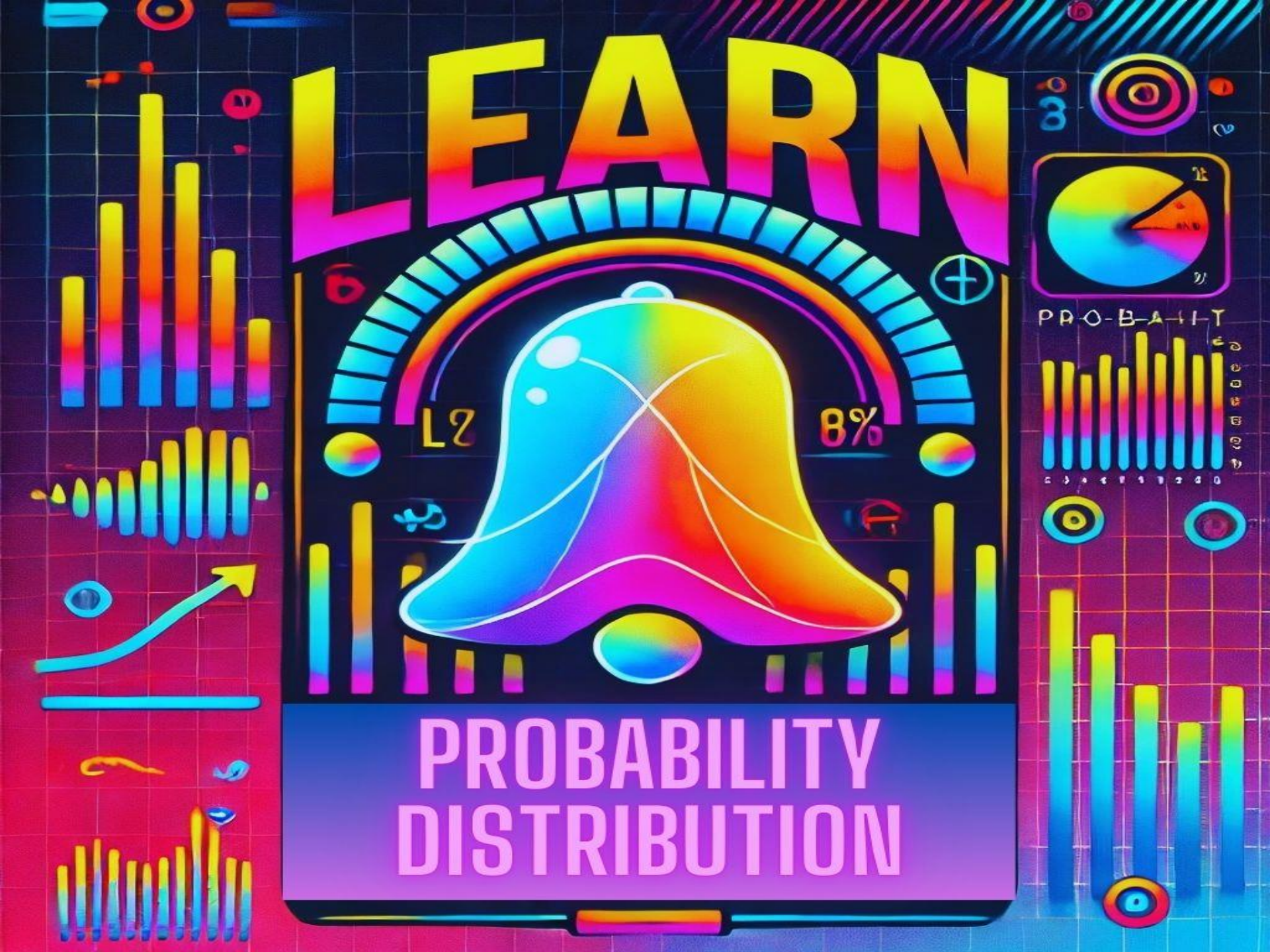
LEARN
PROBABILITY
DISTRIBUTION

Descriptive Statistics: Probability Distribution Concepts Cheat sheet

AN EASY AND ENGAGING EXPLANATION OF DESCRIPTIVE STATISTICS CONCEPTS

- VATSAL SHAH

Algebraic and Random Variables

- **Algebraic Variables:**
  - Think of these like placeholders in math (like x or y) that can hold different values.
  - In algebra, a variable represents an unknown value that can change within the context of a mathematical problem.
  - **For example**, in the equation x + 2 = 5, x is a variable that can be solved to find its value.

- **Random Variables:**
  - In statistics, these are variables that can take on different values based on the outcome of a random event (like rolling a dice).
  - A random variable can be thought of as a way to map the outcomes of a random process to numerical values.
  - **For example**, if we roll a die, we might use a random variable to represent the number that comes up.
  - There are two types of random variables: discrete (finite outcomes) and continuous (infinite outcomes).

Probability Distributions

- **Probability Distribution:**
  - This tells us all the possible outcomes of a random variable and how likely each one is.
  - It provides a function or a table that maps each outcome of the random variable to its probability.
  - **For example**, the probability distribution of rolling a six sided die would list the probabilities of getting each number from 1 to 6.
- **Importance:**
  - Knowing the distribution helps us understand the data's overall behavior and predict future outcomes.
  - **For instance**, if we know the distribution of people's heights in a population, we can predict the probability of a randomly selected person being within a certain height range.
  - It helps in statistical inference, decision making, and hypothesis testing.
- **Types:**
  - There are different types of probability distributions for different types of random variables.
  - Discrete distributions include the binomial and Poisson distributions.
  - Continuous distributions include the normal, exponential, and uniform distributions.
  - Each distribution has its own unique characteristics and parameters.

Probability Mass Function (PMF)

- **PMF:**
  - For discrete variables (like number of heads in coin flips), PMF gives the probability for each possible outcome.
  - It maps each possible value of the discrete random variable to its probability.

- **Conditions:**
  - The probabilities must be nonnegative and add up to 1.
  - This ensures that the PMF is a valid probability distribution.

- **Example:**
  - Suppose we flip a fair coin twice. The random variable X could represent the number of heads observed.
  - X can take on the values 0, 1, or 2. The PMF of X would assign probabilities to these outcomes based on the binomial distribution.
  - **For instance**, the probability of getting 0 heads (TT) is 1/4, 1 head (HT or TH) is 1/2, and 2 heads (HH) is 1/4.
  - The PMF can be represented as a table or a bar graph for better visualization.

Cumulative Distribution Function (CDF)

- **CDF:**
  - This shows the probability that a random variable is less than or equal to a certain value.
  - It provides a running total of probabilities up to a certain point.
- **Mathematical Definition:**
  - For a random variable X and a value x, the CDF F(x) is defined as F(x) = P(X <= x).
- **Example:**
  - If we consider the number of heads in two coin flips again, the CDF would tell us the probability that we get 0, 1, or 2 heads.
  - **For instance**, F(1) would be the probability of getting at most 1 head.
  - The CDF increases as we move from left to right on the number line, starting at 0 and approaching 1.
- **Visualization:**
  - The CDF can be visualized as a step function for discrete variables and as a smooth curve for continuous variables.
  - It is particularly useful in hypothesis testing and determining percentiles in a data set.

Probability Density Function (PDF)

- **PDF:**
  - For continuous variables (like height), PDF shows how the probability is distributed over different values.
  - The PDF itself does not give probabilities directly but rather describes the density of probabilities.
- **Area Under the Curve:**
  - The area under the curve between two points gives the probability of the random variable falling within that range.
- **Example:**
  - For a normally distributed random variable, the PDF is the familiar bell curve.
  - The probability of the variable falling within one standard deviation of the mean can be found by calculating the area under the curve in that range.
  - The total area under the PDF curve is always equal to 1.
- **Uses:**
  - PDFs are used in statistical analyses, such as finding probabilities, expectation, variance, and more.
  - They are also crucial in fields like machine learning, signal processing, and financial modeling.

Density Estimation

- **Density Estimation:**
  - This is about figuring out the PDF from data.
  - It involves estimating the underlying distribution of a set of data points.
- **Methods:**
  - **Parametric:** Assume the data follows a known distribution (like normal distribution). These methods use parameters (like mean and standard deviation for normal distribution) to estimate the density.
  - **Non Parametric:** No assumptions about the distribution. Methods like Kernel Density Estimation (KDE) smooth the data to estimate the PDF directly from the data points.
- **Example:**
  - KDE can be used to create a smooth curve that approximates the PDF of the heights of people in a sample, without assuming it follows any specific distribution.
- **Applications:**
  - Density estimation is used in various fields such as machine learning, data analysis, and econometrics.
  - It helps in visualizing the distribution of data, detecting anomalies, and performing hypothesis testing.
- **Techniques:**
  - Common techniques include histograms, KDE, and Gaussian Mixture Models (GMM).
  - The choice of method depends on the nature of the data and the specific requirements of the analysis.