UNIT-3

PROBABILITY AND STATISTICS

3.1 Overview of Probability

- Probability and statistics are both the most important concepts for Machine Learning. Probability is about predicting the likelihood of future events, while statistics involves the analysis of the frequency of past events.

  Nowadays, Machine Learning has become one of the first choices for most freshers and IT professionals. But, to enter this field, one must have some pre-specified skills, and one of those skills is Mathematics. Mathematics is very important to learning ML technology and developing efficient applications for the business.

- Probability can be calculated by the number of times the event occurs divided by the total number of possible outcomes. Let's suppose we tossed a coin, then the probability of getting head as a possible outcome can be calculated as below formula:

  P (H) = Number of ways to head occur/ total number of possible outcomes

  P (H) = ½

  P (H) = 0.5

  Where;

  P (H) = Probability of occurring Head as outcome while tossing a coin.

3.2 Statistical Tools in Machine Learning

Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, and visualizing empirical data. Descriptive statistics and inferential statistics are the two major areas of statistics. Descriptive statistics are for describing the properties of sample and population data (what has happened). Inferential statistics use those properties to test hypotheses, reach conclusions, and make predictions (what can you expect).

3.3 Descriptive Statistics

It helps in understanding the basic features of the data by summarizing them numerically or graphically. Facts regarding the data involved can be presented by descriptive analysis, however, any kind of generalization or conclusion is not possible.

Descriptive statistics provide a summary of the data, such as the mean, median, standard deviation, and variance. Univariate descriptive statistics are used to describe data containing only one variable. On the other hand, bivariate and multivariate descriptive statistics are used to describe data with multiple variables.

The marks of students in two classes are {70, 85, 90, 65} and {60, 40, 89, 96}. The average marks for each class are 77.5 and 71.25, respectively.

Descriptive statistics can be broadly classified into two categories - measures of central tendency and measures of dispersion.

Types of Descriptive Statistics:

Descriptive statistics are methods used to summarize and describe the main features of a dataset. They provide a way to organize and simplify data, making it easier to understand and interpret. Here are the major types of descriptive statistics, along with examples and visualizations:

Measures of Central Tendency

Mean: The average value of a dataset, calculated by adding all values and dividing by the number of values.

Median: The middle value in a dataset when the values are arranged in order.

Mode: The most frequent value in a dataset.

Example:

Consider the following dataset of scores: 85, 92, 78, 95, 82.

Mean = (85 + 92 + 78 + 95 + 82) / 5 = 86.4

Median = 85 (arranged in order: 78, 82, 85, 92, 95)

Mode = no mode (no value occurs more than once)

Measures of Variability

Range: The difference between the highest and lowest values in a dataset.

Variance: The average of the squared differences from the mean.

Standard Deviation: The square root of the variance, measuring how spread out the values are from the mean.

Example:

Using the same dataset of scores:

Range = 95 - 78 = 17

Variance = 35.36

Standard Deviation = 5.95

3.4 Inferential Statistics

It is simply used for explaining the meaning of descriptive stats. It is simply used to analyze, interpret results, and draw conclusions.

Inferential statistics can be classified into hypothesis testing and regression analysis. Hypothesis testing also includes the use of confidence intervals to test the parameters of a population. Given below are the different types of inferential statistics.

Types of Inferential Statistics:

Hypothesis testing is a part of statistics in which we make assumptions about the population parameter. So, hypothesis testing mentions a proper procedure by analyzing a random sample of the population to accept or reject the assumption.

Z-test

Z-test is mainly used when the data is normally distributed. We find the Z-statistic of the sample means and calculate the z-score. Z-score is given by the formula,

$$\textbf{Z-score} = (x - \mu) / \sigma$$

Z-test is mainly used when the population mean and standard deviation are given.

Confidence interval:

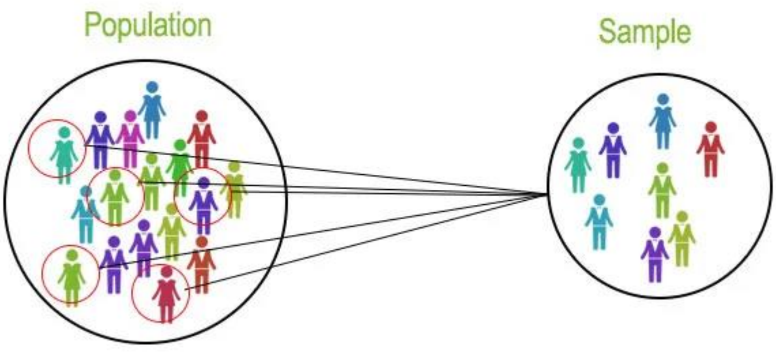
A confidence interval is a range of values that is likely to contain the true population parameter. It is used to estimate the range of values in which the population parameter lies. The confidence interval is calculated from the sample data and is often used in hypothesis testing.

Population

It refers to the collection that includes all the data from a defined group being studied. The size of the population may be either finite or infinite.

Sample

The study of the entire population is always not feasible, instead, a portion of data is selected from a given population to apply the statistical methods. This portion is called a Sample. The size of the sample is always finite



	Descriptive Statistics	Inferential Statistics
Purpose	Describe and summarize the data	Make inferences and draw conclusions about a population based on sample data
Data Analysis	Analyzes and interprets the characteristics of a dataset	Uses sample data to make generalizations or predictions about a larger population
Population vs Sample	Focuses on the entire population or dataset	Focuses on a subset of the population (sample) to draw conclusions about the entire population
Measurements	Provides measures of central tendency and dispersion	Estimates parameters, test hypotheses, and determines the level of confidence or significance in the results
Examples	Mean, median, mode, standard deviation, range, frequency tables	Hypothesis testing, confidence intervals, regression analysis, ANOVA (analysis of variance), chi-square tests, t-tests, etc.
Goal	Summarize, organize, and present data	Generalize findings to a larger population, make predictions, test hypotheses, evaluate relationships, and support decision-making
Population Parameters	Not typically estimated	Estimated using sample statistics (e.g., sample mean as an estimate of population mean)

3.5 Concept of Probability

Probability means possibility. It is a branch of mathematics that deals with the occurrence of a random event. The value is expressed from zero to one.

Experiment as a process that generates well-defined outcomes. On any single repetition of an experiment, one and only one of the possible experimental outcomes will occur.
Examples: Hitting a target, checking the boiling point of a liquid, taking an examination for a student, conducting interviews for some jobs, tossing a coin, rolling a die, hitting a ball with a batsman, sale of products, chemical reaction of elements, are few examples of experiments.

The **sample space** for an experiment is the set of all experimental outcomes.
Example: In the experiment of hitting a target, sample space can be hitting a target, missing the target.

An **Event** is one or more of the possible outcomes of an experiment.
Example: If we toss a coin, getting a head will be one event, and getting a tail will be another event.

For example:- when we toss a coin, either we get Head OR Tail, only two possible outcomes are possible (H, T). But when two coins are tossed then there will be four possible outcomes, i.e. {(H, H), (H, T), (T, H), (T, T)}.

Joint Probability

When the probability of two more events occurring together and at the same time is measured it is marked as Joint Probability. For two events A and B, it is denoted by joint probability is denoted as, $P(A \cap B)$ intersection of two or more events.

Formula: P(A∩B) = P(A) * P(B)

Example: Find the probability that the number three will occur twice when two dice are rolled at the same time.

Solution: Number of possible outcomes when a die is rolled = 6

i.e. {1, 2, 3, 4, 5, 6}

Let A be the event of occurring 3 on first die and B be the event of occurring 3 on the second die.

Both the dice have six possible outcomes, the probability of a three occurring on each die is 1/6.

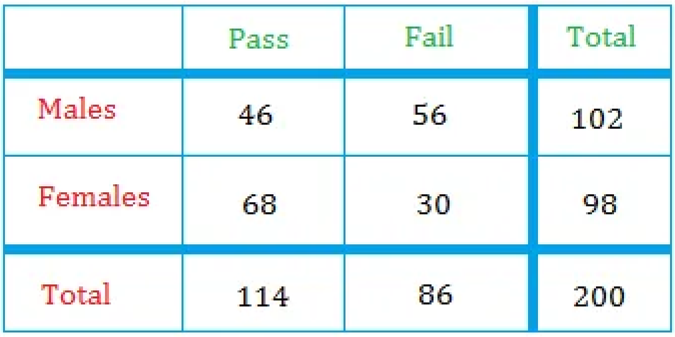
P(A) =1/6

P(B)=1/6

P(A, B) = 1/6 x 1/6 = 1/36

Marginal Probability

The marginal probability is the probability of a single event occurring, independent of other events. Now we have to calculate these probabilities by using a two-way table.



	Pass	Fail	Total
Males	46	56	102
Females	68	30	98
Total	114	86	200

Suppose a company specializes in training students to pass the ML test. The company had 200 students last year.

The following contingency table show their pass rates broken down by gender.

In the table above, the variable that categorizes rows is the gender. This variable can be males or females.

The variable that categorizes columns is the result. This variable can be pass or fail.

Did you notice the followings?

- The number of males + females = 102 + 98 = 200
- The number of students who passed + the number of students who failed = 114 + 86 = 200
- The number of females who failed the ML test is equal to 30

Now that you understand what a contingency table is, what is marginal probability?

In the table above, there are 4 events.

- A student is a male
- A student is a female
- A student has passed
- A student has failed

The probability of each of these 4 events is called marginal probability or simple probability the 4 marginal probabilities can be calculated as follows

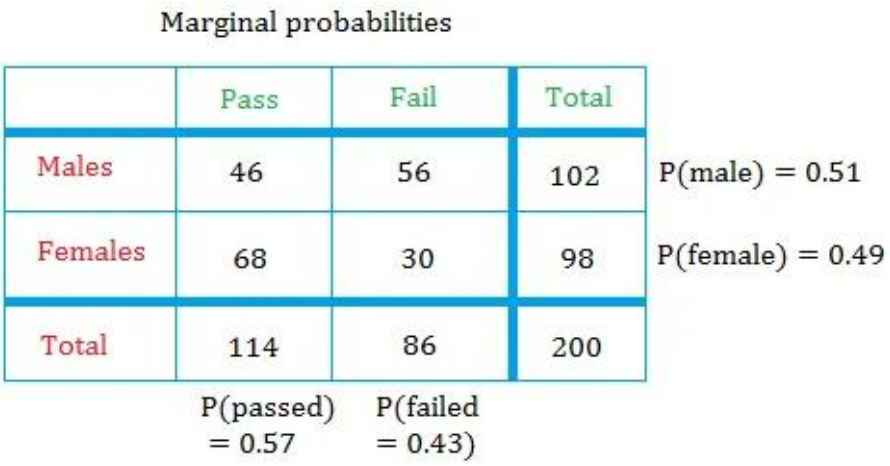
P(A student is a male) = Number of males / Total number of students
= 102 / 200 = 0.51

P(A student is a female) = Number of females / Total number of students
= 98 / 200 = 0.49

P(A student has passed) = No. of students who passed / Total number of students
= 114/200 = 0.57

P(A student has failed) = No. of students who failed / Total number of students = 86 / 200 = 0.43

The marginal probabilities are shown along the right side and along the bottom of the table below.



Marginal probabilities

	Pass	Fail	Total	
Males	46	56	102	P(male) = 0.51
Females	68	30	98	P(female) = 0.49
Total	114	86	200	
	P(passed) = 0.57	P(failed = 0.43)		

Conditional Probability

The probability of an event A based on the occurrence of another event B is termed conditional Probability. It is denoted as P(A|B) and represents the probability of A when event B has already happened.

Here:

P(A | B) = The probability of A given B (or) the probability of A which happens after B

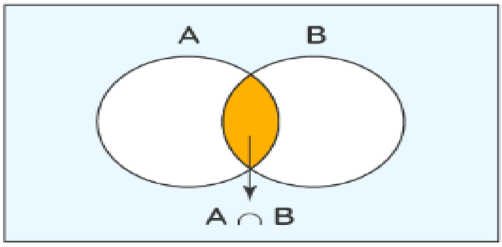
P(B | A) = The probability of B given A (or) the probability of B which happens after A

P(A ∩ B) = The probability of happening of both A and B

P(A) = The probability of A

P(B) = The probability of B

Conditional Probability Formula



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Example: A bag contains 3 red and 7 black balls. Two balls are drawn at random without replacement. If the second ball is red, what is the probability that the first ball is also red?

Solution:

Let A: event of selecting a red ball in first draw

B: event of selecting a red ball in the second draw

P(A ∩ B) = P(selecting both red balls) = 3/10 × 2/9 = 1/15

P(B) = P(selecting a red ball in the second draw) = P(red ball and rad ball or black ball and red ball)

= P(red ball and red ball) + P(black ball and red ball)

= 3/10 × 2/9 + 7/10 × 3/9 = 3/10

∴ P(A|B) = P(A ∩ B)/P(B) = 1/15 ÷ 3/10 = 2/9.

Example: Two dice are rolled, if it is known that atleast one of the dice always shows 4, find the probability that the numbers appeared on the dice have a sum 8.

Solution:

Let,

A: one of the outcomes is always 4

B: sum of the outcomes is 8

Then, A = {(1, 4), (2, 4), (3, 4), (4, 4), (5, 4), (6, 4), (4, 1), (4, 2), (4, 3), (4, 5), (4, 6)}

B{(4, 4), (5, 3), (3, 5), (6, 2), (2, 6)}

n(A) = 11, n(B) = 5, n(A ∩ B) = 1

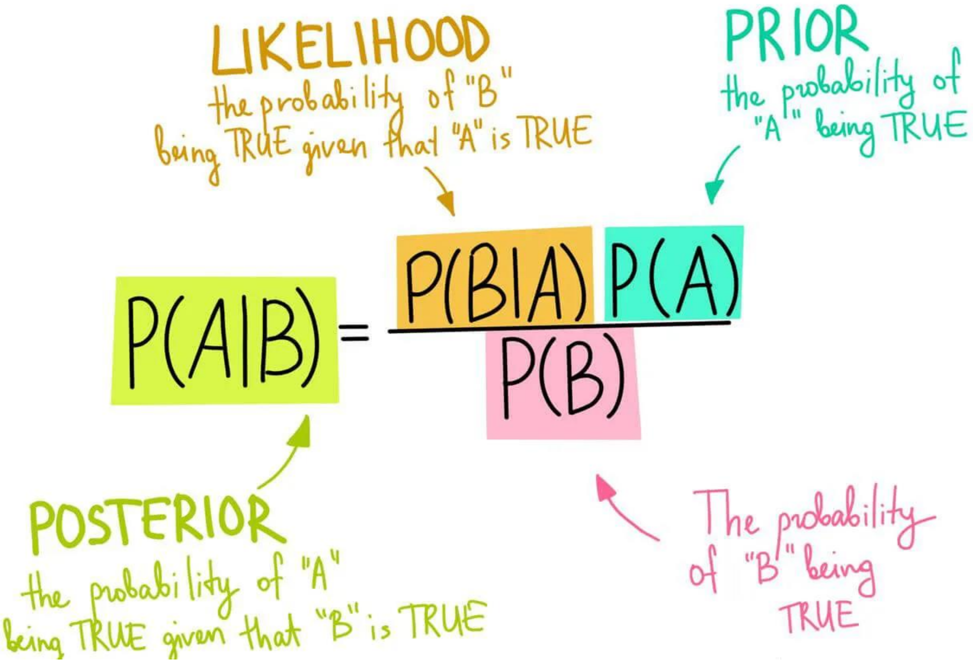
P(B|A) = n(A ∩ B)/n(A) = 1/11.

Actually the basic difference between them is that the joint probability is the probability of two events occurring simultaneously, and in the marginal probability is the probability of an event irrespective of the outcome of another variable, and conditional probability is the probability of one event occurring in the presence of a second event.

Bayes' Theorem

Bayes' theorem is also known as **Bayes' rule, Bayes' law**, or **Bayesian reasoning**, which determines the probability of an event with uncertain knowledge.

In probability theory, it relates the conditional probability and marginal probabilities of two random events.


LIKELIHOOD
the probability of "B" being TRUE given that "A" is TRUE
PRIOR
the probability of "A" being TRUE
P(A|B) = P(B|A) P(A) / P(B)
POSTERIOR
the probability of "A" being TRUE given that "B" is TRUE
The probability of "B" being TRUE

P(A|B) is known as **posterior**, which we need to calculate, and it will be read as Probability of hypothesis A when we have occurred an evidence B.

P(B|A) is called the likelihood, in which we consider that hypothesis is true, then we calculate the probability of evidence.

P(A) is called the **prior probability**, probability of hypothesis before considering the evidence

P(B) is called **marginal probability**, pure probability of an evidence.

Example:

There are two urns containing colored balls. The first urn contains 50 red balls and 50 blue balls. The second urn contains 30 red balls and 70 blue balls. One of the two urns is randomly chosen (both urns have a probability of 50% of being chosen) and then a ball is drawn at random from one of the two urns. If a red ball is drawn, what is the probability that it comes from the first urn?

Solution

In probabilistic terms, what we know about this problem can be formalized as follows:

$$\begin{aligned} P(\text{red} \mid \text{urn } 1) &= \tfrac{1}{2} \\ P(\text{red} \mid \text{urn } 2) &= \tfrac{3}{10} \\ P(\text{urn1}) &= \tfrac{1}{2} \\ P(\text{urn } 2) &= \tfrac{1}{2} \end{aligned}$$

The unconditional probability of drawing a red ball can be derived using the law of total probability:

$$\begin{aligned} P(\text{red}) &= P(\text{red} \mid \text{urn } 1)P(\text{urn } 1) + P(\text{red} \mid \text{urn } 2)P(\text{urn } 2) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{3}{10} \cdot \frac{1}{2} \\ &= \frac{1}{4} + \frac{3}{20} \\ &= \frac{5+3}{20} = \frac{2}{5} \end{aligned}$$

By using Bayes' rule, we obtain

$$\begin{aligned} P(\text{urn } 1 \mid \text{red}) &= \frac{P(\text{red} \mid \text{urn } 1)P(\text{urn } 1)}{P(\text{red})} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{2}{5}} \\ &= \frac{1}{4} \cdot \frac{5}{2} = \frac{5}{8} \end{aligned}$$

3.6 Random Variables

A random variable is a variable which represents the outcome of a trial, an experiment, or an event. It is a specific number which is different each time the trial, experiment, or event is repeated.

- A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes.
- A random variable can be either discrete (having specific values) or continuous (any value in a continuous range).
- The use of random variables is most common in probability and statistics, where they are used to quantify outcomes of random occurrences.
- Risk analysts use random variables to estimate the probability of an adverse event occurring.

Types of Random Variables

Continuous random variable

Continuous random variables take up an infinite number of possible values which are usually in a given range. Typically, these are measurements like weight, height, the time needed to finish a task, etc.

To give you an example, the life of an individual in a community is a continuous random variable. Let’s say that the average lifespan of an individual in a community is 110 years.

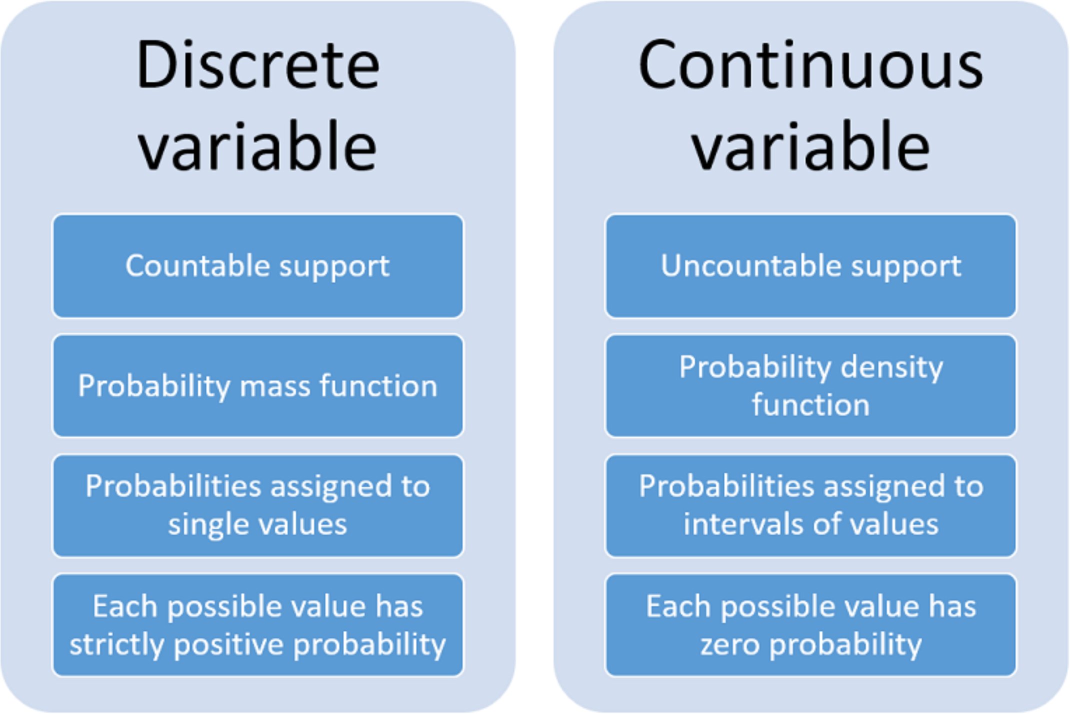
Therefore, a person can die immediately on birth (where life = 0 years) or after he attains an age of 110 years. Within this range, he can die at any age. Therefore, the variable ‘Age’ can take any value between 0 and 110.

Hence, continuous random variables do not have specific values since the number of values is infinite. Also, the probability at a specific value is almost zero.

Discrete random variable

Discrete random variables take on only a countable number of distinct values. Usually, these variables are counts (not necessarily though). If a random variable can take only a finite number of distinct values, then it is discrete.

Number of members in a family, number of defective light bulbs in a box of 10 bulbs, etc. are some examples of discrete random variables.



Discrete variable	Continuous variable
Countable support	Uncountable support
Probability mass function	Probability density function
Probabilities assigned to single values	Probabilities assigned to intervals of values
Each possible value has strictly positive probability	Each possible value has zero probability

3.7 Probability Distribution

Sampling Distribution

A sampling distribution is a probability distribution of a statistic that is based on random samples from a population. It describes the range of possible outcomes for a statistic, such as the mean or mode of a variable.

Discrete Distribution

A discrete probability distribution is a type of probability distribution that shows all possible values of a discrete random variable along with the associated probabilities. In other words, a discrete probability distribution gives the likelihood of occurrence of each possible value of a discrete random variable.

Such a distribution will represent data that has a finite countable number of outcomes

A discrete probability distribution counts occurrences that have countable or finite outcomes.

In finance, discrete distributions are used in options pricing and forecasting market shocks or recessions.

Represented by bars or points, such as in a histogram or probability mass function plot.

Examples: binomial distribution, Poisson distribution, geometric distribution

Continuous Distribution

Continuous Probability Distributions. A continuous distribution describes the probabilities of a continuous random variable's possible values. A continuous random variable has an infinite and uncountable set of possible values (known as the range).

Involves continuous random variables that can take any value within a range. Examples include height, weight, temperature, and time.

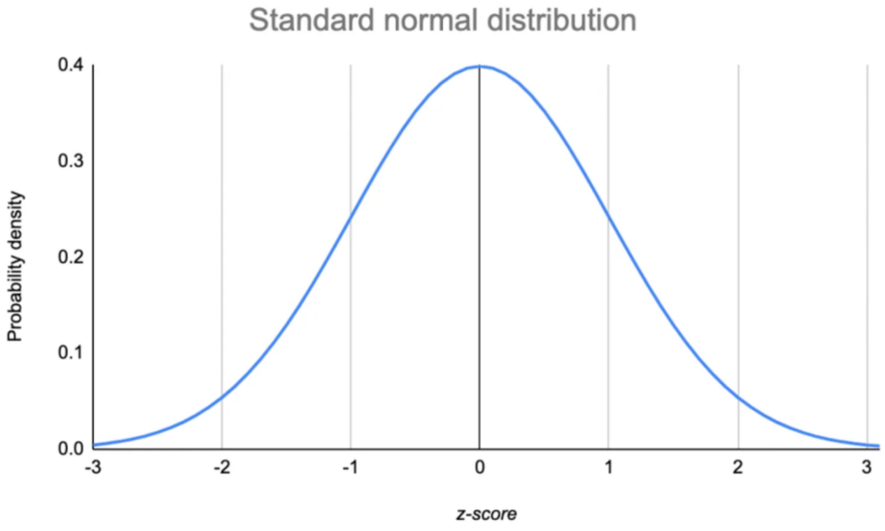
Represented by smooth curves, such as the bell curve of the normal distribution.

Examples: normal distribution, exponential distribution, beta distribution.

Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

The normal distribution appears as a "bell curve" in graphical form.



3.8 Central Limit Theorem

The **central limit theorem,** which is a statistical theory, states that when a large sample size has a finite variance, the samples will be normally distributed, and the mean of samples will be approximately equal to the mean of the whole population.

In other words, the central limit theorem states that for any population with mean and standard deviation, the distribution of the sample mean for sample size N has mean μ and standard deviation σ/√n.

Central limit theorem formula

Fortunately, you don’t need to actually repeatedly sample a population to know the shape of the sampling distribution. The parameters of the sampling distribution of the mean are determined by the parameters of the population:

The standard deviation of the sampling distribution is the standard deviation of the population divided by the square root of the sample size.

$$\mu_{\overline{x}} = \mu$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Where,

N is the normal distribution

μ is the mean of the population

σ is the standard deviation of the population

n is the sample size

3.9 Monte Carlo Approximation

A Monte Carlo simulation is used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. It is a technique used to understand the impact of risk and uncertainty.

A Monte Carlo simulation is used to tackle a range of problems in many fields including investing, business, physics, and engineering. It is also referred to as a multiple probability simulation.

A Monte Carlo simulation requires assigning multiple values to an uncertain variable to achieve multiple results and then averaging the results to obtain an estimate.

Monte Carlo simulations assume perfectly efficient markets.

History of the Monte Carlo Simulation

The Monte Carlo simulation was named after the gambling destination in Monaco because chance and random outcomes are central to this modeling technique, as they are to games like roulette, dice, and slot machines.

The technique was initially developed by Stanislaw Ulam, a mathematician who worked on the Manhattan Project, the secret effort to create the first atomic weapon. He shared his idea with John Von Neumann, a colleague at the Manhattan Project, and the two collaborated to refine the Monte Carlo simulation.