y) = \binom{n_k}{y} p^y (1 - p)^{n_k - y}, \quad y = 0, 1, \ldots, n_k, \quad k = 1, 2$$

Define $Y = Y_1 + Y_2$. We may then define the probability mass function of Y as:

$$\mathbb{P}(Y = y) = \mathbb{P}(Y_1 + Y_2 = y) = \sum_{i=0}^{y} \mathbb{P}(Y_1 = i, Y_2 = y - i)$$

Since Y_1 and Y_2 are independent, we may rewrite the above as:

$$\sum_{i=0}^{y} \mathbb{P}(Y_1 = i, Y_2 = y - i) = \sum_{i=0}^{y} \mathbb{P}(Y_1 = i) \cdot \mathbb{P}(Y_2 = y - i)$$

And now we may substitute in the probability mass function of Y_1 and Y_2:

$$= \sum_{i=0}^{y} \binom{n_1}{i} p^i (1-p)^{n_1-i} \cdot \binom{n_2}{y-i} p^{y-i} (1-p)^{n_2-y+i}$$

$$= \sum_{i=0}^{y} \binom{n_1}{i} \binom{n_2}{y-i} p^y (1-p)^{n_1+n_2-y}$$

$$= p^y (1-p)^{n_1+n_2-y} \sum_{i=0}^{y} \binom{n_1}{i} \binom{n_2}{y-i}$$

And, by Vandermonde's Identity:

$$= p^y (1-p)^{n_1+n_2-y} \binom{n_1+n_2}{y}$$

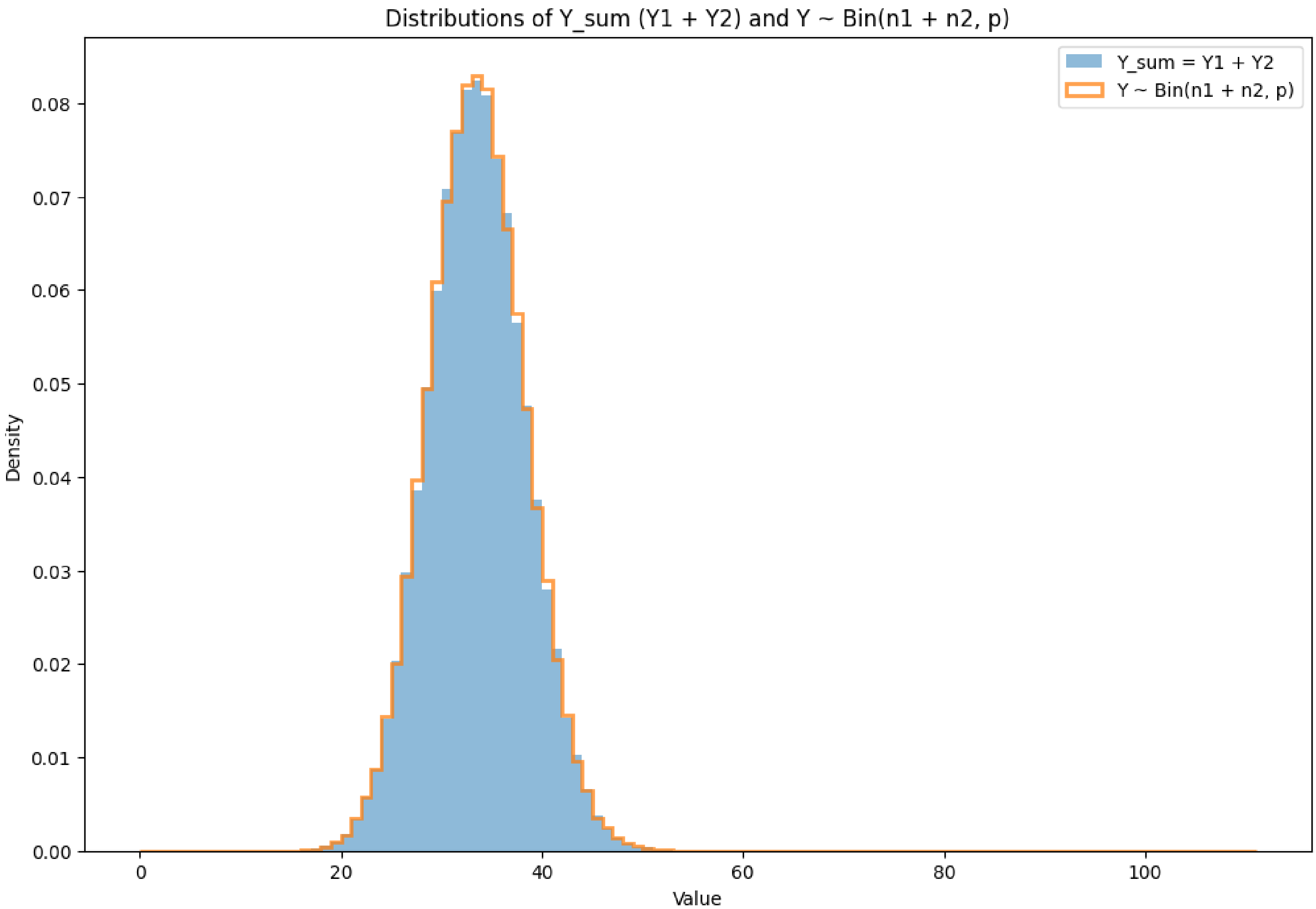
From this probability mass function, we may conclude that $Y \sim Bin(n_1 + n_2, p)$.

Part B

Confirm you answer to part (a) with simulations with $n_1 = 100$, $n_2 = 10$ and $p = 0.3$.

```python
import numpy as np
import matplotlib.pyplot as plt

# Parameters
n1 = 100
n2 = 10
p = 0.3

# Number of simulations
num_simulations = 100000

# Simulating Y1 and Y2
Y1 = np.random.binomial(n1, p, num_simulations)
Y2 = np.random.binomial(n2, p, num_simulations)

# Y = Y1 + Y2
Y_sum = Y1 + Y2

# Simulating Y
Y = np.random.binomial(n1 + n2, p, num_simulations)

# Plotting histograms
plt.figure(figsize=(12, 8))

plt.hist(
    Y_sum, bins=range(n1 + n2 + 2), alpha=0.5, label="Y_sum = Y1 + Y2", density=True
)
plt.hist(
    Y,
    bins=range(n1 + n2 + 2),
    alpha=0.75,
    label="Y ~ Bin(n1 + n2, p)",
    density=True,
    histtype="step",
    linewidth=2,
)
```

```
plt.xlabel("Value")
plt.ylabel("Density")
plt.title("Distributions of Y_sum (Y1 + Y2) and Y ~ Bin(n1 + n2, p)")
plt.legend()

plt.show()
```



Part C

Now suppose $p_1 \neq p_2$. Let

$$Y_3 \sim Bin(n_1 + n_2, \frac{n_1}{n_1 + n_2} p_1 + \frac{n_2}{n_1 + n_2} p_2)$$

Here is an **erroneous** argument for why Y_3 might have the same distribution as $Y_1 + Y_2$ (it doesn't!):

$Y_1 + Y_2$ is the sum of n Bernoulli random variables. Denote these as $X_1, X_2, ... , X_n $ where $n = n_1 + n_2$. Assume they are in order, so that the first n_1 terms are the Bernoulli random variables corresponding to the first binomial distribution (the one with success probability p_1). Note that with this notation we are not specifying whether X_i comes from the first or second Bernoulli sequence. If we randomly select one of these, X_i, then the chance it is equal to 1 is:

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = 1 | i \leq n_1)\mathbb{P}(i \leq n_1) + \mathbb{P}(X_i = 1 | i > n_1)(i > n_1)$$

Now observe that:

$$\mathbb{P}(i \leq n_1) = \frac{n_1}{(n_1 + n_2)}, \quad \mathbb{P}(i > n_1) = \frac{n_2}{(n_1 + n_2)}$$
$$\mathbb{P}(U_i = 1 \mid i \leq n_1) = p_1, \quad \mathbb{P}(U_i = 1 \mid i > n_1) = p_2$$

Plugging these into the formula for $\mathbb{P}(X_i = 1)$ gives the probability of success in the definition of Y_3.

Part D

Explain why this argument above is flawed. **Hint:** Is the variable X_i and X_j independent for all $i \neq j$? If not, why is independence important?

Solution

Part E

Confirm that the argument above is incorrect using simulations, that is, confirm via simulations of an example that Y_3 does not have the same probability distribution as $Y1 + Y2$. You can do this many ways, for example, by plotting $\mathbb{P}(Y_3 > k)$ as a function of k and comparing to $\mathbb{P}(Y_1 + Y_2 > k)$.

Solution