MODULE 2

Probability is the branch of mathematics that deals with measuring and quantifying uncertainty.

Probability distributions describe the likelihood of the possible outcomes of a random event and can be discrete or continuous.

Two types of probability:

- **Objective**
    - Based on statistics, experiments, and mathematical measurements
- **Subjective**
    - Based on personal feelings, experience, or judgment.
    - Does not involves formal calculations, statistical, or scientific experiments

Two types of objective probability:

- **Classical**
    - Based on formal reasoning about events with equally likely outcomes
    - $Classical\ probability = \frac{Number\ of\ desired\ outcomes}{Total\ number\ of\ possible\ outcomes}$
- **Empirical**
    - Based on experimental or historical data
    - Represents the likelihood of an event occurring base on the previous results of experiment of past events
    - $Empirical\ probability = \frac{Number\ of\ times\ a\ specific\ event\ occurs}{Total\ number\ of\ events}$

Fundamental concepts of probability

The probability that event will occur is expressed as number between 0 and 1

- 0 = 0% chance that event will occur
- 1 = 100% chance that event will occur
- 0.5 = 50% chance that event will occur

Random experiment – A process whose outcome cannot be predicted with certainty

All random experiments have **three things in common**:

1. The experiment can have more than one possible outcome
2. You can represent each possible outcome in advance
3. The outcome of the experiment depends on chance

Probability notation

- The probability of event A is written as P(A).
- The probability of event B is written as P(B).

Three basic rules of probability:

1. **Component rule**
   - The event not occurring
   P(A’) = 1 – P(A)
2. **Addition rule**
   Mutually exclusive events
   - They cannot occur at the same time

P(A or B) = P(A) + P(B)

3. **Multiplication rule**
   Independent events
   - The occurrence of one event does not change the probability of the other even
   P(A and B) = P(A) * P(B)

Conditional Probability
Probability of an event occurring given that another event has already occurred

$$P(A\ and\ B) = P(A) * P(B|A)$$

Dependent Events
Two events are dependent if the occurrence of one event changes the probability of the other event. The first event affects the second event.

Bayes' Theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Prior probability P(A) – The probability of an event before new data is collected
Posterior probability P(A|B) – The updated probability of an event based on new data

Bayes' Theorem (expanded version)

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|not\ A) * P(not\ A)}$$

Used for:
1. Medical diagnostic tests
2. Quality control tests
3. Software tests

False positive – Test results that indicates something is present when it really is not
Example: Antivirus software falsely indicate there is a virus even though it's actually safe

False negative – Test result that indicates something is not present when it really is
Example: Spam email filter incorrectly identify a spam email as a legitimate

Random variable - Represents the values for the possible outcomes of a random event

1. **Discrete**
   - Has a countable number of possible values
   - Count the number of outcomes
2. **Continuous**
   - Takes all the possible values in some range of numbers
   - No limit to the number of possible values. Example: Person's height
   - Measure the outcome

Probability Distribution - Describes the likelihood of the possible outcomes of a random event

- **Discrete distributions**
  - **Uniform**

- **Binomial**
- **Bernoulli**
- **Poisson**

- **Continuous distributions**
  - **Normal distribution**

  - Represent continuous random variables
  - Tells the probability that the variable takes on a range of values (intervals)
  - Probability that the variable is exactly any single value is 0
  - Represented as bell curve (normal distribution)

Sample space
The set of all possible values for a random variable

Sample space for a single die roll = {1, 2, 3, 4, 5, 6}

<u>Binomial distribution</u>
A discrete distribution that models the probability of events with only two possible outcomes, success or failure

$$P(X = k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

Where: k = number of success, n = number of trials, p = probability of success on a given trial
Assumes:

- Each event is independent
- Mutually exclusive
- The probability of success is the same for each event

Used in:

- Medicine (new medication causes side effects or not)
- Banking (the credit card is fraudulent or not)
- Investing (stock price rises or falls in value)
- Machine learning (classify data)

Binomial experiment
Attributes:

1. Consists of a number of repeated trials
2. Each trial has only two possible outcomes
3. The probability of success is the same for each trial
4. Each trial is independent

Example of binomial experiment is 10 repeated coin tosses

<u>Bernoulli Distribution</u>
Similar to binomial distribution but the difference is that the Bernoulli distribution only refers to only a single trial of an experiment

<u>Uniform Distribution</u>
Describes whose outcomes are all equally likely, or have equal probability.

Example: Rolling a die can result in equal probability outcomes which are 1-6 each with 16.7% probability.

<u>Poisson Distribution</u>
Models the probability that a certain number of events will occur during a specific time period or space (distance, area, volume)

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

Where: λ = mean number of events that occur during a specific time period, k = number of events

Used in:

- Calls per hours for a customer service call center
- Visitors per hour for a website
- Customers per day at a restaurant
- Severe storms per month in a city

Poisson experiment attributes:

- The number of events in the experiment can be counter
- The mean number of events that occur during a specific time period is known
- Each event is independent

<u>Normal distribution</u>
A continuous probability distribution that is symmetrical on both sides of the mean and bell-shaped.

Also known as Gaussian distribution.

The distance of a data from the mean measured in standard deviations

Normal distributions features:

- The shape is a bell curve
- The mean is located at the center of the curve
- The curve is symmetrical on both sides of the mean
- The total area under the curve equals 1

Empirical rule:

- 68% of values fall within 1 standard deviation of the mean
- 95% of values fall within 2 standard deviation of the mean
- 99.7% of values fall within 3 standard deviation of the mean

Most data professionals considers 3 std as an outlier

Two types of probability functions:

- **Probability Mass Functions (PMFs)** to represent discrete random variables
- **Probability Density Functions (PDFs)** to represent continuous random variables

Z-Score

- Measure of how many standard deviations below or above the population mean a data point is.
- Helps to standardize the data.
- Useful to get an idea of how an individual value compares to the rest of the distribution.

$$Z = \frac{x - \mu}{\sigma}$$

Where: x = data value, μ = mean, σ = standard deviation

Used in application of anomaly detection:

- Fraud financial transactions
- Flaws in manufacturing products
- Intrusions in computer network