Data Science

Inferential Statistics

Descriptive Statistics

- It helps organize data and focuses on main characteristics of data
- It provides summary of data numerically

Inferential Statistics

- Inferential statistics is all about describing the larger picture of the analysis with a limited set of data and deriving conclusions from it.
- Basically, inferential statistics aims at drawing conclusions on populations based on the taken data samples.
- It uses a random sample of data taken from a population to describe and make inferences about the population.

Inferential Statistics

- Inferential Statistics is used to draw inferences beyond the immediate data available.
- In inferential statistics we use methods that rely on probability theory and distribution helping us to predict, in particular, the population's values based on sample data.
- Inferential statistics helps us answer the following questions:
  - Making inferences about a population from a sample
  - Concluding whether a sample is significantly different from the population

Inferential Statistics

- There are two main areas of inferential statistics:
  - Estimating parameters - taking a statistic from your sample data (for example the sample mean) and using it to find something about a population parameter (i.e. the population mean).
  - Hypothesis tests - use sample data to answer research questions. For example, finding if a new cancer drug is effective or not. Or if breakfast helps children perform better in schools.

Inferential Statistics

- Prerequisites for understanding Inferential Statistics -
  - Descriptive Statistics
  - Probability
  - Probability Distributions

Probability Distribution

Type of data

- Discrete
  - Can take only specified values
- Continuous
  - Can take any value within a given range

Terminologies

- Random Variable
  - Whose value is determined by the outcome of a random experiment
- Discrete random variable
  - Whose set of assumed values is countable (arises from counting)
- Continuous random variable
  - Whose set of assumed values is uncountable (arises from measurement.)

Probability Distribution

- In statistics, with distribution we usually mean probability distribution
- A probability distribution is a function that shows the possible values for a variable and how often they occur.
- A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values.
- For eg. rolling a dice, tossing a coin, measuring weight of a student etc

Probability Distribution

- Kind of variable determines the type of probability distribution -
  - Discrete probability distributions for discrete variables
  - Probability density functions for continuous variables

Discrete Probability Distribution

- Also known as **Probability Mass Functions**.
- For example - coin tosses, rolling a dice
- Each possible value has a non-zero likelihood
- The probabilities for all possible values must sum to one

Discrete Probability Distribution

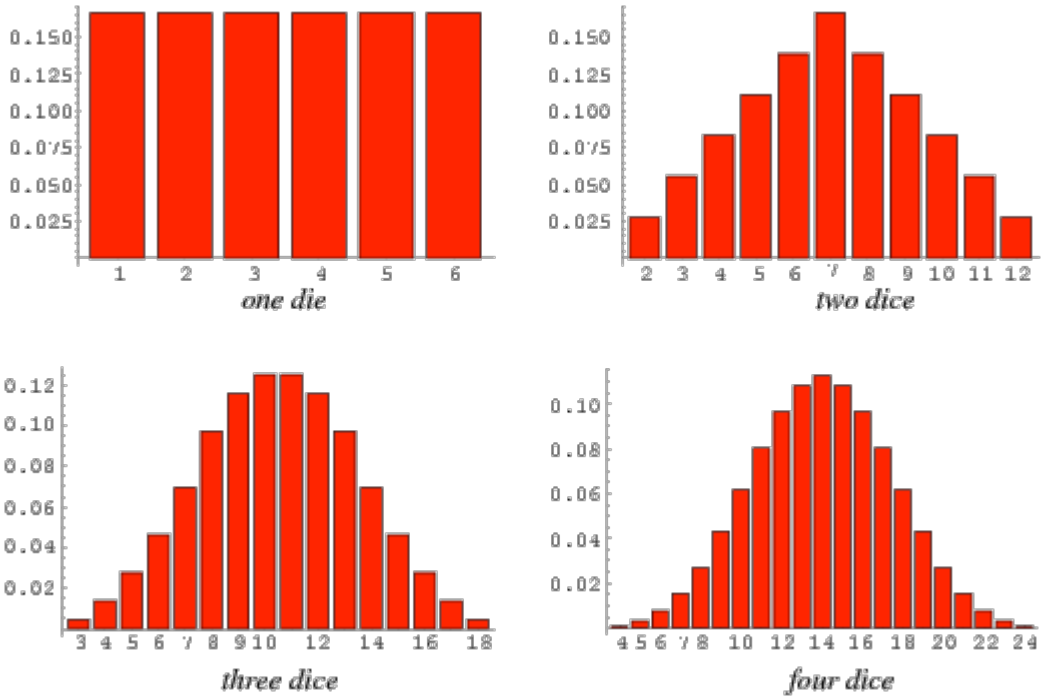


Image Source - http://mathworld.wolfram.com/Dice.html

Discrete Probability Distribution

- There are a variety of discrete probability distributions that you can use to model different types of data. The correct discrete distribution depends on the properties of your data.
- Types of Discrete Distribution
  - Binomial distribution
  - Poisson distribution
  - Uniform distribution

Continuous Probability Distribution

- Also known as **Probability Density Functions**
- For example - often measurements on a scale, such as height, weight, and temperature.
- Specific values in continuous distributions have a zero probability. For example, the likelihood of measuring a temperature that is exactly 32 degrees is zero (because an individual value has an infinitesimally small probability that is equivalent to zero).

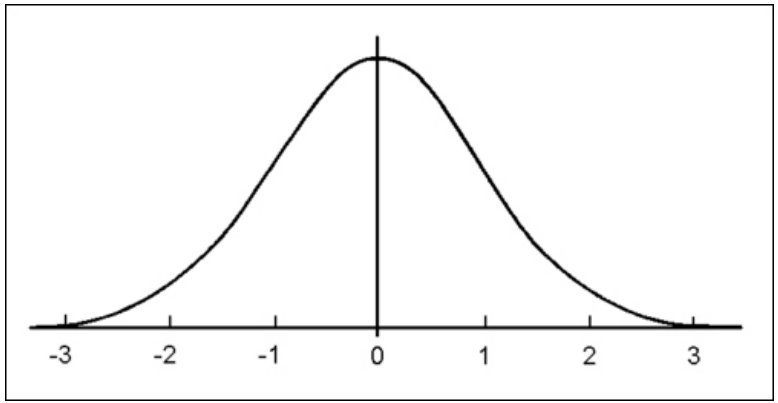
Continuous Probability Distribution

- Probabilities for continuous distributions are measured over ranges of values rather than single points.
- A probability indicates the likelihood that a value will fall within an interval.
- On a probability plot, the entire area under the distribution curve equals 1. This fact is equivalent to how the sum of all probabilities must equal one for discrete distributions.
- The most well-known continuous distribution is the **Normal Distribution**.

Normal Distribution

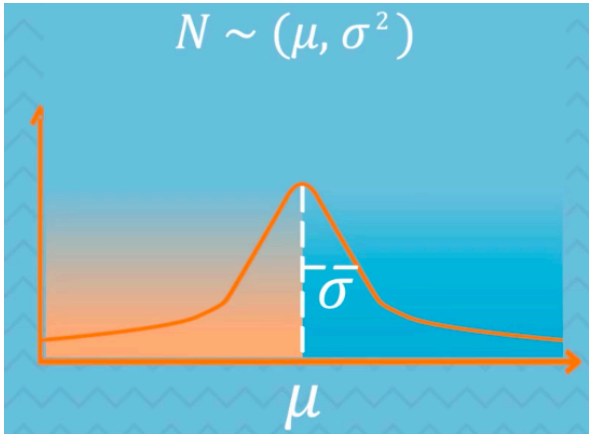
Normal Distribution

- A normal distribution is the most common and widely used distribution in statistics because it approximates to a wide variety of random variables
- It is also called a "bell curve" and "Gaussian curve"

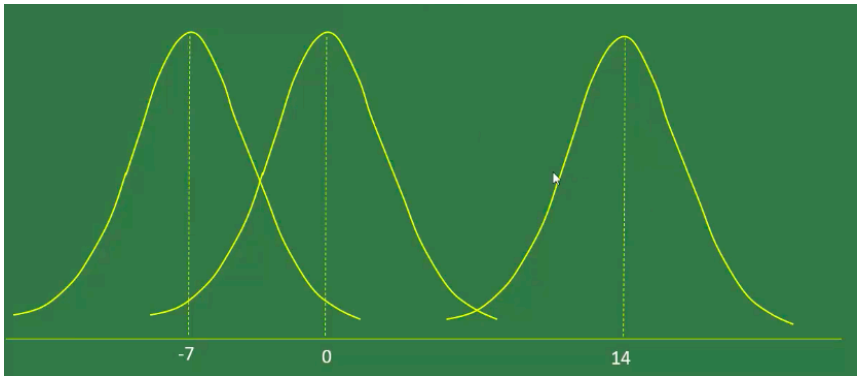


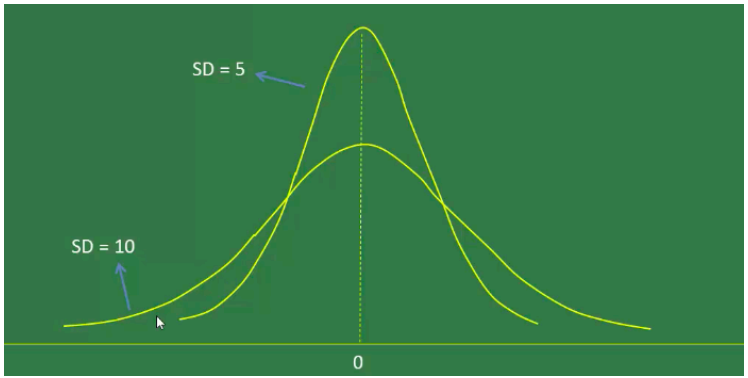
Characteristics of Normal Distribution

- Mean = Median = Mode
- It is symmetric, perfectly centred around mean
- The area under the curve is 1.
- The entire family of normal distribution is differentiated by two parameters -
  - Mean and
  - Standard Deviation
- It is denoted as -
  - N ~ (μ, $\boldsymbol{\sigma}^2$)



Characteristics of Normal Distribution





Standard Normal Distribution

Standard Normal Distribution

- The standard normal distribution is a special case of the normal distribution -
  - when a normal random variable has a mean of zero and
  - a standard deviation of one
- So if we shift the mean by μ and the standard deviation by $\boldsymbol{\sigma}$ for any normal distribution we will arrive at the standard normal distribution. We use the letter Z to denote it
  - $z = (X - \mu) / \sigma$
- The normal random variable of a standard normal distribution is called a standard score or a z score.

Standard Normal Distribution

- Why do we need standard Normal Distribution -
  - Makes predictions and inferences much easier
  - Compare different normally distributed datasets;
  - Detect normality
  - Detect outliers
  - Create confidence intervals
  - Test Hypothesis

Sampling Distribution

Sampling Distribution

- A Sampling Distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population.
- For example, consider a normal population with mean μ and variance σ. Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean for each sample. This statistic is then called the sample mean. Each sample has its own average value, and the distribution of these averages is called the "sampling distribution of the sample mean.

Sampling Distribution

- A Sampling Distribution behaves much like a normal curve and has some interesting properties like :
  - The shape of the Sampling Distribution does not reveal anything about the shape of the population.
  - Sampling Distribution helps to estimate the population statistic, using **Central Limit Theorem**

Central Limit Theorem (CLT)

Sampling Distribution

- Sampling distribution can be very useful in making inferences about the overall population
- To find - how much sample means differ from each other, we'll use standard deviation of the sampling distribution
- This standard deviation is called the **standard error**.
  - Standard error (SE) = s/sqrt(n)

Central Limit Theorem

- The central limit theorem states (given that sample size >= 30) -
  - The sampling distribution of the sample mean has an approximately normal distribution.
  - The mean of the sampling distribution is equals to the population mean
  - The standard deviation of the sampling distribution equals the standard deviation in the population divided by the square root of the sample size (i.e. standard error)

Central Limit Theorem

- Points to note -
  - Central Limit Theorem holds true irrespective of the type of distribution of the population.
  - Now, we have a way to estimate the population mean by just making repeated observations of samples of a fixed size.
  - Greater the sample size, lower the standard error and greater accuracy in determining the population mean from the sample mean.

Central Limit Theorem

- Significance of Central Limit Theorem -
  - Analyzing data involves statistical methods like hypothesis testing and constructing confidence intervals. These methods assume that the population is normally distributed. In case of unknown or non-normal distributions, we treat the sampling distribution as normal according to the central limit theorem
  - If we increase the samples drawn from the population, the standard deviation of sample means will decrease. This helps us estimate the population mean much more accurately