Understanding Distribution in Data Science

Unveiling the power of Data Distribution

“As a data scientist whatever we are learning about mathematical concepts ,we must try relate in context of data understanding, means how such information enhancing our data understanding skills”

Introduction

- Welcome to the presentation on Understanding Distributions in Data Science!
- Today, we'll explore how distributions play a crucial role in data science and why understanding them is important.
- By the end, you'll have a grasp of discrete and random distributions and how they relate to real-world data.

What are Distributions?

- ❖ Distribution refers to the way data is spread out or distributed across different values.

- ❖ Think of it like a recipe: just as ingredients are distributed in different proportions to make a dish, data points are distributed in different ways in a dataset.

- ❖ In data science understanding the distribution of data is fundamental to draw meaningful insights and making informed decisions.

Type of Random Variables

- As distribution talks about data so let us first understand about data points or variables.
- Broadly classified in two category as :

- **Discrete Variables** (Takes Isolated values)
  - No. on a Dice Roll
  - Number of items

- **Continuous Variables** (It contains interval(s))
  - Height
  - Weight
  - Salary
  - These random variables makes the base for classification of distributions as Discrete and Continuous distribution.

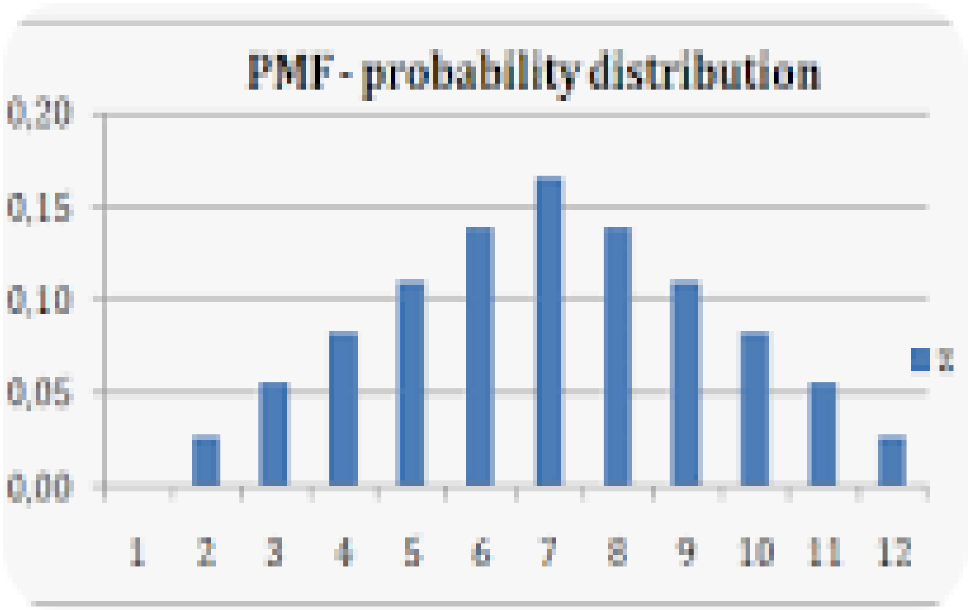
Discrete Distributions

- Discrete distributions are for data that can only take specific, separate values (like integers).
- Examples include coin flips and dice rolls.
- They're described by a Probability Mass Function (PMF), which tells us the probability of each possible outcome.
- Types: Bernoulli Distribution, Binomial Distribution, Poisson Distribution.

Understanding Probability Mass Function (PMF)

- PMF is like a recipe card that tells us the likelihood of getting each outcome.
- If we're rolling a fair six-sided die, the PMF tells us there's a 1/6 chance of rolling each number.
- Visually, PMFs are often represented as bar graphs, where each bar's height represents the probability of a particular outcome.

Probability Mass Function



Random Distributions

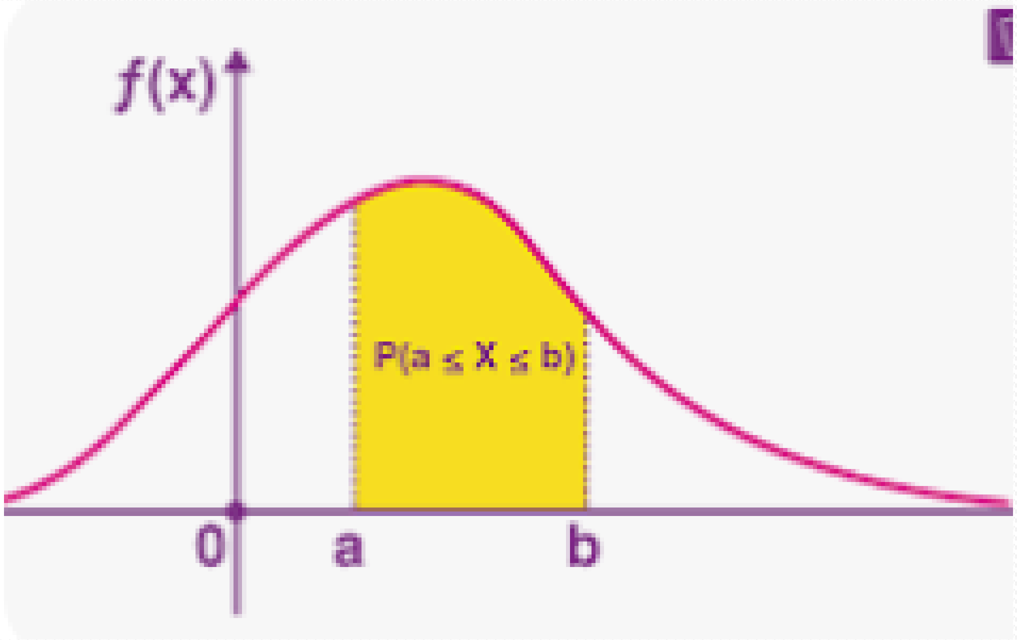
- Random distributions are for continuous data (like heights or weights) or when data can take any value within a range.
- Examples include the Normal Distribution and the Poisson Distribution.
- They're described by a Probability Density Function (PDF), which tells us the likelihood of a range of outcomes.

Probability Density Function (PDF)

- PDF is like a smooth recipe that gives us the probability density at each point.
- Describes the likelihood of continuous random variable. Where likelihood refers to the chance or probability of something happening. It’s a measure of how probable an event or outcome is.
- So when we say ‘likelihood’ we are essentially talking about how likely it is for something to occur.

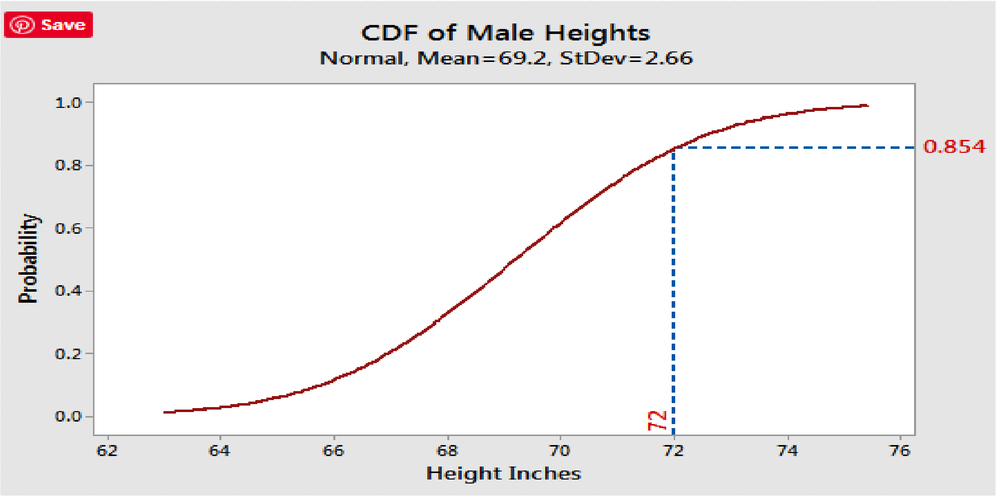
Probability Density Function

- Visually, PDFs are often represented as smooth curves.



Cumulative Distribution Function

- CDF gives the probability that a random variable is less than or equal to a certain value.
- It provides a complete description of the distribution of random variable.



Some Important Distributions in context of Data Science

- **Normal Distributions** :Very popular in data science as we usually assume data is normally distributed.(in case of continuous data).
- **Binomial Distributions**: In number of trials are finite and only two outcomes. (in case of discrete outcome specially in binary).
- **Poisson Distributions**: if we know the occurrence per interval or space.
- **Geometric Distributions**: When we are interested to know that how many trials are needed to get first success in a sequence of independent Bernoulli trials.

Concept of Permutation and Combination

- In general we come across two type of selection in our daily life. For example:
- ✓ let we have five fruits (apple banana, mango, papaya, orange), if we need to arrange them without taking care of duplicate arrangement than we can say there are 5*4*3*2*1 = 120 ways to select 120.
- ✓ While if we are looking for a selection without duplicates than there is only one set of arrangement is possible.
- The case one mathematically known as permutation($_np_r$).

Where n = number of objects , P = permutation, r = number of position.

Concept of Permutation and Combination

- nPr = n!/(n-r)!,where ‘n’ is number of trials and ‘r’ is number of success in ‘n’ trials
- if r is equal to n than number of selection will n! or n factorial.(Where factorial means sequential product of n from n to 1 like n*n-1*n-2*n-3*……….1.)
- While case two mathematically known as combination which talks about unique set. Shown as (nCr)
- nCr = n!/r! (n-r)!
- In general we say combination is subset of permutation.

Central Limit Theorem:

- States that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases regardless of the shape of the population.

Central Limit Theorem Formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample Mean = Population Mean = μ

$$\text{Sample Standard Deviation} = \frac{\text{Standard Deviation}}{n}$$

OR

$$\text{Sample Standard Deviation} = \frac{\sigma}{\sqrt{n}}$$

Expected Mean/Variance/Standard Deviation

- Expected Mean represent the average value of random variable. Denoted as E(X) or as μ.
- Variance/Standard Deviation measure the spread or dispersion of the distribution of a random variable.
- Variance is denoted as Var(X).
- Standard Deviation as σ.

Normal Distributions

- In order of usage as well popularity ‘Normal Distribution’ is first to be understand. Also known as Gaussian Distribution or Bell curve.
- It is one of the most widely used distribution in data science due its occurrence in nature and real world phenomena.
- It talks about symmetric around its mean and has a single peak at the centre of the distribution.

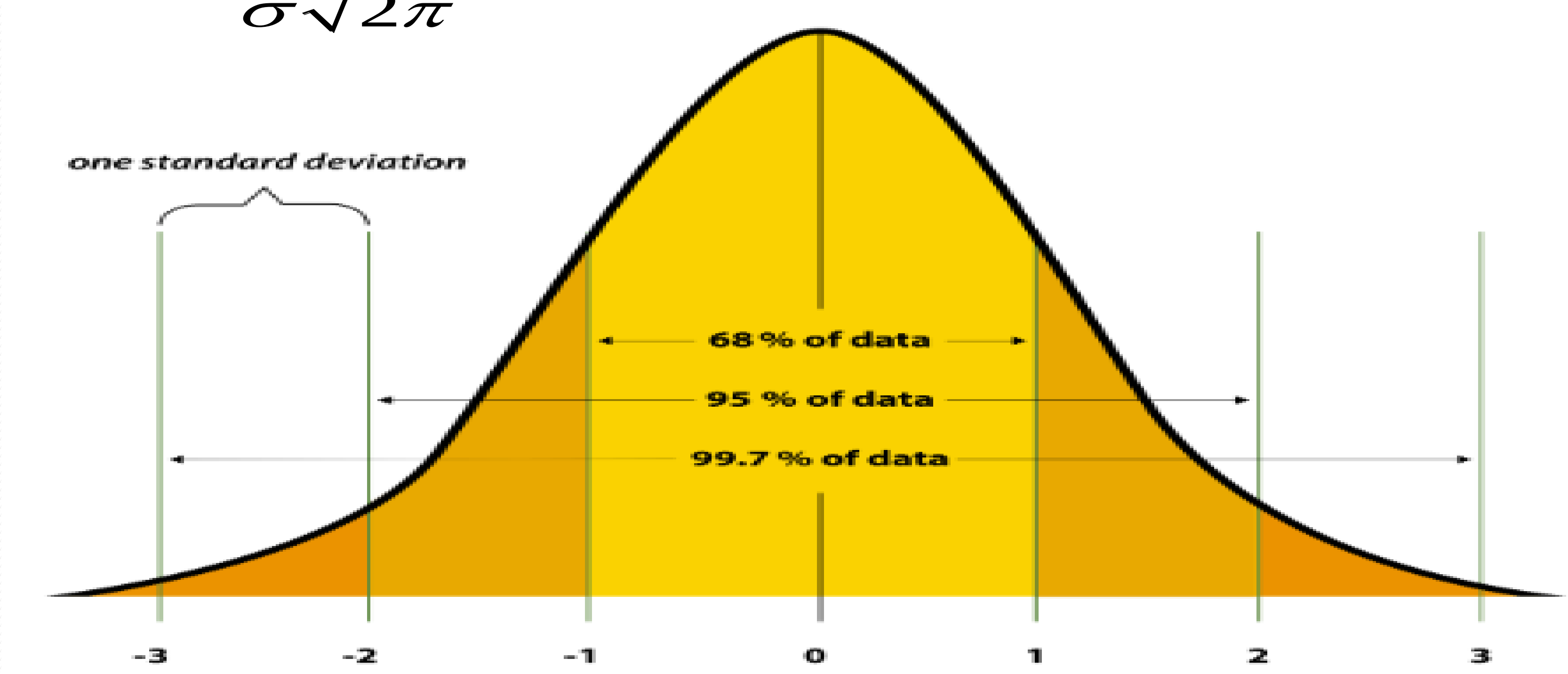
Normal Distributions

- The spread or dispersion of the distribution is determined by the standard deviation.
- The majority of the data falls within one standard deviation(68%).
- The mean ,median ,mode of normal distribution are equal and located at the centre of the distribution.
- It is used for random continuous variables.
- Any PDF curve that look like a bell shape is normally distributed.
- As variance increase peak of the curve fall and gets flatter.

Normal Distributions

- It is assumed that 99.7% of all the values will fall with in 3* S.D. of the mean on either side on curve.
- The pdf of normal distribution is given by;

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Standard Normal Distributions

- It is a special normal distribution with mean value as ‘zero’ and standard deviation as ‘one’.
- It is also called Z distribution.
- Any normal distribution can be standardized by converting its values into Z scores.
- Z scores tell you how many standard deviations from the mean each value lies.
- Z = (X – mean)/ standard deviation.
- The SND is a probability distribution, so the area under the curve between two points tells the probability of variables taking on a range of values.
- The total area under the curve is 1 or 100%.

How to find whether distribution is normal or not

- Histogram: bell shape curve
- QQ plot: straight line
- Mean = Median =Mode

Poisson Distributions

- In data science Poisson distribution is a probability distribution that is commonly used to model the number of events that occur within a fixed interval of time or space, given the average rate of occurance.
- the **Poisson Distribution** is a discrete probability distribution.
- **Count Data**: the distribution is particularly useful for modelling count data. Where number of occurance of an event is being measured.
- Eg: number of customer arrivals at a store within a given hour.
- Number of website visits in a day.
- Number of phone calls received by a call center in an hour.

Poisson Distributions

- **Rare Event:** Occurrence of natural disaster
- Occurrence of defects in manufacturing.
- Number of accidents at a particular intersection.

- The Poisson distribution has only one parameter, called λ.
- The mean of a Poisson distribution is λ.
- The variance of a Poisson distribution is also λ.

Poisson Distributions

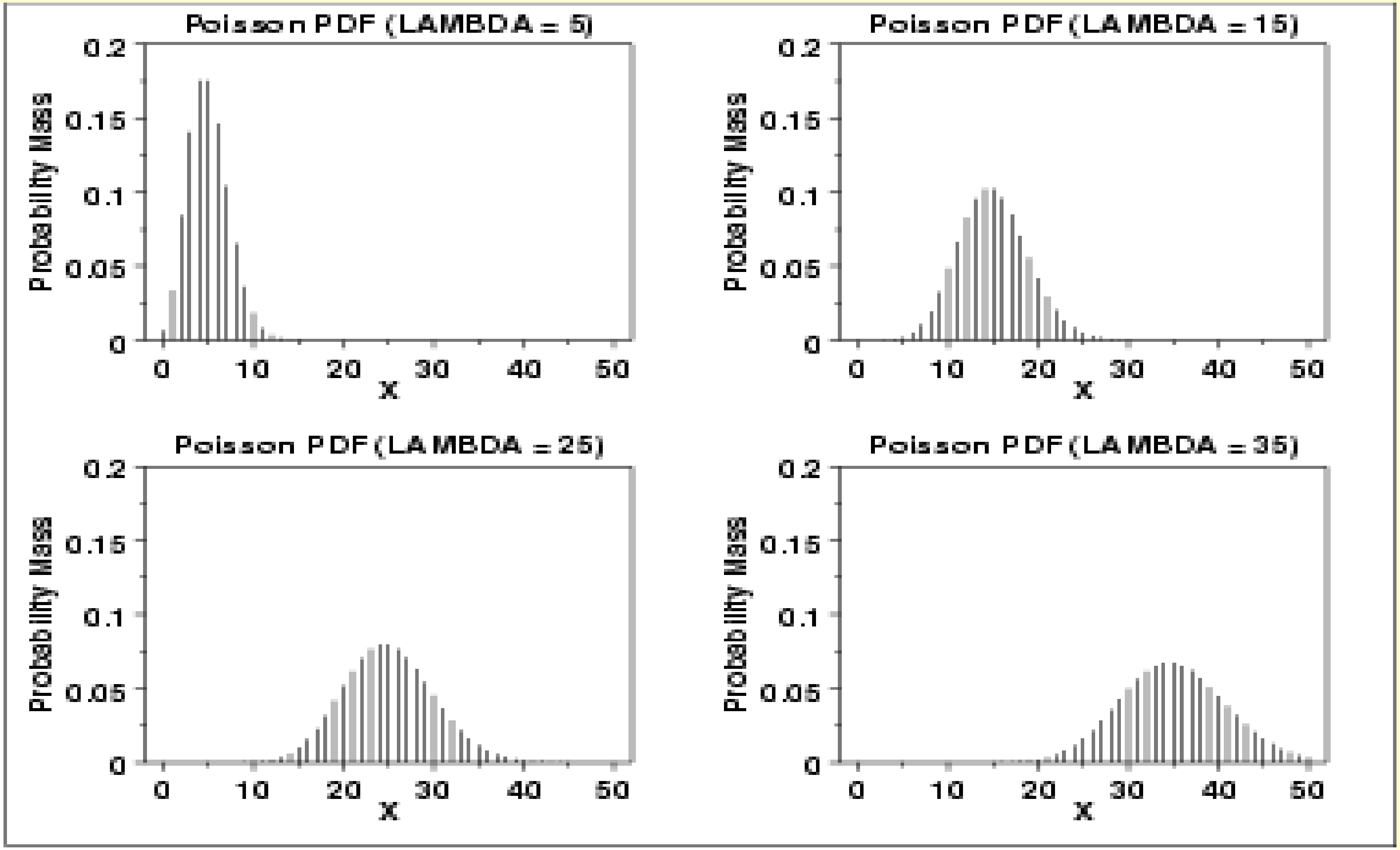
- The probability mass function of the Poisson distribution is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Where:
X is a random variable following a Poisson distribution
k is the number of times an event occurs
P(x=k) is the probability that an event will occur *k* times
e is Euler's constant (approximately 2.718)
λ is the average number of times an event occurs
! is the factorial function

- **For count data** use poisson regression as a machine learning algorithm.

Poisson Distributions



Binomial Distributions

- Binomial distribution is fundamental concept in both stats and data science when dealing with binary outcomes or counting the number of successes in a fixed number of trials.

- In data science its often used in scenario like click through rates in online advertising,conversion rate in marketing campaigns or out come of medical treatments.

Binomial Distributions

- In this distribution number of trial are fixed and finite.
- The probability of success is the same for each trial.
- The shape of distribution depends upon n,p.
- The closer to 0.5 the shape is more symmetrical.
- n = number of trials
- p = success

Generalised Equation of Binomial Distribution

$$P(x; p, n) = \binom{n}{x} (p)^x (1-p)^{(n-x)} \quad \text{for } x = 0, 1, 2, \cdots, n$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

n is the number of trials (occurrences)
x is the number of successful trials
p is the probability of success in a single trial
${}^{n}C_{x}$ is the combination of n and x.
E(X) = np
Variance = np(1-p)

Geometric Distribution

- When we are interested to know that how many trials are needed to get first success in a sequence of independent Bernoulli trials.(Where Bernoulli trials is a random experiment with only two possible outcomes eg: Flipping a coin, Rolling a dice or working of a bulb.)
- It has a distinctive shape showing max probability of success in Ist trial and sequentially reducing as number of trial increases.
- Expectation(E) = 1/p
- Variation V(x) = q/p
- P(x = r) = p.q**(r-1)