Probability Introduction

Chance is the occurrence of events in the absence of any obvious intention or cause. It is, simply, the possibility of something happening. When the chance is defined in Mathematics, it is called probability.

Probability is the extent to which an event is likely to occur, measured by the ratio of the favourable cases to the whole number of cases possible.
Mathematically, the probability of an event occurring is equal to the ratio of a number of cases favourable to a particular event to the number of all possible cases.

The theoretical probability of an event is denoted as P(E).

$$P(E) = \frac{\text{Number of Outcomes Favourable to E}}{\text{Number of all Possible Outcomes of the Experiment}}$$

Importance of Probability

The concept of probability is of great importance in everyday life. Statistical analysis is based on this valuable concept. Infact the role played by probability in modern science is that of a substitute for certainty.

The following discussion explains it further:

i. The probability theory is very much helpful for making prediction. Estimates and predictions form an important part of research investigation. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.

ii. It has also immense importance in decision making.

iii. It is concerned with the planning and controlling and with the occurrence of accidents of all kinds.

iv. It is one of the inseparable tools for all types of formal studies that involve uncertainty.

v. The concept of probability is not only applied in business and commercial lines, rather than it is also applied to all scientific investigation and everyday life.

vi. Before knowing statistical decision procedures one must have to know about the theory of probability.

vii. The characteristics of the Normal Probability. Curve is based upon the theory of probability.

Random Experiment

A random experiment is defined as an experiment whose outcome cannot be predicted with certainty

An activity that produces a result or an outcome is called an experiment. It is an element of uncertainty as to which one of these occurs when we perform an activity or experiment. Usually, we may get a different number of outcomes from an experiment. However, when an experiment satisfies the following two conditions, it is called a random experiment.

(i) It has more than one possible outcome.

(ii) It is not possible to predict the exact outcome in advance.

Outcome

A possible result of random experiment is called a possible outcome of the experiment.

Sample Space

The set of all possible outcomes of a random experiment is called the sample space. The sample space is denoted by S or Greek letter omega (Ω). The number of elements in S is denoted by n(S). A possible outcome is also called a sample point since it is an element in the sample space.

Event

A subset of the sample space is called an event.

Favourable Outcome

An outcome that belongs to the specified event is called a favourable outcome.

Types of Events

Elementary Event: An event consisting of a single outcome is called an elementary event.

Certain Event: The sample space is called the certain event if all possible outcomes are favourable outcomes. i.e. the event consists of the whole sample space.

Impossible Event: The empty set is called impossible event as no possible outcome is favorable

Union of Two Events

Let A and B be two events in the sample space S. The union of A and B is denoted by A∪B and is the set of all possible outcomes that belong to at least one of A and B.

Let S = Set of all positive integers not exceeding 50;

Event A = Set of elements of S that are divisible by 6;

Event B = Set of elements of S that are divisible by 9.

A = {6,12,18,24,30,36,42,48}

B = {9,18,27,36,45}

∴ A∪B = {6,9,12,18,24,27,30,36,42,45, 48} is the set of elements of S that are divisible by 6 or 9.

Exhaustive Events

Two events A and B in the sample space S are said to be exhaustive if A∪B = S

Intersection of Two Events

Let A and B be two events in the sample space S.

The intersection of A and B is the event consisting of outcomes that belong to both the events A and B.

Let S = Set of all positive integers not exceeding 50,

Event A = Set of elements of S that are divisible by 3,

Event B = Set of elements of S that are divisible by 5.
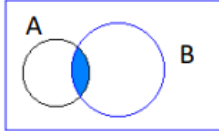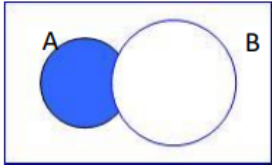
Then A = {3,6,9,12,15,18,21,24,27,30,33, 36,39,42,45,48},

B = {5,10,15,20,25,30,35,40,45,50}

∴ A∩B = {15,30,45} is the set of elements of S that are divisible by both 3 and 5.

Mutually Exclusive Events

Event A and B in the sample space S are said to be mutually exclusive if they have no outcomes in common. In other words, the intersection of mutually exclusive events is empty. Mutually exclusive events are also called disjoint events.

If two events A and B are mutually exclusive and exhaustive, then they are called **Complementary events**.

S.NO	Operator	Symbol	Example	Meaning
1	Union	$\cup$	$A \cup B$	The event of either A or B occurring.
2	Finite union	$\bigcup\limits_{i=1}^{n} A_i$	$\bigcup\limits_{i=1}^{3} A_i$	The event of any one of the events A_1, A_2 and A_3 occurring.
3	Countable union	$\bigcup\limits_{i=1}^{\infty} A_i$	$\bigcup\limits_{i=1}^{\infty} A_i$	The event of any one of the events A_1, A_2 ….. occurring.
4	Intersection	$\cap$	$A \cap B$	The event of both A and B occurring.
5	Finite intersection	$\bigcap\limits_{i=1}^{n} A_i$	$\bigcap\limits_{i=1}^{3} A_i$	The event of all the events A_1, A_2 and A_3 occurring.
6	Countable intersection	$\bigcap\limits_{i=1}^{\infty} A_i$	$\bigcap\limits_{i=1}^{\infty} A_i$	The event of all the events A_1, A_2 ….. occurring.
7	Complementation	c or $-$	A^c or $\overline{A}$	The event of A not occurring.
8	Subtraction	-	A-B	The event of A occurring and B not occurring.

Operation	Interpretation
A', A or A^c	Not A.
$A \cup B$	At least, one of A and B
$A \cap B$	Both A and B
$(A' \cap B) \cup (A \cap B')$	Exactly one of A and B
$(A' \cap B') = (A \cup B)'$	Neither A nor B

Elementary Properties of Probability:

1) A' is complement of A and therefore $P(A') = 1 - P(A)$

2) For any event A in S, $0 \leq P(A) \leq 1$

3) For the impossible event φ, $P(\varphi) = 0$

4) For the certain event S, $P(S) = 1$

5) If A1 and A2 two mutually exclusive events then $P(A1 \cup A2) = P(A1) + P(A2)$

6) If $A \subseteq B$, then $P(A) \leq P(B)$ and $P(A' \cap B) = P(B) - P(A)$

7) Addition theorem: For any two events A and B of a sample space S,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

8) For any two events A and B, $P(A \cap B') = P(A) - P(A \cap B)$

9) For any three events A, B and C of a sample space S,

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - (P(A \cap C) + P(A \cap B \cap C)$

10) If A_1, A_2, ..., A_m are mutually exclusive events in S, then

$P(A_1 \cup A_2 \cup, ... \cup A_m = P(A_1) + P(A_2) + ... + P(A_m)$

Conditional Probability

Let S be a sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S.

Then the probability of occurrence of event A under the condition that event B has already occurred and P(B) $\neq$ 0 is called conditional probability of event A given B and is denoted by P(A/B).

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0$$

Multiplication theorem

Let S be sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S.

Then the probability of occurrence of both the events is denoted by P(A∩B)

and is given by P(A∩B) = P(A).P(B/A)

Independent Events

Let S be sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S. If the occurrence of either event, does not affect the probability of the occurrence of the other event, then the two events A and B are said to be independent.

Thus, if A and B are independent events then,

P(A/B) = P(A/B') = P(A) and P(B/A) = P(B/A') = P(B)

If A and B are independent events then P(A∩B) = P(A).P(B)

(P(A∩B) = P(A).(B/A) = P(A).P(B) ∴ P(A∩B) = P(A).P(B))

If A and B are independent events then

      a) A and B' are also independent event

      b) A' and B' are also independent event

Bayes Theorem

Bayes' Theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence

- Bayes' Theorem allows you to update the predicted probabilities of an event by incorporating new information.
- Bayes' Theorem was named after 18th-century mathematician Thomas Bayes.
- It is often employed in finance in calculating or updating risk evaluation.
- The theorem has become a useful element in the implementation of machine learning.
- The theorem was unused for two centuries because of the high volume of calculation capacity required to execute its transactions.
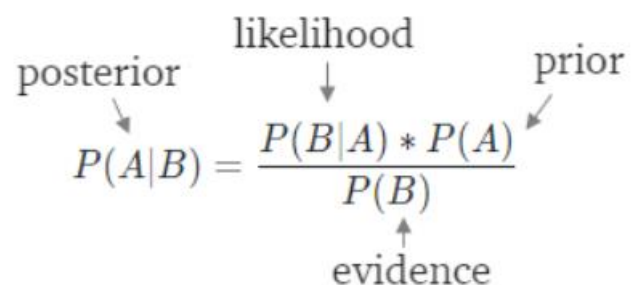
Applications of Bayes' Theorem are widespread and not limited to the financial realm. For example, Bayes' theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test. Bayes' theorem relies on incorporating prior probability distributions in order to generate posterior probabilities.

Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed.

Posterior probability is the revised probability of an event occurring after taking into consideration the new information. Posterior probability is calculated by updating the prior probability using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

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

Bayes' Theorem formula

$$\underset{\text{posterior}}{P(A|B)} = \frac{\overset{\text{likelihood}}{P(B|A)} * \overset{\text{prior}}{P(A)}}{\underset{\text{evidence}}{P(B)}}$$

ODDS (Ratio of two complementary probabilities):

Let n be number of distinct sample points in the sample space S. Out of n sample points, m sample points are favourable for the occurrence of event A. Therefore remaining (n-m) sample points are favourable for the occurrence of its complementary event A'.

$\therefore P(A) = \dfrac{m}{n}$ and $P(A') = \dfrac{n-m}{n}$

Ratio of number of favourable cases to number of unfavourable cases is called as odds in favour of event A which is given by $\dfrac{m}{n-m}$ i.e. P(A):P(A').

Ratio of number of unfavourable cases to number of favourable cases is called as odds against event A which is given by $\dfrac{n-m}{n}$ i.e. P(A'):P(A

Random Variable

A random variable is a variable whose values can be determined from the outcomes of a random experiment example in an experiment of throwing a pair of dice,

S = {(1,1), (1,2), ……, (6,6), if we define a variable,

X = sum of the numbers on uppermost faces. From the outcomes of sample space we can determine values assume by the variable that is X assumes the values 2,3,….,12 and hence X is a random variable.

According to the values assumed by the random variable we have two tpes of random variable.

- A **discrete random variable** is one whose set of assumed values is countable (arises from counting).

- A **continuous random variable** is one whose set of assumed values is uncountable (arises from measurement.).

Examples:

(1) If the random experiment is throwing an unbiased dice, the random variable associated with this experiment can be defined as the number on the uppermost face, then possible values of X are 1, 2, 3, 4, 5, 6.

(2) If two cards are taken from pack of fifty-two cards, X = number of red cards then X can take the values 0, 1,2.

(3) The number of missed calls received on a particular day.

(4) Suppose milk contents in one liter bags are measured then X = milk content in bag, then X can take values 990 ml $<X \leq 1100$ ml

In examples one and two the sample space of random variable is finite and in third it is countably infinite so these are Discrete random variables where as in fourth sample space interval so values are uncountable hence it is continuous random variable. is

Notation: The random variable is denoted by uppercase letter and its value in lowercase letter. E.g. X denotes a random variable whereas x denotes its value.

Discrete variable: If the number of possible values of the variable is finite or countably infinite then the variable is called as a discrete variable. Thus discrete random variable takes only isolated values.

For example:

(i)   Age in completed years.

(ii)   Number of arrivals at the Clinic.

(iii)   Number of Jobs completed per day.

(iv)   Number of accidents on a particular spot per day.

(v)   Number of heads in 4 flips of a coin (possible outcomes are 0, 1, 2, 3, 4).

(vi)   Number of classes missed last week (possible outcomes are 0, 1, 2, 3, maximum number).

Discrete variables obtained by counting.

PROBABILITY MASS FUNCTION

Let X be a discrete random variable. The set of all possible values of X are denoted by the sample space S. Then probability function

$P(X = x)$ for all S is known as probability mass function if function satisfies the following two conditions.

(i) $0 < P(X = x) < 1$ for all S

(ii) $\Sigma P(X = x) = 1$

CUMULATIVE PROBABILITY DISTRIBUTION FUNCTION

$F(x) = P(X \leq x) = \sum_{-\infty}^{x} P(X = x)$

Discrete Uniform Distribution

A discrete uniform is the simplest type of all the probability distributions. Discrete uniform distribution is a symmetric probability distribution whereby a finite number of values are equally likely to be observed; such that each value among n possible values has equal probability 1/n. In other words discrete uniform distribution could be expressed as "a known, finite number of outcomes equally likely to happen".

Consider the case of throwing a die. The following observations can be made with this action.

(1) The die is numbered from 1 to 6. In probability language we call it as six outcomes.

(2) The die can roll on any number.

(3) The total number of possible events is 6. That is, the die can roll on any of the numbers from 1 to 6. In other words, the events are 'equally likely' with the same probability of 1/6.

(4) The numbers engraved on the die are 6. In other words, the number of values of the random variable is finites.

(5) The numbers are equally spaced or we can say the values of the variables are equally spaced. The experiment leads to discrete Uniform distribution, satisfying all above stated characteristics. Thus a discrete uniform distribution is a probability distribution of equally likely events with equal probability and with finite number of equally spaced outcomes. The discrete uniform distribution is essentially non-parametric, i.e. it does not involve any parameter.

A discrete random variable is said to follow uniform distribution over the range 1,2,2,,,,n if tis pmf is given by

$$P(X = x) = \frac{1}{n} \qquad x = 1,2,3, \ldots \ldots, n$$

$$= 0 \qquad \text{otherwise}$$

$$\text{Mean} = \frac{n+1}{2}$$

$$\text{Variance} = \frac{n^2 - 1}{12}$$

Binomial Distribution

Bernoulli experiment: Suppose we perform an experiment with two possible outcomes; either success or failure. Success happens with probability p, hence failure occurs with probability q=1-p then such experiment is referred to as Bernoulli experiment.

Probability mass function of Bernoulli variate x is given by

$$P_x(x) = P(X = x) = p^x q^{1-x} \quad x = 0,1; \ 0<p<1; p+q=1$$

$$= 0 \text{ otherwise } 0<p<1 p+q=1$$

The binomial experiment means Bernoulli experiment which is repeated n times. The binomial distribution is used to obtain the probability of observing x successes in n trials, with the probability of success on a single trial denoted by p. The binomial distribution assumes that p is fixed for all trials. Here n and p are called as parameters of binomial distribution.

In the above definition notice that the following conditions need to be satisfied for a binomial experiment:

(1) n trials are carried out where n is fixed number.

(2) The outcome of a given trial is either a "success" or "failure".

(3) The probability of success (p) remains constant from trial to trial.

(4) The trials are independent; the outcome of a trial is not affected by the outcome of any other trial.

A discrete random variable X is said to follow **Binomial distribution** with parameters (n,p) if tis pmf given by

$$P_x(x) = P(X = x) = nC_x p^x q^{n-x} \quad x = 0,1,2,\ldots,n; \ 0<p<1; p+q=1$$

$$= 0 \text{ otherwise } 0<p<1 p+q=1$$

Mean = np

Variance = npq

Real Life Applications of Binomial Distributions

.

- There are a lot of areas where the application of binomial theorem is inevitable, even in the modern world areas such as computing. In computing areas, binomial theorem has been very useful such as in distribution of IP addresses. With binomial distribution, the automatic allocation of IP addresses is possible.
- Another field that uses Binomial distribution as the important tools is the nation's economic prediction. Economists use binomial theorem to count probabilities to predict the way the economy will behave in the next few years. To be able to come up with realistic predictions, binomial theorem is used in this field.
- Binomial distribution has also been a great use in the architecture industry in design of infrastructure. It allows engineers, to calculate the magnitudes of the projects and thus delivering accurate estimates of not only the costs but also time required to construct them. For contractors, it is a very important tool to help ensuring the costing projects is competent enough to deliver profits.
- The binomial distribution is used when a researcher is interested in the occurrence of an event, not in its magnitude. For instance, in a clinical trial, a patient may survive or die. The researcher studies the number of survivors, and not how long the patient survives after treatment.
- Another example is whether a person is ambitious or not. Here, the binomial distribution describes the number of ambitious persons, and not how ambitious they are.
- Other situations in which binomial distributions arise are quality control, public opinion surveys, medical research, and insurance problems.

Poisson Distribution

The Poisson Distribution was developed by the French mathematician Simeon Denis Poisson in 1837. The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).If we let X = the number of events in a given interval and the mean number of events per interval as λ, then distribution of X is given by Poisson distribution.

A discrete random variable is said to follow Poisson Distribution with parameter λ, if its p.m.f is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0,1,2,\ldots.$$

$$= 0 \qquad \text{otherwise}$$

Mean = λ

Variance = λ

Comparison between Binomial and Poisson Distribution

Even though both Binomial and Poisson distribution give probability of X successes, the differences between binomial and Poisson distribution can be drawn clearly on the following grounds:

(1) The binomial distribution is one in which the probability of X successes among n Bernoulli trials is studied. A probability distribution that gives the probability of the count of a number of independent events that occur randomly within a given period, is called Poisson distribution.

(2) Binomial Distribution is biparametric, i.e. it is marked by two parameters n and p whereas Poisson distribution is uniparametric, i.e. described by a single parameter λ.

(3) There are a fixed number of attempts in the binomial distribution. On the other hand, an unlimited number of trials are there in a Poisson distribution.

 (4) The success probability is constant in binomial distribution but in Poisson distribution, there are an extremely small number of success.

(5) Poisson distribution can be considered as limiting form of binomial distribution. When success is a rare event i.e. probability of success p is small, p→0, number of trials is large i.e.

n→∞, but mean np is finite, binomial distribution tends to Poisson distribution.

(6) In binomial distribution Mean > Variance while in Poisson distribution mean = variance.

Apart from the above differences, there are a number of similar aspects between these two distributions i.e. both are the discrete theoretical probability distribution. Further, on the basis of the values of parameters, both can be unimodal or bimodal. Moreover, the binomial distribution and success probability (p) tends to 0 but A = np is constant. Infinity

Applications of Poisson distribution:

Poisson distribution is applied whenever we observe rare events. Some examples are given below:

- The number of deaths by horse kicking in the Prussian army (first application).
- Birth defects and genetic mutations.
- Rare diseases (like Leukemia, but not AIDS because it is infectious and so not independent). Car accidents.
- Traffic flow and ideal gap distance.
- Number of typing errors on a page. Hairs found in McDonald's hamburgers. Spread of an endangered animal in Africa.
- Failure of a machine in one month.

Continuous Random Variable

Continuous variable: A variable which assumes infinitely many values included in some range is defined as continuous random variable.

We call X a continuous random variable in a ≤ x ≤ b, if X can take on any value in this interval. An example of a random variable is the height of a person; say an adult male, selected randomly from a population. (This height typically takes on values in the range 1.64 < x < 9.84 feet, say, so a = 1.64 and b = 9.84.) If we select a male at random from a large population, and measure his height, the measured height can take on any real number within the interval of interest. This compels us to redefine our idea of a distribution, related to continuous variable, using a continuous function in place of the discrete function to represent probability of continuous variable.

A Continuous variable takes all possible values in a range set which is in the form of interval. On the other hand discrete random variable takes only specific or isolated values. Generally values of discrete random variable are obtained by counting while values of continuous variable are obtained by measuring to any degree of accuracy. The values of continuous variable move continuously from one possible value to another, without having to jump as in case of a discrete variable. E.g. Random variable X which is number of F.Y.B.Sc. students is discrete. While random variable Y which is height of a student is continuous variable. Continuous Variables would (literally) take forever to count. In fact, you would get to "forever" and never finish counting them. For example, take age. You can't count "age". Why not? Because it would literally take forever. For example, you could be: 20 years, 10 months, 2 days, 5 hours, 10 minutes, 4 seconds, 4 milliseconds, 8 nanoseconds, 99 picoseconds...and so on. Even though age is a continuous could turn age into a discrete variable and then you could count it. For example:

- A person's age in years.
- A baby's age in months.

Hence age, height, weight, time, income are continuous variables but we can turn them into discrete variable by appropriate definition.

Examples of continuous random variable:

- Consumption of cooking oil in a house.
- Waiting time on a ticket window.
- Life of an electronic gadget.
- Yield of a crop in certain area.
- Temperature in Mumbai
- Time it takes a computer to complete a task.

Properties of pdf of continuous random variable

i)   f(x)>0, for all x belongs to sample Space S since probability >0
ii) $\int_{-\infty}^{\infty} f(x)df = 1$ implied that total area bounded by the curve of density function and X – axis equal to 1, when computed over entire range of variable X.

Continuous Uniform Distribution

The simplest continuous probability distribution is the Uniform distribution also known as

Continuous Uniform distribution. It is observed in many situations. In general, if the probabilities of various classes of a continuous variable are more or less the same, the situation is best described by uniform distribution. In this distribution the p.d.f. of random variable remains constant over the range space of variable.

Definition:

A continuous random variable is said to follow uniform distribution over the interval (a, b), if

its p.d.f. is given by

$$f(x) = \frac{1}{b-a}; \quad a \leq x \leq b$$
$$= 0 \qquad \text{otherwise}$$

And is represented as X~ U(a, b)

Mean $= \frac{a+b}{2}$

Variance $= \frac{(b-a)^2}{12}$

Normal Distribution

Normal distribution is an important continuous distribution because a good number of random variables occurring in practice can be approximated to it. If a random variable is affected by many independent causes, and the effect of each cause is not overwhelmingly large as compared to other effects, then the random variable will closely follow a normal distribution.

Pioneers of Normal distribution:

Normal distribution was first mentioned by De-Moivre in 1733 and was also known to Laplace in 1774. Independently, the mathematicians Adrain in 1808 and Gauss in 1809 developed the formula for the normal distribution and showed that errors in astronomy were fit well by this distribution. Quételet and Galton were the first to apply the normal distribution to human as well as animal characteristics. He noted that characteristics such as height, weight, and strength were normally distributed. Anyhow, the credit of Normal distribution has been given to Gauss and is often called as Gaussian distribution. Normal distribution is the maximally used distribution in the theory of Statistics.

Definition:

A continuous random variable X is said to follow normal distribution with parameters μ and o^2, if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}; \quad -\infty < x < \infty; -\infty < \mu < \infty; \sigma > 0$$

It is denoted as X~N (μ, σ^2)

Descriptive Statistical measures

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics. The mean, or the average, is calculated by adding all the figures within the data set and then dividing by the number of figures within the set.

For example, the sum of the following data set is 20: (2, 3, 4, 5, 6). The mean is 4 (20/5). The mode of a data set is the value appearing most often, and the median is the figure situated in the middle of the data set. It is the figure separating the higher figures from the lower figures within a data set.

Central Tendency

Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables and general discussions to help people understand the meaning of the analysed data.

Measures of central tendency describe the centre position of a distribution for a data set. A person analyses the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analysed data set.

Measures of Variability

Measures of variability (or the measures of spread) aid in analysing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.

So, while the average of the data maybe 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).

MERITS AND DEMERITS OF MEAN

Merits:

1. It is rigidly defined.

2.It is easy to understand and easy to calculate.

3.It is based on all the observations.

4.It is capable of further algebraic treatment.

5.Of all the averages, A.M. is least affected by sampling fluctuations i.e. it is a stable average.

Demerits:

1. It cannot be obtained by mere inspection nor can it be located graphically.

2. It cannot be obtained even if a single observation is missing. It is affected by extreme values.

3. It is affected by extreme values

4.It cannot be calculated for frequency distribution having open end class- intervals e.g. class-intervals like below 10 or above 50 etc.

5.It may be a value which may not be present in the data.

6. Sometimes, it gives absurd results. e.g. Average number of children per family is 1.28.

7.It cannot be used for the study of qualitative data such as intelligence, honesty, beauty etc.

Even though A.M. has various demerits, it is considered to be the best all averages as it satisfies most of the requisites of a good average. A.M. is called the Ideal Average.

MERITS AND DEMERITS OF MEDIAN

Merits:

1.It is easy to understand and easy to calculate.

2.It is quite rigidly defined.

3.It can be computed for a distribution with open-end classes.

4.In majority of the cases, it is one of the values in the data.

5.It can be determined graphically.

6.Since median is a positional average, it can be computed even if the observations at the extremes are unknown.

7.It is not highly affected by fluctuations in sampling.

8.It can be calculated even for qualitative data.

Demerits:

1.When the number of observations is large, the pre-requisite of arranging observations in ascending/descending order of magnitude is a difficult process.

2. It is not based on all observations and hence, may not be a proper representative.

3. It is not capable of further mathematical treatment.

4. Since it does not require information about all the observation, it is insensitive to some changes

MERITS AND DEMERITS OF MODE

Merits:

1. It is easy to understand and simple to calculate.

2.It is not affected by extreme values or sampling fluctuations.

3.It can be calculated for distribution with open-end classes.

4.It can be determined graphically.

5.It is always present within the data and is the most typical value of the given set of data.

6. It is applicable to both, qualitative and quantitative data.

Demerits

1.It is not rigidly defined.

2.It is not based on all observations.

3.It is not capable of further mathematical treatment.

4.It is indeterminate if the modal class is at the extreme of the distribution.

5. If the sample of data for which mode is obtained is small, then such mode has no significance.

Comparison between Central Tendencies

We have studied five different measures of central tendency. It is obvious that no single measure can be the best for all situations. The most commonly used measures are mean, median and mode. It is not desirable to consider any one of them to be superior or inferior in all situations. The selection of appropriate measure of central tendency would largely depend upon the nature of the data; more specifically, on the scale of measurement used for representing the data and the purpose on hand.

The data obtained on nominal scale, we can count the number of cases in each category and obtain the frequencies. We may then be interested in knowing the class which is most popular or the most typical value in the data. In such cases, mode can be used as the appropriate measure of central tendency.

e.g. Suppose in a genetical study, for a group of 50 family members, we want to know most common colour of eyes. Then we count the number of persons for each different colour of eye. Suppose 3 persons have light eyes, 6 persons have brown eyes, 12 with dark grey eyes and 29 persons are with black eyes. Then the most common colour of eyes (i.e. mode) for this group of people is 'black'.

When the data is available on ordinal scale of measurement i.e. the data is provided in rank order, use of median as a measure of central tendency is appropriate. Suppose in a group of 75 students, 10 students have failed, 15 get pass class, 20 secure second class and 30 are in first class. The average performance of the students will be the performance of the middlemost student (arranged as per rank) i.e. the performance of 38th student i.e. second class; which is the median of the data. Median is only a point on the scale of measurement, below and above which lie exactly 50% of the data. Median can also be used (i) for truncated (incomplete) data, provided we know the total number of cases and their positions on the scale and (ii) when the distribution is markedly skewed.

Arithmetic Mean is the most commonly used measure of central tendency. It can be calculated when the data is complete and is represented on interval or ratio scale. It represents the centre of gravity of the data i.e. the measurements in any sample are perfectly balanced about the mean. In computation of simple A.M., equal importance is given to all observations in the data. It is preferred because of its high reliability and its applicability to inferential statistics. Thus, A.M. is more precise, reliable and stable measure of central tendency.

Range

Definition: If L is the largest observation in the data and S is the smallest observation, then range is the difference between L and S. Thus,

$$\text{Range = L-S}$$

For a frequency distribution, range may be considered as the difference between the largest and the smallest class-boundaries.

Range is a crude and simplest measure of dispersion. It measures the scatter of observations among themselves and not about any average.

The corresponding relative measure is

Coefficient of range $= \dfrac{L - S}{L + S}$

Note:

1.Range is a suitable measure of dispersion in case of small groups. In the branch of statistics known as Statistical Quality Control, range is widely used. It is also used to measure the changes in the prices of shares. Variation in daily temperatures at a certain place are measured by recording maximum temperature and minimum temperature. Range is also used in medical sciences to check whether blood pressure, haemoglobin count etc. are normal.

2. The main drawback of this measure is that it is based on only two extreme values, the maximum and the minimum, and completely ignores all the remaining observations.

Quartile Deviation

We have seen earlier that range, as a measure of dispersion, is based only on two extreme values and fails to take into account the scatter of remaining observations within the range. To overcome this drawback to an extent, we use another measure of dispersion called Inter-Quartile Range. It represents the range which includes middle 50% of the distribution. Hence,

Inter-Quartile Range $= Q_3 - Q_1$

where, Q_3 and Q_1 represent upper and lower quartiles respectively.

Half of Inter-Quartile-Range i.e. Semi-Inter-Quartile Range $= \frac{Q_3 - Q_1}{2}$ is also

used as absolute measure of dispersion. The semi-inter-quartile range is popularly known as Quartile Deviation (Q.D.)

$$QD = \frac{Q_3 - Q_1}{2}$$

The corresponding relative measure of dispersion is called coefficient of quartile deviation and is defined as

Coefficient of Q.D $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Note:

1. Q.D. is independent of extreme values. It is a better representative and more reliable than range.

2. Q.D. gives an idea about the distribution of middle half of the observations around the median.

3. Whenever median is preferred as a measure of central tendency, quartile deviation is preferred as a measure of dispersion. However, like median, quartile deviation is also not capable of further algebraic treatment, as it does not take into consideration all the values of the distribution.

4. For a symmetric distribution,

$Q_1 =$ Median - Q.D. and $Q_3 =$ Median + Q.D.

STANDARD DEVIATION (S.D.)

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical techniques.

Definition:

Standard Deviation (S.D.) is defined as the positive square root of the arithmetic mean of the squares of the deviations of the observations from their arithmetic mean.

The arithmetic mean of the squares of the deviations of the observations from their A.M. is called variance

Thus $SD = +\sqrt{variance}$

Note: The coefficient of variation is considered to be the most appropriate measure for comparing variability of two or more distributions. The importance of C.V. can be explained with the following example: Suppose milk bags are filled with automatic machine, the amount of milk being 1 litre per bag. Setting the machine for C.V. = 0 (i.e. zero variability) is impossible, because of chance causes which exist in any process and are beyond human control. Let us assume that the machine is set for C.V. less than or equal to 1. Then, using statistical law, we can expect approximately 99.73% of the bags to contain milk quantity ranging between atleast 970 mls. and at the most 1030 mls. Usually, this variation is not noticable and hence acceptable to customers But, if the machine is set for C.V. say equal to 5, we can see that about 16% of the bags will contain 900 mls. or less of milk, which is definitely not acceptable. Thus, one has to take utmost care to reduce C.V.

In manufacturing process, with reference to quality control section and in pharmaceutical industries, C.V. plays a very important role. In quality control section, efforts are made to improve the quality by producing items as per given specifications. The extent of deviation from given specifications can be measured using C. V. The lower is the value of C.V., better is the quality of the items produced. Due to competition, almost all industries have reduced the C.V. of their goods to a considerable extent in last few years.

In pharmaceutical industries, C.V. is as low as 1 or less than 1. The variation in the weights of tablets is almost negligible.

In industrial production, C.V. depends upon raw material used. A good quality of raw material will result in homogeneous end product. In chemical and pharmaceutical industries, C.V. can be reduced by thorough mixing and pounding of the raw material.

MERITS AND DEMERITS OF STANDARD DEVIATION

Merits:

1.It is rigidly defined.

2.It is based on all observations.

3.It is capable of further algebraic treatment.

4.It is least affected by sampling fluctuations.

Demerits:

1.As compared to other measures it is difficult to calculate.

2.It cannot be calculated for distribution with open-end class-intervals.

3.It gives more importance (weightage) to extreme values and less importance to the values close to A.M. that is unduly affected due to extreme observations

4.It cannot be calculated for qualitative data.

Sampling

When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a sampling method. There are two primary types of sampling methods that you can use in your research:

- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

First, you need to understand the difference between a population and a sample, and identify the target population of your research.

- The population is the entire group that you want to draw conclusions about.
- The sample is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, or many other characteristics.

It can be very broad or quite narrow: maybe you want to make inferences about the whole adult population of your country; maybe your research focuses on customers of a certain company, patients with a specific health condition, or students in a single school.

It is important to carefully define your target population according to the purpose and practicalities of your project.

If the population is very large, demographically mixed, and geographically dispersed, it might be difficult to gain access to a representative sample. A lack of a representative sample affects the validity of your results, and can lead to several research biases, particularly sampling bias.

Sampling frame

The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

Example

You are doing research on working conditions at a social media marketing company. Your population is all 1000 employees of the company. Your sampling frame is the company's HR database, which lists the names and contact details of every employee.
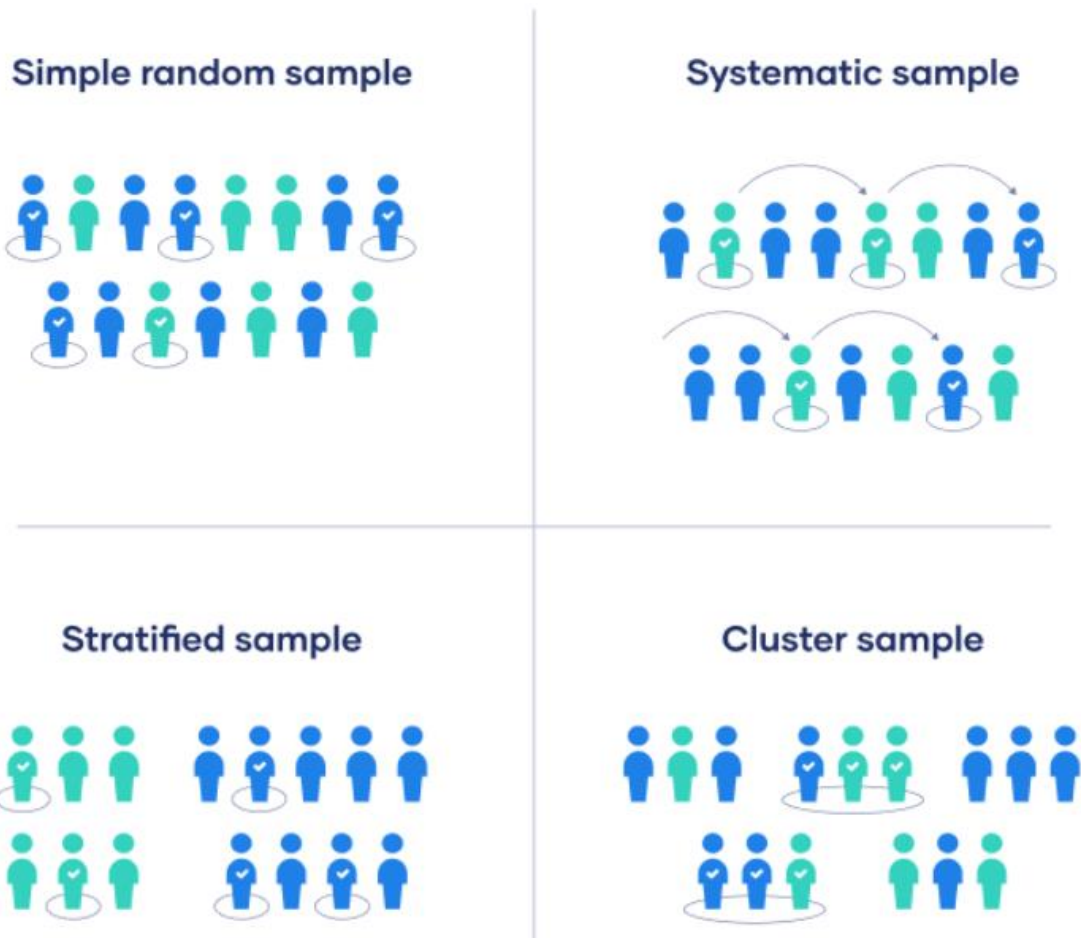
Sample size

The number of individuals you should include in your sample depends on various factors, including the size and variability of the population and your research design. There are different sample size calculators and formulas depending on what you want to achieve with statistical analysis.

Probability sampling methods

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.

1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example: Simple random sampling You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: Systematic samplingAll employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: Stratified sampling the company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: Cluster sampling the company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Non-probability sampling is a sampling method that uses non-random criteria like the availability, geographical proximity, or expert knowledge of the individuals you want to research in order to answer a research question.

Non-probability sampling is used when the population parameters are either unknown or not possible to individually identify. For example, visitors to a website that doesn't require users to create an account could form part of a non-probability sample.

Note that this type of sampling is at higher risk for research biases than probability sampling, particularly sampling bias.

Note

Be careful not to confuse probability and non-probability sampling.

- In **non-probability sampling**, each unit in your target population does not have an equal chance of being included. Here, you can form your sample using other considerations, such as convenience or a particular characteristic.
- In **probability sampling**, each unit in your target population must have an equal chance of selection.

Convenience sampling

Convenience sampling is primarily determined by convenience to the researcher.

This can include factors like:

- Ease of access
- Geographical proximity
- Existing contact within the population of interest

Convenience samples are sometimes called "accidental samples," because participants can be selected for the sample simply because they happen to be nearby when the researcher is conducting the data collection.

Example: Convenience samplingYou are investigating the association between daily weather and daily shopping patterns. To collect insight into people's shopping patterns, you decide to stand outside a major shopping mall in your area for a week, stopping people as they exit and asking them if they are willing to answer a few questions about their purchases.

Quota sampling

In quota sampling, you select a predetermined number or proportion of units, called a quota. Your quota should comprise subgroups with specific characteristics (e.g., individuals, cases, or organizations) and should be selected in a non-random manner.

Your subgroups, called strata, should be mutually exclusive. Your estimation can be based on previous studies or on other existing data, if there are any. This helps you determine how many units should be chosen from each subgroup. In the data collection phase, you continue to recruit units until you reach your quota.

TipYour respondents should be recruited non-randomly, with the end goal being that the proportions in each subgroup coincide with the estimated proportions in the population.

There are two types of quota sampling:

1. Proportional quota sampling is used when the size of the population is known. This allows you to determine the quota of individuals that you need to include in your sample in order to be representative of your population.

Example: Proportional quota samplingLet's say that in a certain company there are 1,000 employees. They are split into 2 groups: 600 people who drive to work, and 400 who take the train.
You decide to draw a sample of 100 employees. You would need to survey 60 drivers and 40 train-riders for your sample to reflect the proportion seen in the company.

2. Non-proportional quota sampling is used when the size of the population is unknown. Here, it's up to you to determine the quota of individuals that you are going to include in your sample in advance.

Example: Non-proportional quota samplingLet's say you are seeking opinions about the design choices on a website, but do not know how many people use it. You may decide to draw a sample of 100 people, including a quota of 50 people under 40 and a quota of 50 people over 40. This way, you get the perspective of both age groups.
Note that quota sampling may sound similar to stratified sampling, a probability sampling method where you divide your population into subgroups that share a common characteristic.

The key difference here is that in stratified sampling, you take a random sample from each subgroup, while in quota sampling, the sample selection is non-random, usually via convenience sampling. In other words, who is included in the sample is left up to the subjective judgment of the researcher.

Example: Quota samplingYou work for a market research company. You are seeking to interview 20 homeowners and 20 tenants between the ages of 45 and 60 living in a certain suburb.

You stand at a convenient location, such as a busy shopping street, and randomly select people to talk to who appear to satisfy the age criterion. Once you stop them, you must first

determine whether they do indeed fit the criteria of belonging to the predetermined age range and owning or renting a property in the suburb.

Sampling continues until quotas for various subgroups have been selected. If contacted individuals are unwilling to participate or do not meet one of the conditions (e.g., they are over 60 or they do not live in the suburb), they are simply replaced by those who do. This approach really helps to mitigate nonresponse bias.

Self-selection (volunteer) sampling

Self-selection sampling (also called volunteer sampling) relies on participants who voluntarily agree to be part of your research. This is common for samples that need people who meet specific criteria, as is often the case for medical or psychological research.

In self-selection sampling, volunteers are usually invited to participate through advertisements asking those who meet the requirements to sign up. Volunteers are recruited until a predetermined sample size is reached.

Self-selection or volunteer sampling involves two steps:

1. Publicizing your need for subjects
2. Checking the suitability of each subject and either inviting or rejecting them

Example: Self-selection samplingSuppose that you want to set up an experiment to see if mindfulness exercises can increase the performance of long-distance runners. First, you need to recruit your participants. You can do so by placing posters near locations where people go running, such as parks or stadiums.
Your ad should follow ethical guidelines, making it clear what the study involves. It should also include more practical information, such as the types of participants required. In this case, you decide to focus on runners who can run at least 5 km and have no prior training or experience in mindfulness.
Keep in mind that not all people who apply will be eligible for your research. There is a high chance that many applicants will not fully read or understand what your study is about, or may possess disqualifying factors. It's important to double-check eligibility carefully before inviting any volunteers to form part of your sample.

Snowball sampling

Snowball sampling is used when the population you want to research is hard to reach, or there is no existing database or other sampling frame to help you find them. Research about socially marginalized groups such as drug addicts, homeless people, or sex workers often uses snowball sampling.

To conduct a snowball sample, you start by finding one person who is willing to participate in your research. You then ask them to introduce you to others.

Alternatively, your research may involve finding people who use a certain product or have experience in the area you are interested in. In these cases, you can also use networks of people to gain access to your population of interest.

Example: Snowball samplingYou are studying homeless people living in your city. You start by attending a housing advocacy meeting, striking up a conversation with a homeless woman. You explain the purpose of your research and she agrees to participate. She invites you to a parking lot serving as temporary housing and offers to introduce you around.

In this way, the process of snowball sampling begins. You started by attending the meeting, where you met someone who could then put you in touch with others in the group.

When studying vulnerable populations, be sure to follow ethical considerations and guidelines.

Purposive (judgmental) sampling

Purposive sampling is a blanket term for several sampling techniques that choose participants deliberately due to qualities they possess. It is also called judgmental sampling, because it relies on the judgment of the researcher to select the units (e.g., people, cases, or organizations studied).

Purposive sampling is common in qualitative and mixed methods research designs, especially when considering specific issues with unique cases.

Note:

Unlike random samples—which deliberately include a diverse cross-section of ages, backgrounds, and cultures—the idea behind purposive sampling is to concentrate on people with particular characteristics, who will enable you to answer your research questions.

The sample being studied is not representative of the population, but for certain qualitative and mixed methods research designs, this is not an issue.