Discrete Probability Distributions in Data Analytics. Statistical Modeling with Discrete Random Variables.

Prepared By: Siman Giri, Instructor: Ronit and Shiv for Herald Center for AI.

Summer, 2025

1 Learning Objectives.

- To help students understand and apply the concepts of discrete random variables, probability mass functions (PMFs), empirical vs theoretical distributions, and basic distribution fitting using real-world data analytics scenarios.
- Use a real-world dataset to apply concepts of discrete random variables, PMFs, empirical vs theoretical distributions, parameter estimation, and model fitting in business analytics scenarios.

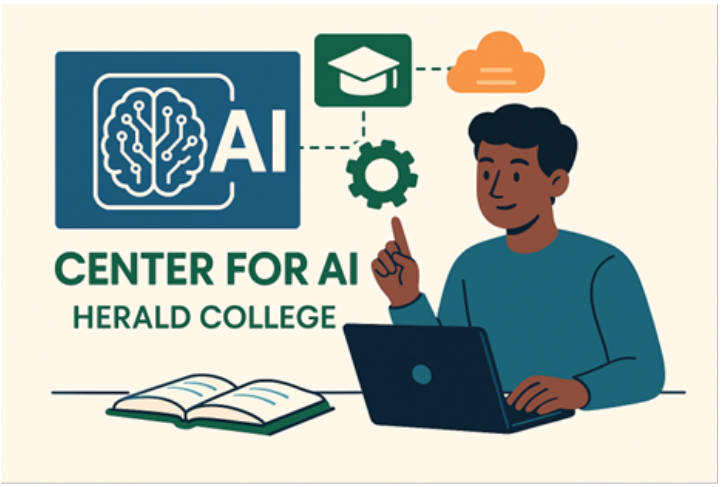

image generated via copilot.

2 Conceptual Understanding [Short Answers].

1. Explain what a random varaible is.Give a real - world business example where mdoeling a random variable helps in decision-making.
2. Define and differentiate between:
   - **Empirical PMF.**
   - **Theoretical PMF.**
3. In a business analytics scenario, explain the difference between:
   - $\hat{\mathbf{p}}$ (sample proportion)
   - $\mathbf{P}(\mathbf{X} = \mathbf{x})$ (probability mass function value)
4. Explain the purpose of a probability distribution.Why is it useful in data analytics?

3 Case - Based Questions.

3.1 Case - Email Marketing Campaign:

You are running a campaign and send emails to 500 batches of 3 users each. You record the number of users who click in each batch.
Suppose the click data across batches is:

Table 1: Clicks per Batch Frequency Distribution

Clicks per Batch	Frequency
0	90
1	180
2	120
3	110

1. Compute the empirical PMF for the number of clicks in a batch.
2. Plot the empirical PMF as bar chart (you may use Excel, Python, or draw by hand.)
3. What is the estimated probability that exactly 2 users click in a batch? Interpret this result.
4. Estimate the average number of users who click per batch (i.e. expected value of the empirical distribution).
5. Suppose you believe the number fo clicks, per batch follows a Binomial Distribution, Estimate parameter $\hat{\mathbf{p}}$ from the data. Justify your choice.

Sample solution - 5:

Scenario: Each email batch consists of 3 users. The click data across 500 batches is given as:

Clicks per Batch (x)	**Frequency (f_x)**
0	90
1	180
2	120
3	110

Step 1: Estimate $\hat{p}$ (Empirical Probability of Click)

The total number of batches is:

$$N = 90 + 180 + 120 + 110 = 500$$

The total number of users:

$$\text{Total trials} = 500 \text{ batches} \times 3 \text{ users} = 1500$$

The total number of clicks:

$$0 \cdot 90 + 1 \cdot 180 + 2 \cdot 120 + 3 \cdot 110 = 0 + 180 + 240 + 330 = 750$$

Estimated probability of a user clicking:

$$\hat{p} = \frac{750}{1500} = 0.5$$

Step 2: Binomial Model Assumption

Assume $X \sim \text{Binomial}(n = 3, p = 0.5)$. The PMF of a Binomial distribution is:

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

For $n = 3$ and $p = 0.5$:

$$P(0) = \binom{3}{0} (0.5)^0 (0.5)^3 = 0.125$$
$$P(1) = \binom{3}{1} (0.5)^1 (0.5)^2 = 0.375$$
$$P(2) = \binom{3}{2} (0.5)^2 (0.5)^1 = 0.375$$
$$P(3) = \binom{3}{3} (0.5)^3 (0.5)^0 = 0.125$$

Step 3: Empirical vs Theoretical PMF

Clicks (x)	**Empirical PMF**	**Theoretical PMF**
0	$\frac{90}{500} = 0.18$	0.125
1	$\frac{180}{500} = 0.36$	0.375
2	$\frac{120}{500} = 0.24$	0.375
3	$\frac{110}{500} = 0.22$	0.125

Conclusion and Justification

- The empirical distribution approximates a Binomial distribution but with some deviations (e.g., $P(2)$ is lower in data).
- The Binomial assumption is reasonable since:
  - Each batch involves 3 independent user trials.
  - All users are assumed to have the same click probability.
- Therefore, modeling with $X \sim \text{Binomial}(n = 3, \hat{p} = 0.5)$ is justified and helps estimate probabilities without listing all outcomes.

3.2 Case - Online Order Fulfillment Center:

You are analyzing operations at a warehouse where customer orders are packaged and shipped. On a given day, the system records how many **packing errors** were found in 600 randomly selected order batches, each containing 5 items.

Error Count Data (per Batch of 5 items):

Table 2: Error Count Distribution per Batch (Size = 5 Items)

Packing Errors in a Batch	Frequency
0	240
1	190
2	110
3	40
4	15
5	5

Questions:

1. Compute the empirical PMF of the number of packing errors per batch.
2. Plot the empirical PMF using a bar chart (you may use Python, Excel, or draw manually).
3. What is the estimated probability that a randomly selected batch has 2 packing errors? Interpret this result in the context of quality control.(Hint: use the empirical mean)
4. You suspect that the number of packing errors follows a Binomial distribution with number of trials $n = 5$ (since each batch has 5 items
   Estimate the unknown parameter $\hat{\mathbf{p}}$(probability of an item being packed incorrectly).
   Justify your modeling choice and compare the empirical PMF with the theoretical Binomial distribution using the estimated $\hat{\mathbf{p}}$.

4 Design, Collect & Analyze Your Own Data

4.1 Objective:

Design a simple data collection process (e.g., survey or observation), gather your own dataset with discrete outcomes, and apply the analysis techniques we've discussed.

4.2 Step -by- Step Instructions:

1. **Design a Survey or Experiment**
   - Choose a topic with measurable, discrete outcomes.
   - Examples:
     - Number of messages sent per day
     - Number of times students use AI tools in a week
     - Whether they clicked on a resource shared in class (0 or 1)
     - Rating of coffee quality in the cafeteria (1 to 5)
     - Number of ads watched before skipping
   - Clearly define:
     - The random variable
     - The question or metric
     - Possible outcomes (finite set)
2. **Collect Data**
   - Aim to collect data from at least 20 people.
   - Record data in a table or CSV format.
   - Ensure the variable is discrete and quantitative or binary.
3. **Analyze Your Data** Apply the concepts covered in class:
   - Create a frequency table and empirical PMF
   - Plot the empirical distribution (bar plot)
   - Compute sample mean and sample variance
   - Choose a simple theoretical model (e.g., Binomial, Poisson) and overlay its PMF
   - Compare and interpret fit (visual comparison only — no parameter estimation required)
4. **Reflect and Report** In a few sentences, explain:
   - What your variable measured and why it matters
   - Any challenges you faced in survey design or data collection
   - Whether your theoretical model fit the data well
   - What you would do differently if you repeated this experiment

4.3 Submission Checklist:

- Survey description and method
- Cleaned dataset (CSV or table)
- Jupyter Notebook with:
    1. PMF computation
    2. Plots
    3. Summary statistics
    4. Reflection and business insight

——————— The - End ———————-