Statistics for Data Science

Let's Talk Statistics

Statistics is a method for summarizing info/data insight in order to draw conclusions or for decision-making

Statistics is widely use in our daily life and how we represent our data

Statistics is useful in data science for purposes like data collection, data preparation, data transformation and data interpretation

Dimensions of Statistics

Descriptive Statitistics

Method of summarizing information. This can be in form of graphs or numbers

Inferential Statitistics

is about drawing conclusions about a population on the basis of a limited number of cases(sample)

Descriptive Statistics

Univariate analysis , Bivariate analysis and Multi-variate analysis

Univariate analysis is used to describe the distribution for a single variable with use of graphs, centre tendency, dispersion and distribution shape

Bivariate/Mutivariate analysis is used to understand the relationship between two or more variables. Popularly used methods is scatterplot and correlation

Some of the above tools are also useful for outlier detection.

Refresher Questions

- What are the dimensions of statistics?
- Descriptive Statitics can be split into three analysis areas. What are they?

Things to take note of

Apart from some new python syntax

Tips of some data processing steps

like the data shape etc

Central Tendency and Dispersion

Mean, Mode, variance etc

Outlier Detection method

Boxplot etc

Others

Data type, correlation etc

Types of Data

Categorical data and Numerical Data

Categorical Data can be divided into : Nominal (Gender) and ordinal (Position rank)

Numerical data can be divided into : Discrete (no of people) and Continous (height & weight)

Discrete - Set of seperate/whole numbers
Continous - Infinite ranges of values

Summary Statistics

Central Tendency

It can seen as the center value of a dataset. The goal is to best describe a dataset with a single value/number e.g mean, mode and median

Dispersion

describes how spread out the datasets are from each other

Mean

average of a dataset

Median

the middle value

Mode

The mode is the most frequent value.

Refresher Questions

- Can you think of practical scenarios where the mean, median,mode would be a valuable measure?

- Can you think of scenarios where you would use the median over the mean and vice versa?

Tip: Mean and Median for Numeric Data. Mode for Categorical Data

Range

The difference between the max and the min value

Variance

the square of the sum of deviation

Standard Deviation

The squareroot of the sum of deviation

Refresher Questions

- Why do we need measures of dispersion, what role do they play in helping us understand our data?

- What is the difference between standard deviation and variance?

Tip: Because of this squaring, the variance is no longer in the same unit of measurement as the original data. Taking the root of the variance means the standard deviation is restored to the original unit of measure and therefore much easier to interpret.

Correlation co-efficient

measures how strong a relationship is between two variables

This co-efficient ranges from -1 to 1

Refresher Questions

- Does anyone want to discuss everyday examples of a correlation and the value of understanding correlation

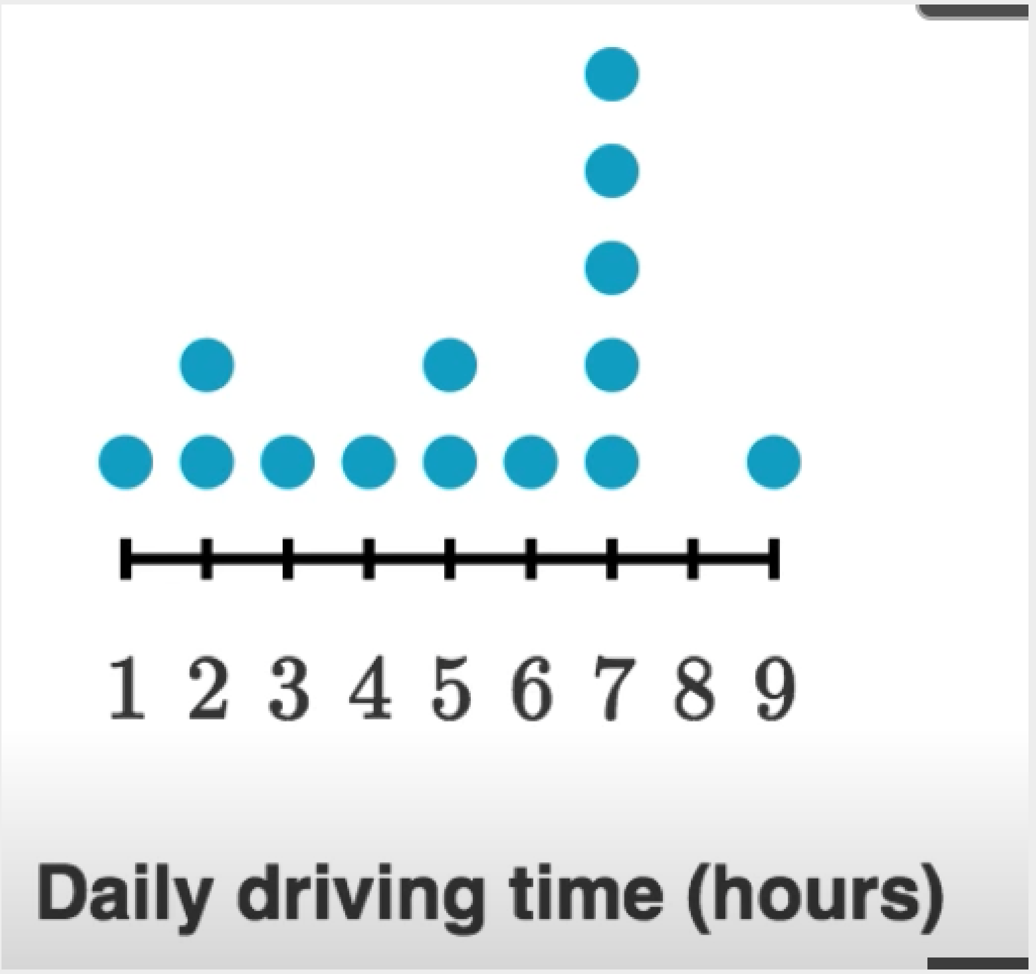
- Does correlation imply causation?

Statistics Part 2

- Percentile
- Probability Mass Function
- Cummulative Distribution Function
- Distributions

Review on Data Types/Percentiles

Percentile is the percentage of the data that is below the amount in question or at or below the amount in question



Probability / Probability Distribution

Probability is the likelihood of something happening or an evening. Probabilities are between 0 and 1. Frequency measure how often an event occurs.

Probability Distribution describes the probabilities of occurance of multiple events.

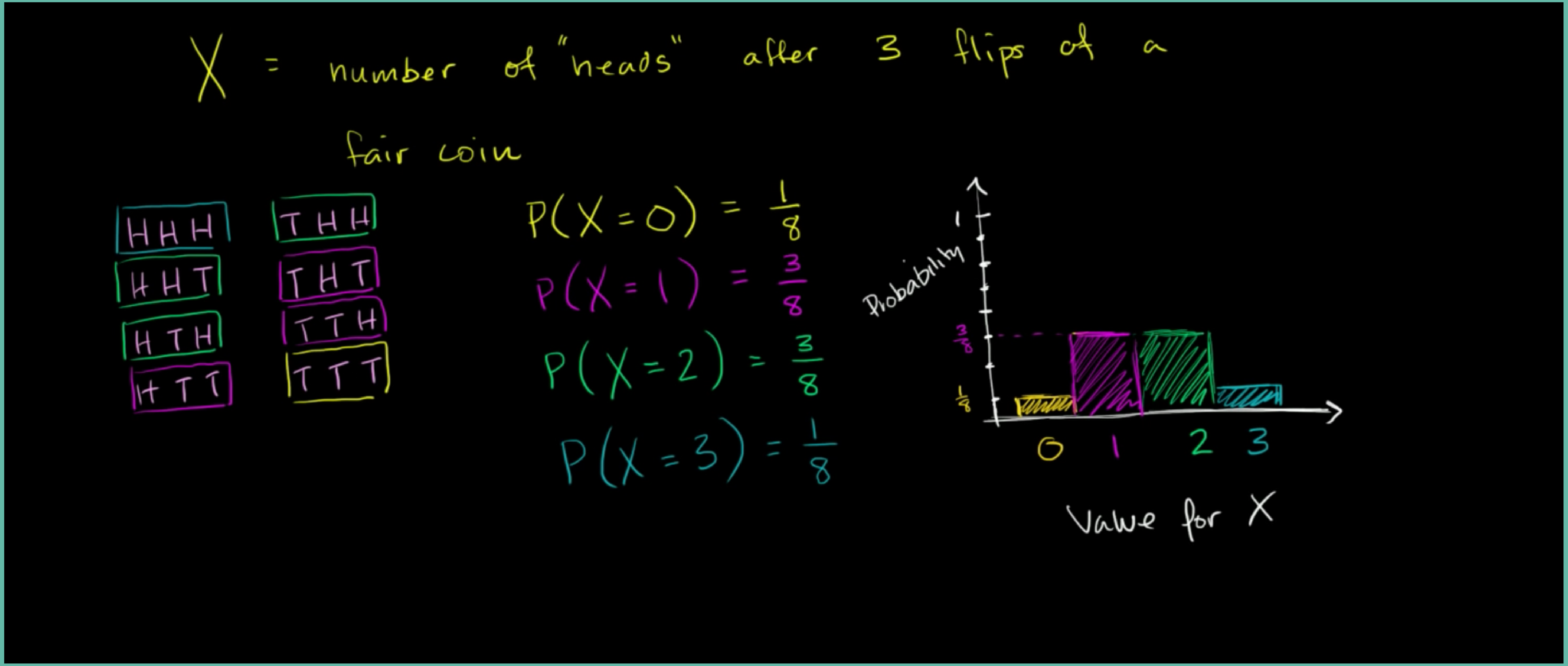
To define probability distributions for the specific case of random variables (so the sample space can be seen as a numeric set), it is common to distinguish between discrete and continuous random variables.

Probability mass functions gives you probabilities for discrete random variables. While Probability density function gives you probabilities for continous random variables (X).

The Cumulative distribution function (CDF) of a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x

Probability / Probability Distribution

Probability Mass Distribution



X = number of "heads" after 3 flips of a fair coin

HHH	THH
HHT	THT
HTH	TTH
HTT	TTT

$P(X=0) = \frac{1}{8}$

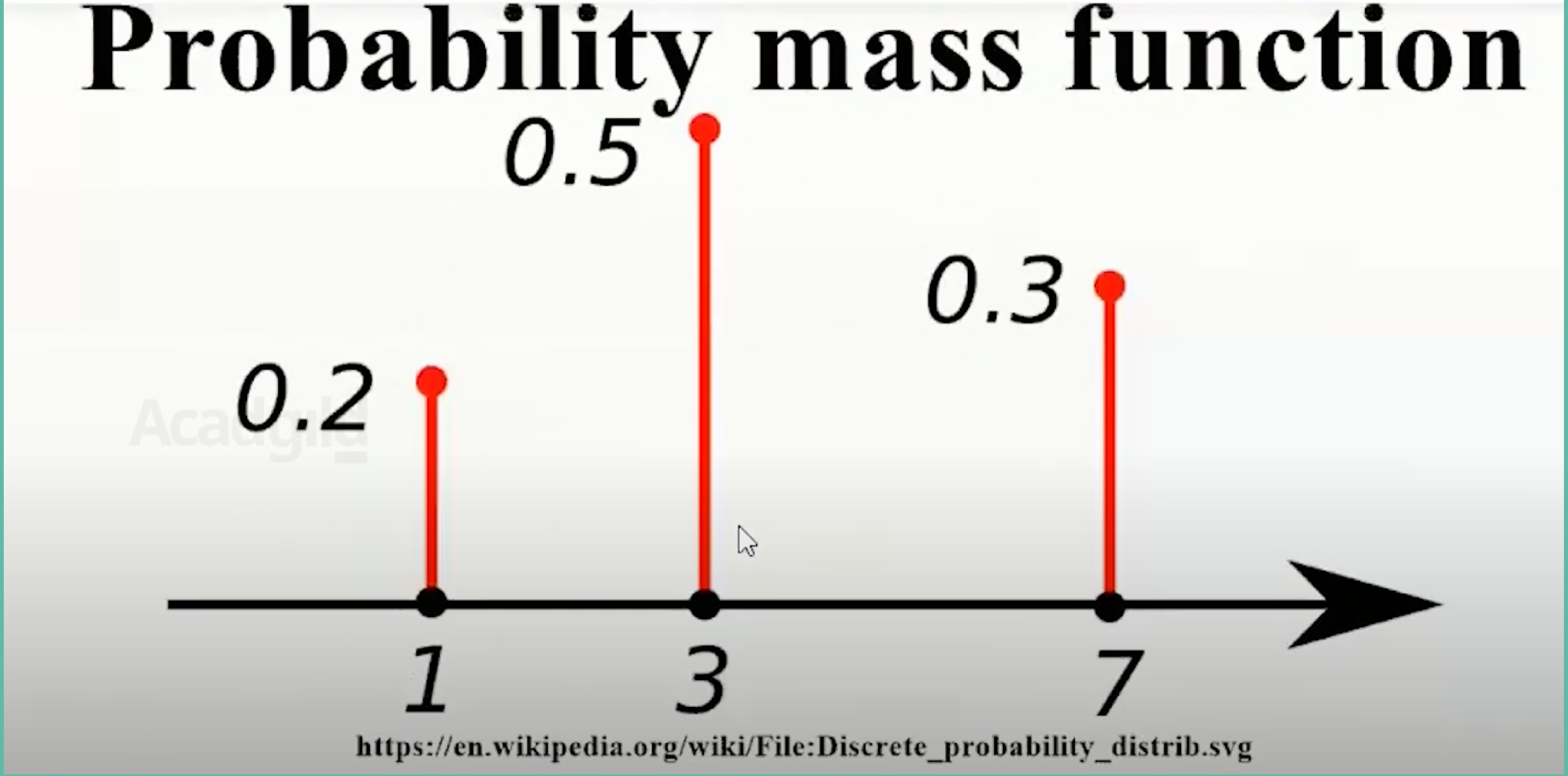
$P(X=1) = \frac{3}{8}$

$P(X=2) = \frac{3}{8}$

$P(X=3) = \frac{1}{8}$

Probability / Probability Distribution

Probability Mass Distribution



https://en.wikipedia.org/wiki/File:Discrete_probability_distrib.svg

Types of Probability Distribution

Distribution for Continous Data

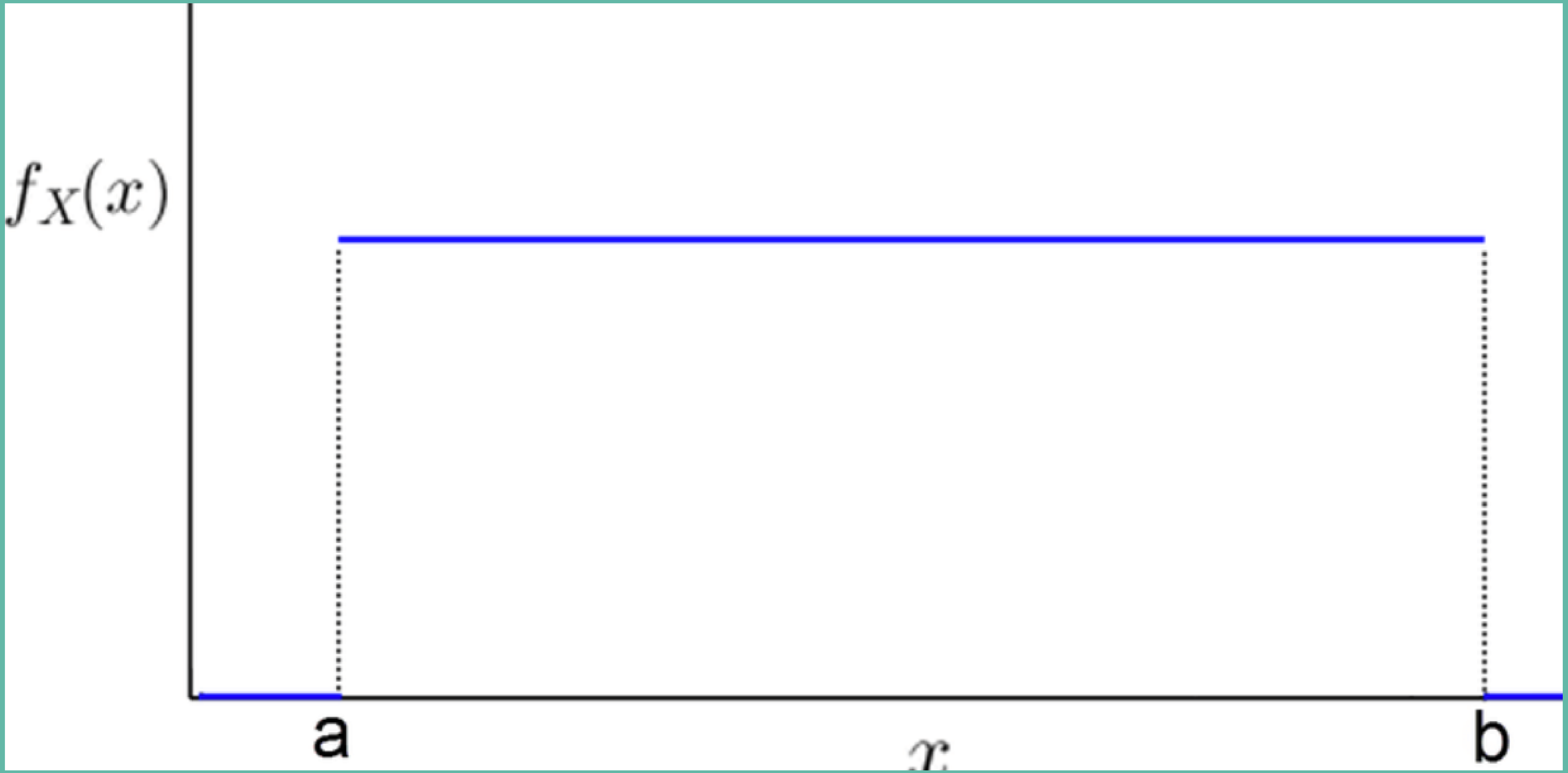
- Uniform Distribution
- Normal distribution(gaussian)
- Exponential Distribution

Distribution for Discrete data

- Uniform Distribution
- Bernoulli Distribution
- Binomial Distribution
- Poisson Distribution

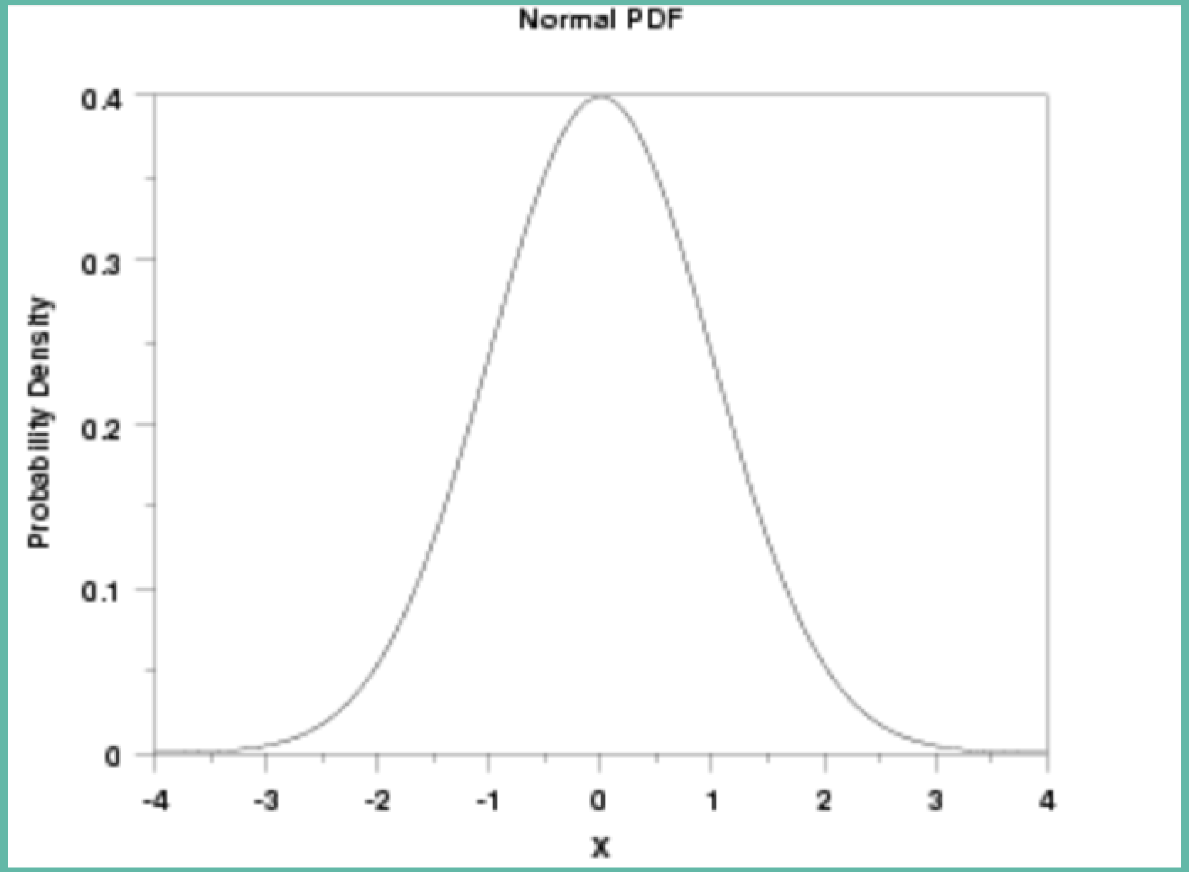
Uniform Distribution

A uniform distribution can either be discrete or continuous. In a uniform distribution, every event is equally likely to occur. If you think of the die, each of the six outcomes has an equal probability of occurring at 1/6



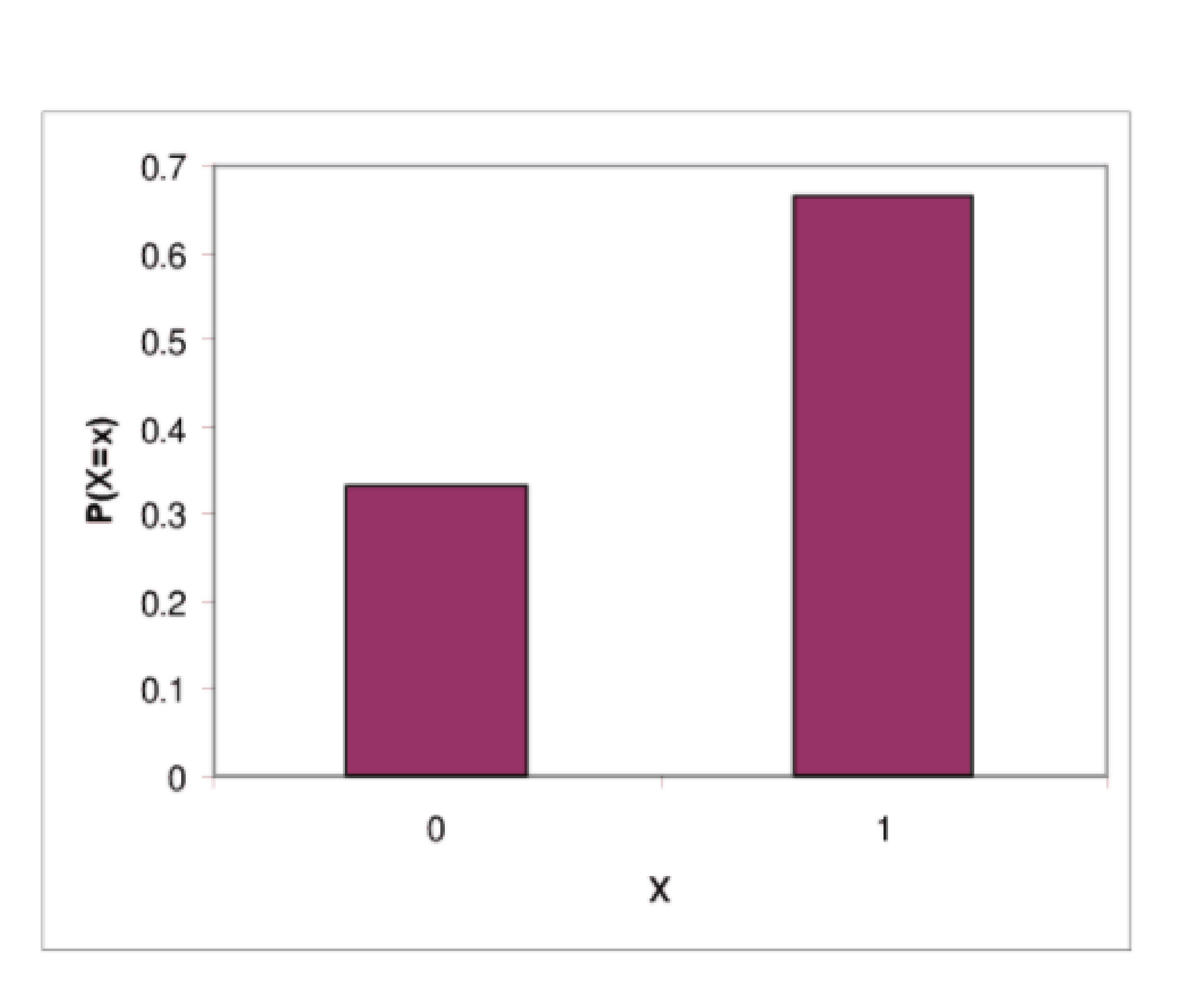
Normal Distribution /Gaussian

The normal distribution is the most commonly seen continuous distribution in nature. In the normal distribution the mean, median, and mode all line up such that the center of the distribution is the mean. Because of this, exactly half of the results fall to either side of the mean. The normal distribution is also identifiable by its bell shape and may sometimes be referred to as a bell curve.



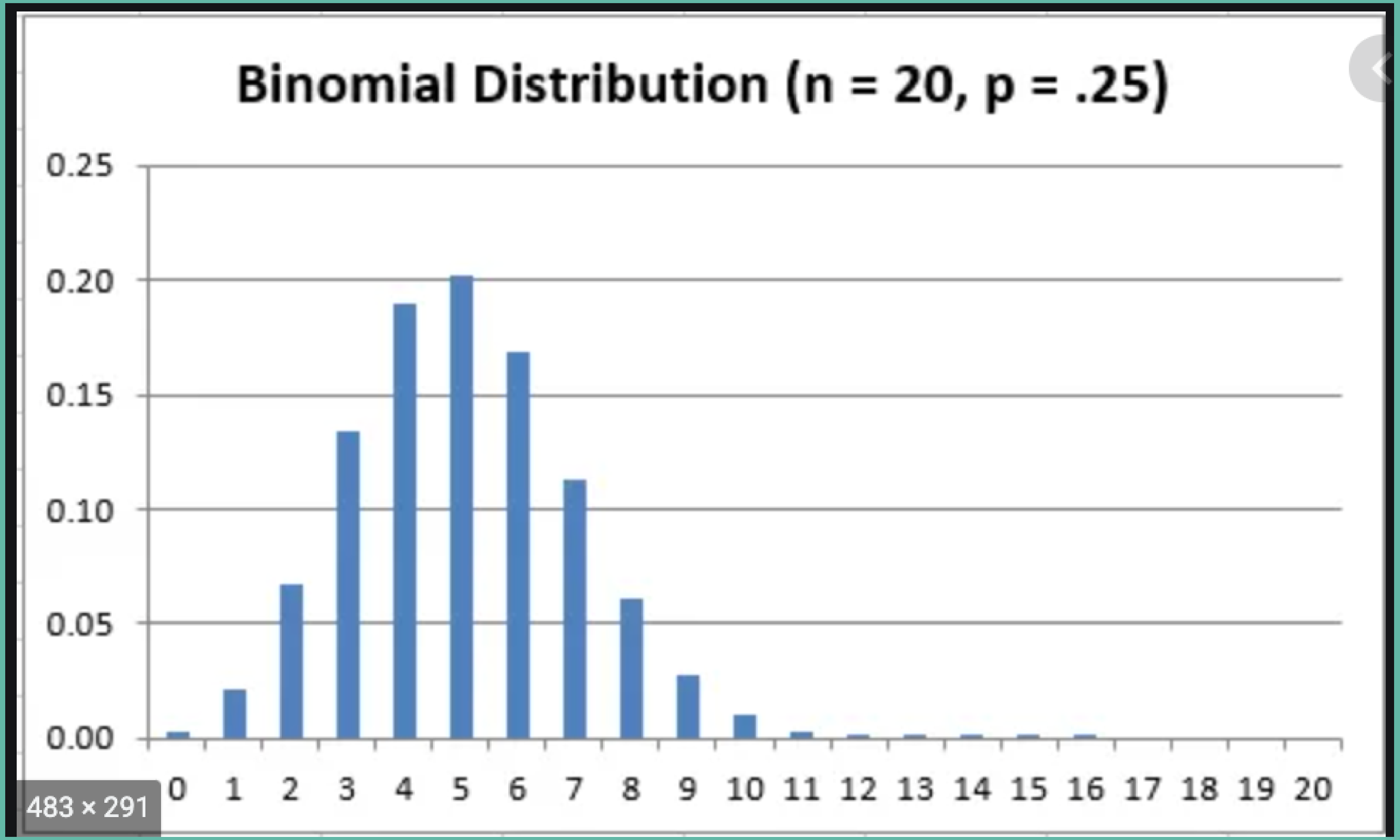
Bernoulli Distribution

The Bernoulli distribution is a kind of discrete probability distribution- a random trial that has two results(outcomes). The results are classified as successes or failure. Here the number of trials is 1 i.e n is 1. outcome is 2. e.g a single toss of coin



Binomial Distribution

The binomial distribution is a type of discrete distribution. It is just like a bernoulli distribution in terms of outcomes but with n number of trials where n > 1.
We then measure the number of successes over these n trials.



Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate. Poisson has to do with an element of time.(l)

This answers questions such as:
How many customers will come through the door in an hour?
How many pages will have typos in a book?
How many calls will a call center receive in a day?

Exponential Distribution

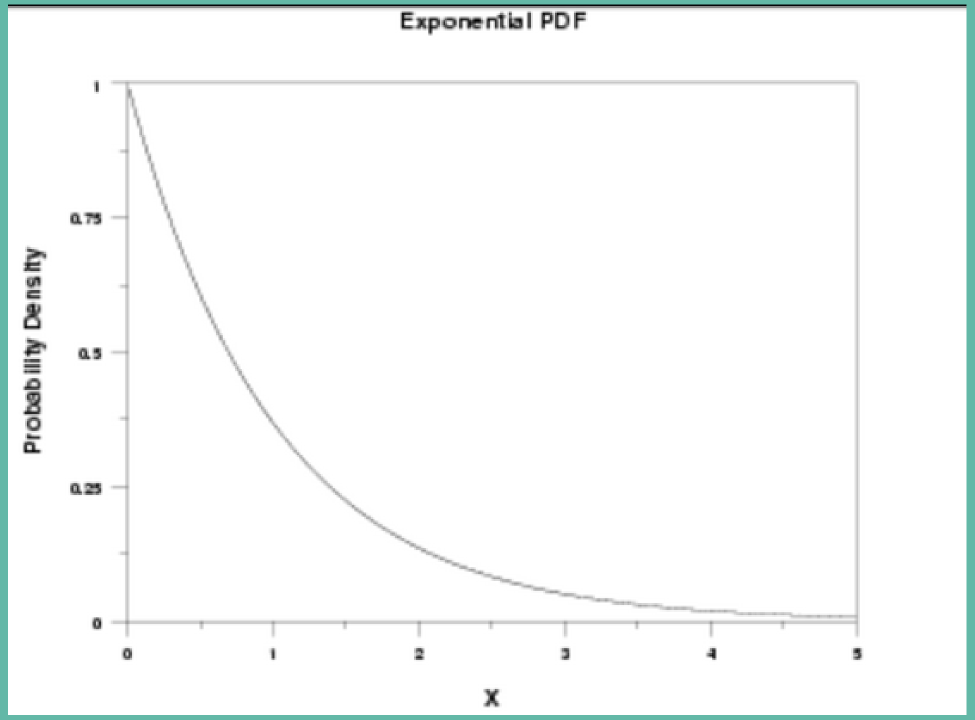
The probability of an amount of time passing before an event occurs or between two events occuring/events in a Poisson point process.

This answers questions such as:

The amount of time until the next bus arrives

The length of time I have to wait for my food at a restaurant

How long until I have to replace my computer


Exponential PDF
Probability Density
X