

Prof. (Dr.) R.M. KAPILA RATHNAYAKA



Fundamentals of Statistics



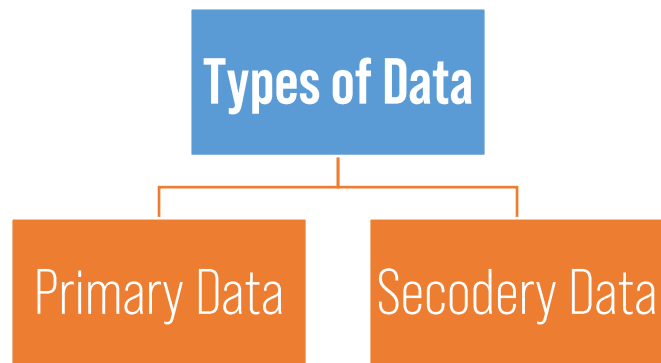
- Introduction to statistics
- Descriptive statistics
- Elementary Probability
- Random variables and Probability Distributions
- Special Probability Distributions
- Introduction to statistical software

PST 12209

Equation Lists

Equation 1: Arithmetic Mean	13
Equation 2: Direct Method (Arithmetic mean)	14
Equation 3: Sort-Cut Method (Arithmetic mean)	14
Equation 4: Coding method (Arithmetic mean)	15
Equation 5: Median formula for classified data	15
Equation 6: Quartiles (1st method)	16
Equation 7: Quartiles for classified data	16
Equation 8: Mode for classified data	17
Equation 9: Mean Deviation	17
Equation 10: Mean Deviation for classified data	18
Equation 11: Inter quartile range	18
Equation 12: Inter Quartile Deviation (semi-Interquartile Range)	18
Equation 13: Coefficient of quartile deviation	18
Equation 14: Sample Variance	19
Equation 15: Standard Deviation	19
Equation 16: The Karl Pearson's coefficient of skewness (1)	19
Equation 17: The Karl Pearson's coefficient of skewness (2)	19
Equation 18: Geometric Mean	21
Equation 19: Harmonic Mean	22
Equation 20: Weighted Arithmetic Mean (Z-Score)	22
Equation 21: Probability	24
Equation 22: Law of Algebra of Sets	26
Equation 23: Conditional Probability	28
Equation 24: Independent events	29
Equation 25: Multiplication Rule-01	29
Equation 26: Multiplication Rule -02	29
Equation 27: The Law of Total Probability	30
Equation 28: Bayes' Theorem	31
Equation 29: Properties of probability density function	33
Equation 30: Continuous Probability Distribution	34
Equation 31: Mathematical Expectation (Mean Value of the probability Distribution)	36
Equation 32: If Random variable becoming a function	36
Equation 33: Relationship between $E[g(x)]$ and $E(x)$	36
Equation 34: Variance in Random Variables	38
Equation 35: Variance Equation (Simplified)	38
Equation 36: Relationship between $\text{Var}(aX+b)$ and $\text{Var}(x)$	38
Equation 37: Summarization of Expectation and Variance	39
Equation 38: Discrete Uniform distribution for Equal probabilities variables	41
Equation 39: Mean and Variance for Discrete Uniform Distribution	41
Equation 40: Probability density function	42
Equation 41: Mean & Variance for the distribution	43
Equation 42: Probability Distribution for Poisson distribution	44
Equation 43: Means and Variance for Poisson Distribution	44
Equation 44: Binomial to Poisson Distributions	45
Equation 45: mean & standard deviation for Uniform Distribution	46
Equation 46: Normal probability density function for a continuous random variable	47
Equation 47: Relationship between variable Z and variable X, μ and σ	48

Sources of data



- The data that you collect may be either primary data or secondary data.
- It can be gathered by organizations using experiments or surveys, or by individual workers.
- The main difference between primary and secondary data is related to the way that the data itself is gathered.

Primary data

- When you create the data you want by yourself, it's called primary data.
- Data observed or collected directly from first-hand experience is called primary data. Also known as raw data.
- In primary data collection, you collect the data by yourself using methods such as ;
 - Questionnaires
 - Interviews
 - Diaries
 - Portfolios

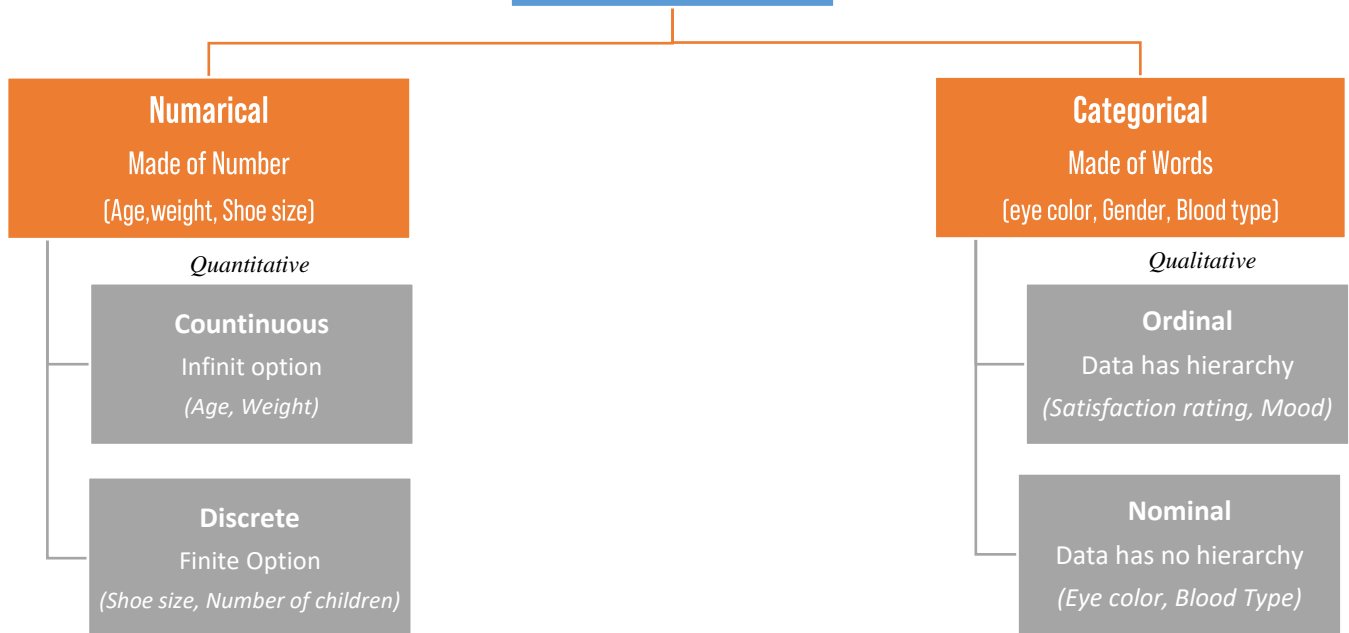


Secondary data

Published data and the data collected in the past or other parties is called secondary data. Some examples of Secondary sources

- Newspapers and popular magazine articles. (may also be Primary)
- Dictionaries and encyclopedias
- Organizational records
- Social science include censuses

Data



Numerical Data (Data that is Numbers):

Continuous Variables

Continuous variables are a variable whose value is obtained by measuring.

Ex:

- height of students in class
- weight of students in class
- time it takes to get to school
- distance traveled between classes

Discrete Variables

A discrete variable is a variable whose value is obtained by counting. All continuous variables are numeric, but not all numeric variables are continuous.

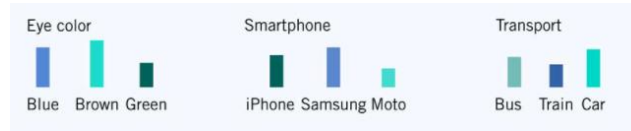
Ex:

- number of students present
- number of red marbles in a jar
- number of heads when flipping coin
- students' grade level

Categorical Data (Data that is not numbers)

Nominal Variable

Nominal variables are basically labels. Think of them as categories with no specific order. Examples include things like colors (red, blue, green), types of fruit (apple, banana, orange), or even gender (male, female, non-binary). With nominal variables, there's no inherent ranking or order to the categories.



Ordinal Variable

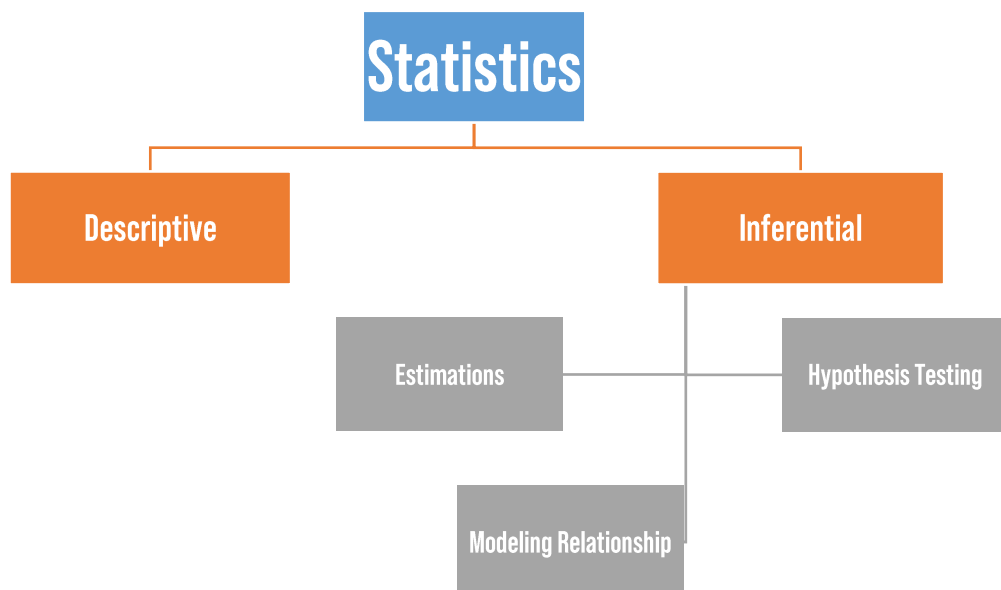
Ordinal variables, on the other hand, are categories that do have a specific order or ranking. Examples of ordinal variables include things like t-shirt sizes (small, medium, large), education levels (high school, bachelor's, master's), or even satisfaction ratings (very unsatisfied, unsatisfied, neutral, satisfied, very satisfied).



Kinds of Statistics

We can divide statistics in to two parts.

- Descriptive statistics
- Inferential statistics



Population

A population is the set of all the individuals of interest in a particular study.

Ex:

- Advertisements for IT jobs in the Sri Lanka
- Songs from the VOICE Song Contest
- Undergraduate students in SUSL
- All countries of the world



Mainly the term population can be divided in to two parts.

- Finite population
- Infinite population

Finite population

If a population consists of fixed number of values, then it is said to be finite.

Ex: Number of days per month.

Infinite population

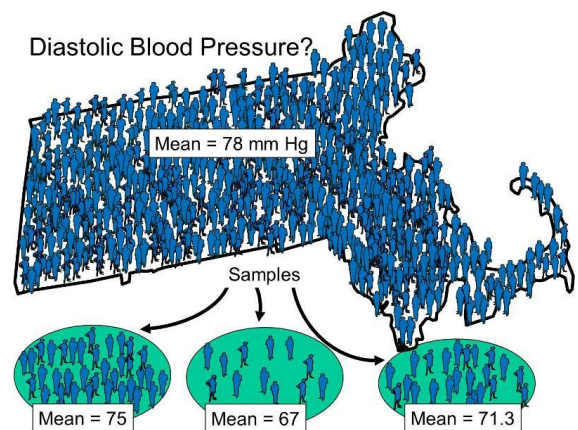
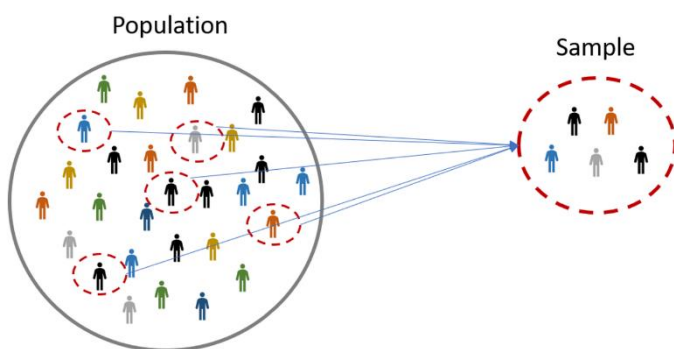
If a population consists of an endless succession of values, it is said to be infinite.

Ex: Number of insects in a certain region.

Sample

A sample is a set of data drawn from the population. (A sample is a small segment of the population)

- Potentially very large, but less than the population.
- In other words, a sample is a subset of a population.



Population has Parameters, Samples have Statistics.

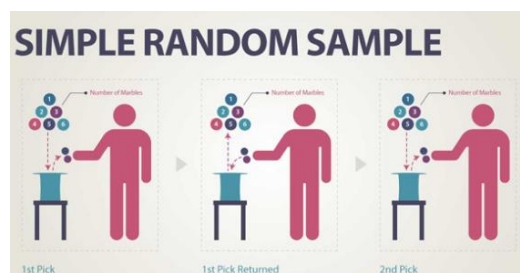
Reason for Sampling

- **Necessity:** Sometimes it's simply not possible to study the whole population due to its size or inaccessibility.
- **Practicality:** It's easier and more efficient to collect data from a sample.
- **Cost-effectiveness:** There are fewer participant, laboratory, equipment, and researcher costs involved.
- **Manageability:** Storing and running statistical analyses on smaller datasets is easier and reliable

Note:

Random Sample

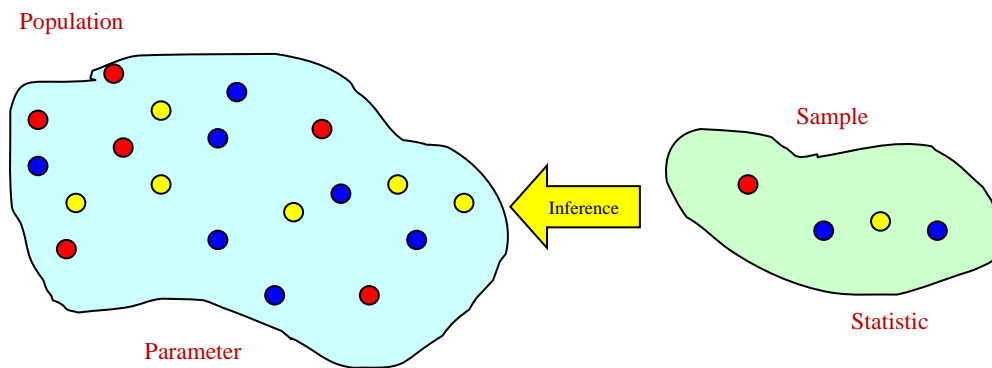
It is a sample chosen in a very specific way and has been selected in such a way that every element in the population has an equal opportunity of being included in the sample.



Population	Sample
Advertisements for IT jobs in the Sri Lanka	The top 50 search results for advertisements for IT jobs in the Sri Lanka on January 1, 2021
Songs from the VOICE Song Contest	Winning songs from the VOICE Song Contest that were performed in English
Undergraduate students in the Sri Lanka	300 undergraduate students from three Sri Lanka universities who volunteer for your psychology research study
All countries of the world	Countries with published data available on birth rates and GDP since 2000

Statistical Inferences

Statistical inference is the process of making an estimate, prediction, or decision about a population based on a sample.



Inferential statistics is used to draw conclusions or inferences about characteristics of populations based on data from a sample.

Characteristics

Characteristics are of two kinds.

- **Attributes:** Attributes are the non-measurable characteristics which cannot be numerically express in terms of units. These are qualitative objects.
Ex: Religion, nationality, literacy.
- **Variables:** A variable is a measurable characteristic that changes or has different values for different individuals.
Ex: weight of a person, height of a student.

Variables

Variables are of two kinds.

- Discrete variable
- Continuous variable

Discrete variable

A discrete variable consists of separate indivisible categories. No values can exist between two neighboring categories.

Ex: No. of books, no. of students.

Continuous variable

For a continuous variable there are an infinite number of possible values that fall between any two observed values. A continuous variable is divisible in to an infinite number of fractional parts.

Ex: Life time of an insect, life time of a bulb, weight of a child.

CLASSIFICATION AND TABULATION OF DATA

Difference Between Classification and Tabulation

Basis for comparison	Classification	Tabulation
Meaning	Classification is the process of grouping data into different categories, on the basis of nature, behavior, or common characteristics.	Tabulation is a process of summarizing data and presenting it in a compact form, by putting data into statistical table.
Order	After data collection	After classification
Arrangement	Attributes and variables	Columns and rows
Purpose	To analyze data	To present data

Definition Of classification

Classification refers to a process, wherein data is arranged based on the characteristic under consideration, into classes, groups, as per resemblance of observations. Classification puts the data in a condensed form, as it removes unnecessary details that helps to easily comprehend data.

The classification of data reduces the large volume of raw data into homogeneous groups, i.e. data having common characteristics or nature are placed in one group.

There are four types of classification

- Qualitative Classification or Ordinal Classification
- Quantitative Classification
- Chronological or Temporal Classification
- Geographical or Spatial Classification

Definition Of Tabulation

Tabulation refers to a logical data presentation, where in raw data is summarized and displayed in a compact form, i.e. in statistical tables.

After data classification is over, those data are represented in a tabular form which can be called tables.

Tables are made up of various numbers of rows and columns.

Type of Tabulations

There are three types of tabulation:

Simple Tabulation:

It can also be called a One-way tabulation.

This happens when the tabulation of data is done based on only one individual characteristic.

For example: Tabulating the data based on only one characteristic like height, weight, religion, etc.

Double Tabulation:

It is also known as two-way tabulation.

This happens when the tabulation of data is done based on two characteristics.

For example: Tabulating the based on two characteristics like height and weight etc.

Complex Tabulation:

When the tabulation of data is done based on two or more characteristics.

For example: Tabulating the data on height, weight, and age, etc.

Simple Tabulation

The Nicotine contents, in milligrams for 50 cigarettes (sample) of a certain brand were selected randomly from their population (1000) and recorded in Table (01).

31	22	25	24	28
41	43	47	34	38
04	08	16	28	21
36	27	33	20	33
26	31	23	27	18
22	29	28	34	36
29	42	39	28	49
29	14	27	24	09
31	35	48	35	27
38	36	11	32	27

Step 01: Arranged in an array

Arranged the data set in ascending order (smallest to largest) or descending (largest to smallest) order. When the dataset is arranged in order from smallest to largest or largest to smallest, it is known as array.

Step 02: Calculate Range

Difference between the largest and smallest observations.

The Sample Range (r) = $\max(x_i)$ (Largest number) – $\min(x_i)$ (Smallest Number).

Step 03: Estimates the number of intervals

It is desirable to have class intervals (of the same length if desired) to categorize data.

The number of class intervals depends on the data set.

The number of class intervals,

- can be guessed.
- can be taken in between 5-20.
- can be taken as the smallest integer k such that $2^k \geq n$; where n is the sample size.

ex: Let $n=50$, then, so $k=6$.

This means we can divide the data set in to 6 class intervals.

Step 04: Estimate the Class Width

- Approximate class width can be found as follows

$$\begin{aligned}\text{The Class width} &= \text{The Range} / \text{Number of classes to be own} \\ &= 45 / 6 = 7.5 \approx 8\end{aligned}$$

- This can be rounded off to a convenience number.

Step 05: Class Interval

Class	Class Interval
1	3-10
2	11-18
3	19-26
4	27-34
5	35-42
6	43-50

Step 06: Frequency Table

The number of observations in any particular class is called the class frequency of the data.

Class	Class Interval	Tolly Marks	Frequency
1	3-10	///	3
2	11-18	////	4
3	19-26	### ///	9
4	27-34		
5	35-42		
6	43-50		

Lower class limit

Upper class limit

Step 07: Class Boundary

Lower Class boundary = $\frac{1}{2}$ (Upper class limit of lower class + Lower class limit of given class)

Upper Class boundary = $\frac{1}{2}$ (Lower class limit of upper class + Upper class limit of given class)

Class	Class Interval	Class Boundary	Frequency
1	3-10	2.5-10.5	3

Step 08: Mid-point

Mid-point = $\frac{1}{2}$ (Upper Class boundary + Lower Class boundary)

The class mid-point is called the class mark.

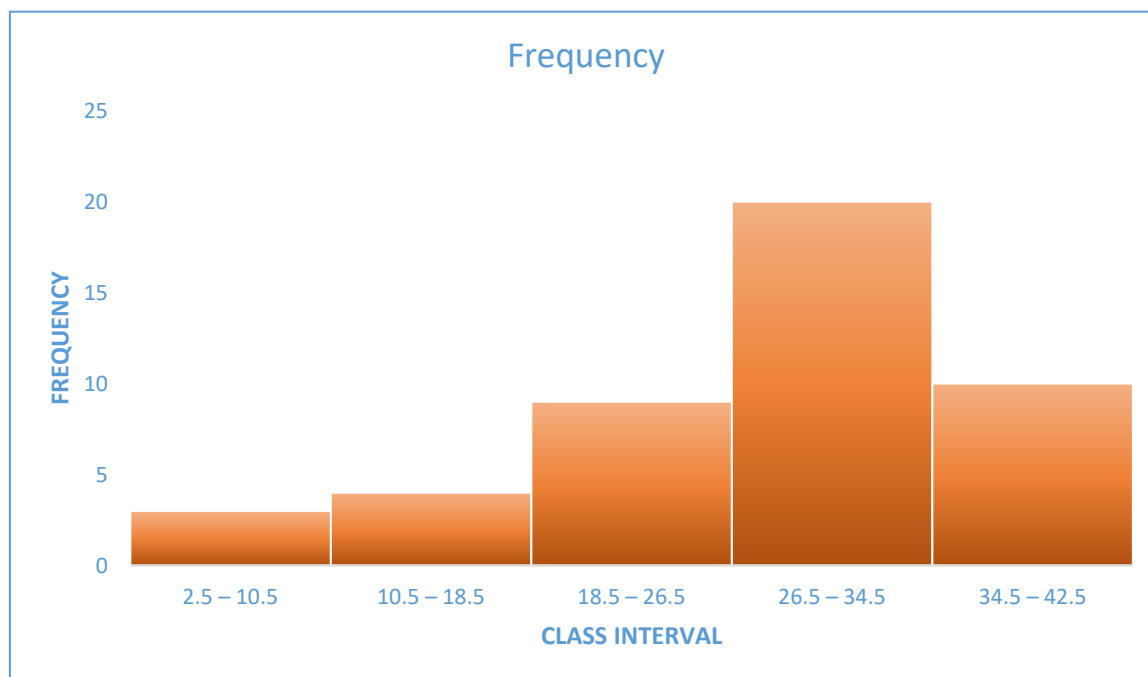
Class	Class Interval	Class Boundary	Mid-Value	Frequency
1	3-10	2.5-10.5	6.5	3

Graphical Representation of Numeric Data

Histograms

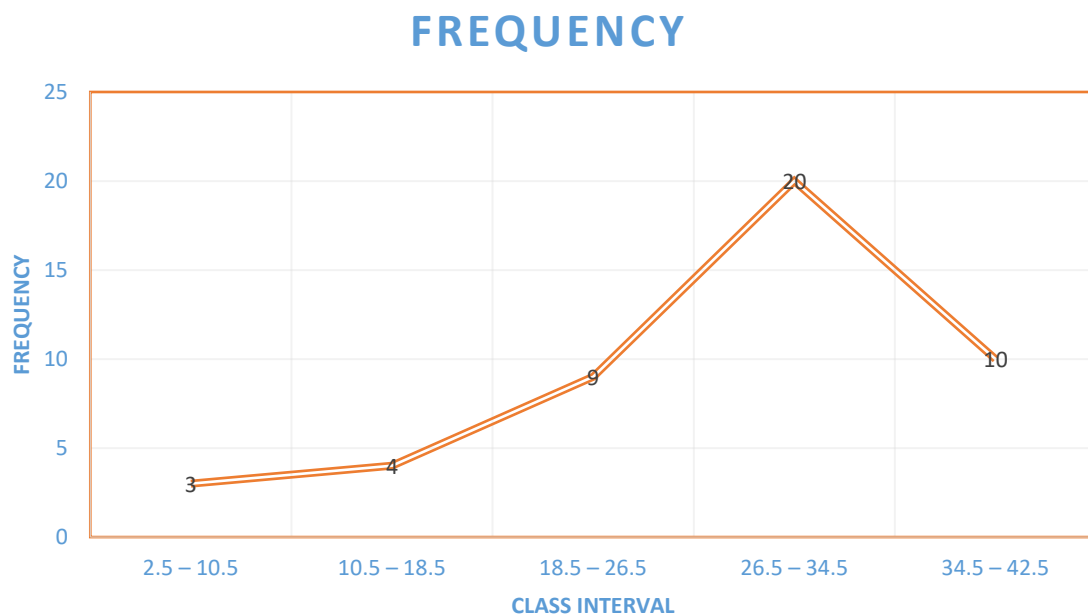
- This is most common form of diagrammatic representation of grouped frequency distribution of both continuous and discontinuous type.
- It consists of a set of rectangles drawn on a horizontal base line. i.e. frequency is marked on the vertical line.
- The area of each rectangle is proportional to the frequency in the respective class interval.
- The size of class intervals is drawn on x-axis with equal width and their respective frequencies on y-axis.

Class	Class Interval	Class Boundaries	Mid-Values	Frequency
1	3 – 10	2.5 – 10.5	6.5	3
2	11 – 18	10.5 – 18.5	14.5	4
3	19 – 26	18.5 – 26.5	22.5	9
4	27 – 34	26.5 – 34.5	30.5	20
5	35 – 42	34.5 – 42.5	38.5	10



Frequency Polygon

- This is another way of displaying data graphically. In the horizontal axis(X), class midpoints are marked.
- A dot is placed at the mid – point of each top of the rectangles corresponding to class marks.
- Dots are connected by the straight lines
- To close the polygon, dots are placed on the X axis, one half class interval to the left lower class, and one-half class interval to the right of the highest-class interval.



Other graphical Representation

- Bar Chart
- Pie charts

Cumulative frequency

Cumulative frequency is the number of observations falling up to that value or range.

Ex-: Consider the following frequency table.

Class	Class Interval	Frequency	Cumulative frequency
1	$03 \leq x < 10$	3	3
2	$10 \leq x < 18$	4	7
3	$18 \leq x < 26$	9	16
4	$26 \leq x < 34$	20	36
5	$34 \leq x < 42$	10	46

MEASURES OF CENTRAL TENDENCY

Introduction

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.

As such, measures of central tendency are sometimes called measures of central location.

They are also classed as summary statistics. In addition to central tendency, the variability and distribution of your data set is important to understand when performing descriptive statistics.

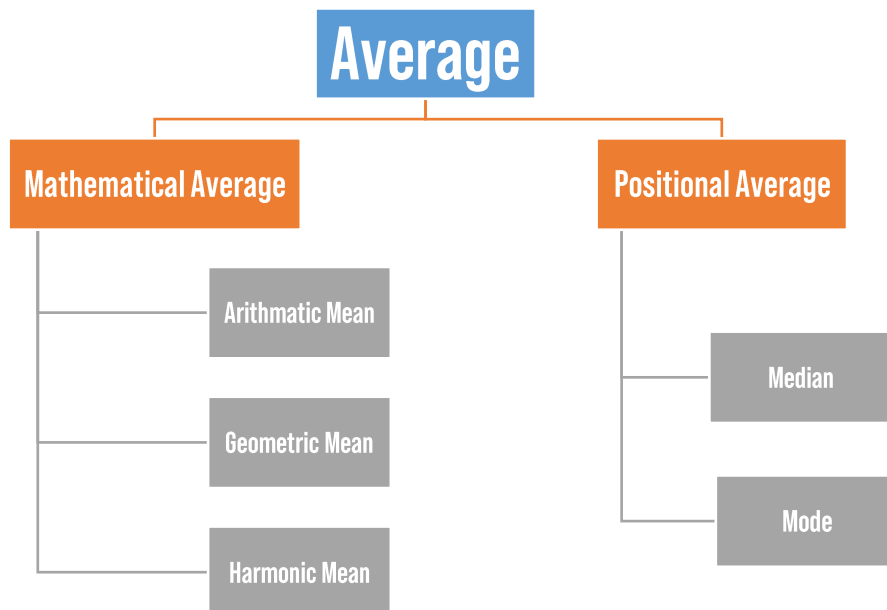
In statistics, a central tendency (or, more commonly, a measure of central tendency) is a central or typical value for a probability distribution.

It may also be called a center or location of the distribution. It helps you find the middle, or the average, of a data set.

The most common measures of central tendency are the

- Arithmetic mean
- Median
- Mode

A central tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution.



THE ARITHMETIC MEAN (MEAN/ AVERAGE)

This is the most widely used measure of central tendency of any group of data.

If x_1, x_2, \dots, x_n are sample observations for a variable x then the sample mean (arithmetic mean) is usually denoted by;

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Equation 1: Arithmetic Mean

Grouped Data and Calculations

There are three methods to find the arithmetic mean in grouped data.

- Direct method
- Short-cut method
- Coding method or step-deviation method.

Direct Method

In the case of continuous series, each class frequency is multiplied by the mid-value of the class-interval, the products added together and the total divided by the number of observations.

If numbers x_1, x_2, \dots, x_n occurs f_1, f_2, \dots, f_n time (frequency) respectively.

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i}$$

Equation 2: Direct Method (Arithmetic mean)

Short-cut method

$$\bar{x} = A + \frac{\sum_{i=1}^n f_i d_i}{\sum_{i=1}^n f_i}$$

$$d_i = x_i - A$$

A = Assumed Mean

d_i = Deviation of Mid-point

Equation 3: Sort-Cut Method (Arithmetic mean)

Class Boundaries	Mid-Values	Deviation	Frequency	$f_i d_i$
2.5 – 10.5	6.5	16	3	48
10.5 – 18.5	14.5	8	4	32
18.5 – 26.5	22.5	0	9	0
26.5 – 34.5	30.5	-8	20	-160
34.5 – 42.5	38.5	-16	10	-160
Totals			46	-240

$$\begin{aligned}\bar{x} &= A + \frac{\sum_{i=1}^n f_i d_i}{\sum_{i=1}^n f_i} \\ &= 22.5 + \frac{-240}{46} \\ \bar{x} &= 17.2826\end{aligned}$$

Coding method or Step – deviation method

The above short-cut method can further be simplified in practice by what is known as coding method. The deviations from the assumed mean are divided by a common factor to reduce their size.

Class Boundaries	Mid-Values	d_i	$u_i=d_i/c$	f_i	$f_i u_i$
2.5 – 10.5	6.5	16	2	3	6
10.5 – 18.5	14.5	8	1	4	4
18.5 – 26.5	22.5	0	0	9	0
26.5 – 34.5	30.5	-8	-1	20	-20
34.5 – 42.5	38.5	-16	-2	10	-20
Total				46	-30

$$\bar{x} = A + \frac{C \sum_{i=1}^n f_i u_i}{\sum_{i=1}^n f_i}$$

$$d_i = x_i - A$$

C = Class width

A = Assumed Mean

$$u_i = \frac{d_i}{C}$$

Equation 4: Coding method (Arithmetic mean)

$$\begin{aligned}\bar{x} &= A + \frac{C \sum_{i=1}^n f_i u_i}{\sum_{i=1}^n f_i} \\ &= 22.5 + \frac{8 \times -30}{46} \\ &= 17.2826\end{aligned}$$

Sample Median

When the items of a series are arranged in ascending or descending order of magnitude, the value of the middle item in the series is known as median in the case of individual observations.

If the total number of items in a series are odd then the value of the $\left(\frac{n+1}{2}\right)^{th}$ item gives the median.

On the other hand, if the total number of items in a series are even then the value of the $\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$ item gives the median.

For grouped data the Median is obtained by the following formula:

$$\text{Median} = L_1 + \frac{\frac{N}{2} - \sum f_m}{f_{\text{median}}}$$

Equation 5: Median formula for classified data

L_1 = Lower class boundary of median class

N = Total number of Data

f_{median} = Frequency of median class

$\sum f_m$ = Sum of frequencies of all classes lower than median class

Quartiles (1st method)

They are the points that divide the ordered (*ascending or descending*) data set in to three equal points. They denoted by,

$$Q_1 = \left(\frac{n+1}{4} \right)^{th} \text{ item} \quad Q_2 = \left(\frac{n+1}{2} \right)^{th} \text{ item} \quad Q_3 = \left(\frac{3[n+1]}{4} \right)^{th} \text{ item}$$

Equation 6: Quartiles (1st method)

Quartiles for classified data

$$Q_1 = L_1 + \left(\frac{\frac{N}{4} - \sum f_{Q_1-1}}{f_{Q_1}} \right) C$$

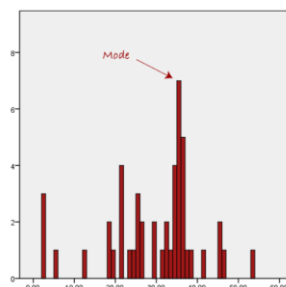
$$Q_2 = L_1 + \left(\frac{\frac{N}{2} - \sum f_{Q_2-1}}{f_{Q_2}} \right) C = \text{median}$$

$$Q_3 = L_1 + \left(\frac{\frac{3N}{4} - \sum f_{Q_3-1}}{f_{Q_3}} \right) C$$

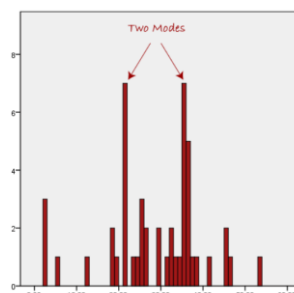
Equation 7: Quartiles for classified data

Mode

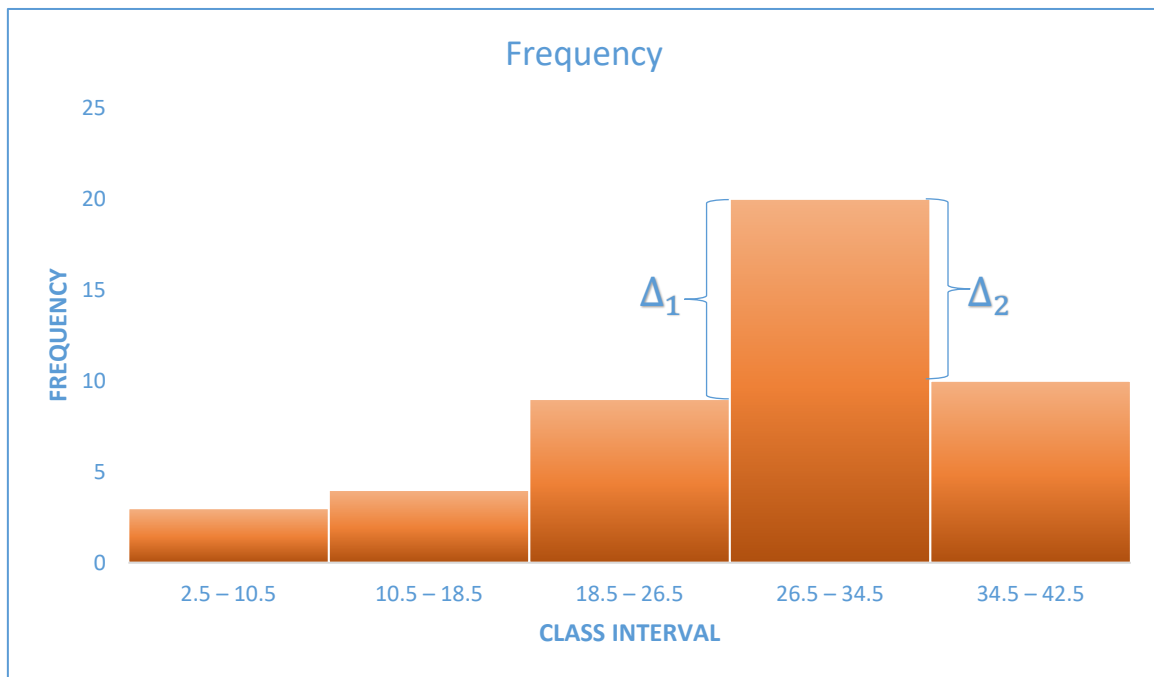
The mode of the set of number is that value which occurs most frequently.



However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency, such as below:



Finding mode in classified data



$$Mode = L_1 + C \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right)$$

Equation 8: Mode for classified data

MEASURES OF VARIABILITY OR SPREAD OR DISPERSION OF A DATA SET

Measures of Variability

A measure of variability is a summary statistic that represents the amount of dispersion in a dataset. Variability refers to how "spread out" a group of scores is. In statistics, variability, dispersion, and spread are synonyms that denote the width of the distribution.

Range

Range is the simplest method of studying dispersion. It is simply the distance between the largest (L) and the smallest (S) value in a group of items.

Mean deviation

The deviations of the variety of values from the mean are another method of measuring variability.

We have numbers x_1, x_2, \dots, x_n Then

$$Mean\ deviation = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Equation 9: Mean Deviation

The mean Deviation for classified data

$$\text{Mean deviation} = \frac{\sum_{i=1}^n f_i |x_i - \bar{x}|}{n}$$

Equation 10: Mean Deviation for classified data

Inter quartile range

$$I_{QR} = Q_3 - Q_1$$

Equation 11: Inter quartile range

Inter Quartile Deviation (semi-Interquartile Range)

The dependence of range on extreme items can be avoided by adopting this measure.

$$Q_D = \frac{(Q_3 - Q_1)}{2}$$

Equation 12: Inter Quartile Deviation (semi-Interquartile Range)

Coefficient of quartile deviation

Quartile deviation is an absolute measure of dispersion. The relative measures corresponding to quartile deviation is called the coefficient of quartile deviation and is expressed as follows.

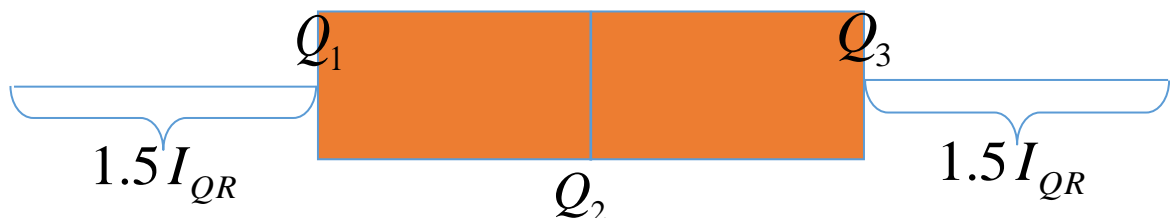
$$CQD = \frac{(Q_3 - Q_1)}{(Q_3 + Q_1)}$$

Equation 13: Coefficient of quartile deviation

Box plot

The box plot is a graphical display, that simultaneously displays important features of the data, such as,

- location
- Central Tendency (Median)
- Spread or variability (Variance and Standard deviation, Inter quartile Range)
- unusually observations (outliers)



Standard Deviation and Variance

A commonly used measure of dispersion is the standard deviation, which is simply the square root of the variance. The variance and the standard deviation are measures of how spread out a distribution is. In other words, they are measures of variability.

Variance

Population

Sample

Sample Variance

$$\text{Variance}(s^2) = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n f_i x_i^2 - \frac{\left(\sum_{i=1}^n f_i x_i\right)^2}{n}}{n-1}$$

Equation 14: Sample Variance

Standard Deviation

$$S = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^n f_i x_i^2 - \frac{\left(\sum_{i=1}^n f_i x_i\right)^2}{n}}{n-1}}$$

Equation 15: Standard Deviation

Measures of Skewness

Skewness, like measures of central value and dispersion is a measure for the study of frequency distributions. Thus, skewness measures the degree of departure from symmetry. When a distribution is symmetrical, skewness is absent and the values of the mean, median and mode coincide. The mean and median are pulled away from the mode, either to the right or to the left. When mean and median are pulled towards right, the skewness is positive and otherwise negative.

There are three important measures of skewness.

- The Karl Pearson's coefficient of skewness
- The bowly's coefficient of skewness
- Coefficient of skewness, based on moments

The Karl Pearson's coefficient of skewness

The Karl Pearson's coefficient of skewness is denoted by,

$$S_{kp} = \frac{\text{mean} - \text{mode}}{\text{Standard Deviation}}$$

Equation 16: The Karl Pearson's coefficient of skewness (1)

It is also known as Pearson's coefficient of skewness. If the mode is ill defined,

$$S_{kp} = \frac{3(\text{mean} - \text{median})}{\text{Standard Deviation}}$$

Equation 17: The Karl Pearson's coefficient of skewness (2)

- If $S_{kp} = 0$, the distribution of data is symmetric
- If $S_{kp} > 0$, the distribution of data is skewed to the Right
- If $S_{kp} < 0$, the distribution of data is skewed to the Left

EXERCISE

The Nicotine contents, in milligrams for 50 cigarettes (sample) of a certain brand were selected randomly from their population (1000) and recorded).

31	22	25	24	28
41	43	47	34	38
04	08	16	28	21
36	27	33	20	33
26	31	23	27	18
22	29	28	34	36
29	42	39	28	49
29	14	27	24	09
31	35	48	35	27
38	36	11	32	27

Class Interval	Frequency(f)	Mid-Value(x)	fx	fx ²	c.f
02.5-10.5	03	6.5	19.5	126.75	03
10.5-18.5	04	14.5	58	841	07
18.5-26.5	09	22.5	202.5	4556.25	16
26.5-34.5	20	30.5	610	18605	36
34.5-42.5	10	38.5	385	14822.5	46
42.5-50.5	04	46.5	186	8649	50
Total			1461	47600.5	

Calculate the,

1. Sample means

$$\begin{aligned}
 \text{Mean} &= \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} \\
 &= \frac{1461}{50} \\
 &= 29.22
 \end{aligned}$$

3. Mode

$$\begin{aligned}
 \text{Mode} &= L_1 + C \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \\
 &= 26.5 + 8 \left(\frac{11}{11 + 10} \right) \\
 &= 30.69
 \end{aligned}$$

2. Median

$$\begin{aligned}
 \text{Median} &= L_1 + C \left(\frac{\frac{n}{2} - \sum_{i=1}^n f_{Q_i-1}}{f_{Q_i}} \right) \\
 &= 26.5 + 8 \left(\frac{25 - 16}{20} \right) \\
 &= 30.1
 \end{aligned}$$

4. sample variance and standard deviation.

$$\begin{aligned}
 \text{Variance}(S^2) &= \frac{\sum_{i=1}^n f_i x_i^2 - \left(\sum_{i=1}^n f_i x_i \right)^2}{n - 1} \\
 &= \frac{47600.5 - \frac{1461^2}{50}}{49} \\
 &= 100.206 \\
 \text{Standard deviation} &= \sqrt{\text{Variance}} \\
 &= 10.01
 \end{aligned}$$

5. Q1 and Q3

$$Q_1 = L_1 + C \left(\frac{\frac{n}{4} - \sum_{i=1}^n f_{Q_1-1}}{f_{Q_1}} \right)$$

$$= 18.5 + 8 \left(\frac{12.5 - 7}{9} \right)$$

$$= 23.389$$

$$Q_3 = L_1 + C \left(\frac{\frac{3n}{4} - \sum_{i=1}^n f_{Q_3-1}}{f_{Q_3}} \right)$$

$$= 34.5 + 8 \left(\frac{37.5 - 36}{10} \right)$$

$$= 35.7$$

6. Discuss the skewness

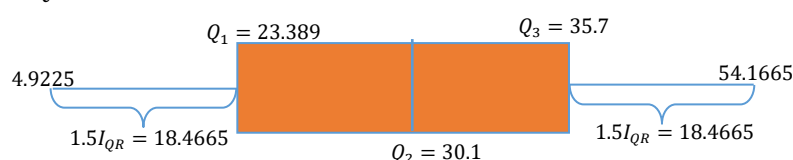
$$S_{kp} = \frac{\text{Mean-Mode}}{\text{Standard Deviation}}$$

$$= \frac{29.22 - 30.69}{10.01}$$

$$= -0.146853147$$

7. Draw a box plot and identify the outliers

$$I_{QR} = Q_3 - Q_1 = 35.7 - 23.389 = 12.311$$



GEOMETRIC MEAN

The geometric mean is most useful when numbers in the series are dependent of each other or if numbers tend to make large fluctuations.

Geometric Mean is well defined only for sets of positive real numbers.

Applications of the geometric mean are most common in business and finance, where it is frequently used when dealing with percentages to calculate growth rates and returns on a portfolio of securities.

This is calculated by multiplying all the numbers (call the number of numbers n), and taking the nth root of the total.

$$\text{Geometric Mean} = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

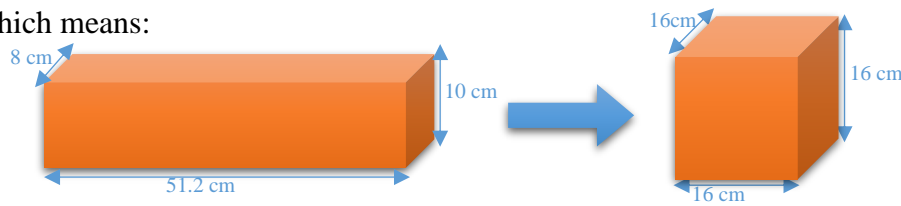
Equation 18: Geometric Mean

Example:

What is the Geometric Mean of 10, 51.2 and 8?

$$\text{Geometric mean} = \sqrt[3]{10 \times 51.2 \times 8} = 16$$

Which means:



HARMONIC MEAN

The harmonic mean is a very specific type of average. It's generally used when dealing with averages of units, like speed or other rates and ratios. Add the reciprocals of the numbers in the set. Divide the number of items in the set.

For General Dataset.....

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

For Classified Dataset.....

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^n \frac{f_i}{x_i}}$$

Equation 19: Harmonic Mean

Example: we travel 10 km at 60 km/h, then another 10 km at 20 km/h, what is our average speed?

$$\text{Harmonic Mean} = \frac{20 \text{ km}}{\frac{10 \text{ km}}{60 \text{ km/h}} + \frac{10 \text{ km}}{20 \text{ km/h}}} = 30 \text{ km/h}$$

Check: The 10 km at 60 km/h takes 10 minutes, the 10 km at 20 km/h takes 30 minutes, so the total 20 km takes 40 minutes, which is 30 km per hour

WEIGHTED ARITHMETIC MEAN

The weighted arithmetic mean is calculated after assigning appropriate weights to the values of the data. If all the weights are equal, then the weighted mean is the same as the arithmetic mean.

$$\bar{x} = \frac{\sum_{i=1}^n w_i \times x_i}{\sum_{i=1}^n w_i}$$

w_i = The weights
 x_i = The Values

Equation 20: Weighted Arithmetic Mean (Z-Score)

Example:

Given that $A = 4$, $B+ = 3.4$, $B = 3$, $C+ = 2.5$, $C = 2$, $F = 0$, Determine the grade point average carried by a student in particular semester, based on the following grades.

Course	Grade	Credit Hours (w)	x	wx
Accounting	A	4	4	16
Finance	$C+$	2	2.5	5
Marketing	$B+$	3	3.4	10.5
Statistics	B	4	3	12
Management	C	3	2	6
		16		49.5

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^n w_i \times x_i}{\sum_{i=1}^n w_i} \\ &= \frac{49.5}{16} \\ &= 3.094\end{aligned}$$

ELEMENTARY PROBABILITY THEORY

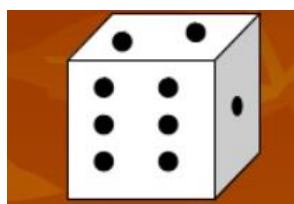
Probability means possibility. It measures of ‘uncertainty’, randomness or likelihood of occurrence of events. The meaning of probability is basically the extent to which something is likely to happen. There are several ways for determining the probability. Usually, we use the Classical method to obtain the probability of simple or basic events. In the classical method, probability of an event is calculated based on the sample space.

RANDOM EXPERIMENT

A random experiment is a mechanism that produces a definite outcome that cannot be predicted with certainty.

Example:

Consider the simple random experiment of rolling a die. Before rolling a die you do not know the result.



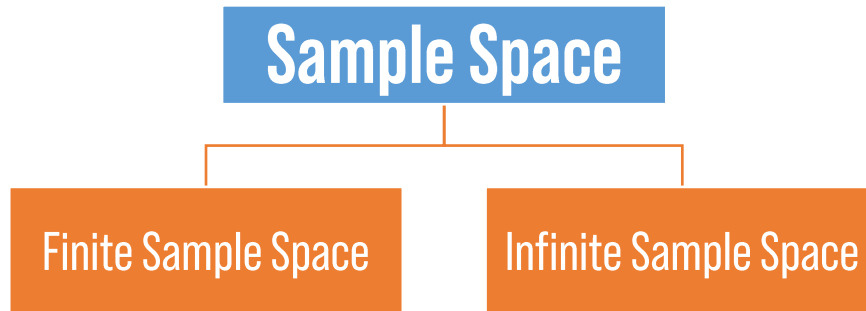
1. Observe a **1**
2. Observe a **2**
3. Observe a **3**
4. Observe a **4**
5. Observe a **5**
6. Observe a **6**

SAMPLE SPACE

The sample space associated with a random experiment. The set of all possible outcomes of a statistical experiment is called the sample space and is represented by the symbol S . The sample space S of possible outcomes when a coin is tossed may be written,

$$S = \{H, T\}$$

Where H and T correspond to “heads” and “tails” respectively. Each outcome in a sample space is called an element or a member of the sample space or simply a sample point.



Finite Sample Space

- A finite sample space contains a finite number of outcomes.
- Sample space Ω is the set of all possible sample points $\omega \in \Omega$
- Tossing a coin: $\Omega = \{H, T\}$
- Casting a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- A pair of six-sided dice is tossed twice.

Infinite Sample Space

- Number of customers in a queue: $\Omega = \{0, 1, 2, \dots\}$
- Call holding time (e.g. in minutes): $\Omega = \{x \in \mathbb{R} \mid x > 0\}$

Event and event Space

- An event is a subset of the sample space (S).
- The class of all events associated with a given experiment is define to be the event space (ε).

$$S = \{Head, Tail\}$$
$$\varepsilon = \{H, T, \{H, T\}, \phi\}$$

This method is valid only if

- the sample space is finite, and
- all the outcomes in the sample space are equally likely.

If the above two conditions are satisfied, the probability of an event E is calculated as

$$P(E) = \frac{\text{number of outcomes in } E}{\text{total number of outcomes in the sample space}} = \frac{n(E)}{n(S)}$$

Equation 21: Probability

The probability of an event E is usually denoted by $\Pr(E)$ or $P(E)$.

Example:

We define Probability of an event A to be to be a successful outcome

$$P(A) = \frac{\text{Number of successful outcomes (r)}}{\text{Total number of outcomes (n)}} = \frac{r}{n}$$

Note:

The elementary probability is the measure of relative frequency of certain repeatable events. The probability P(E) of some event E is a number that lies between zero and one.

$$0 < P(E) < 1$$

- If $P(E) = 0$, then the event E is impossible to occur
- If $P(E) = 1$, then the event E will surely occur.

Example (01)

Let event A be the number 3 or 4 turning up in a single throw of a die. Find P(A).

Sample Space: {1,2,3,4,5,6}

A = {3,4}

So, $P(A) = \frac{1}{3}$

Probabilities are representing in many ways.

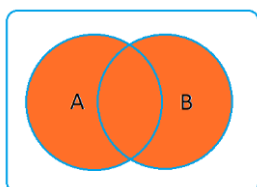
- Venn diagram
- Tree diagram
- Probability space diagram

Venn diagram

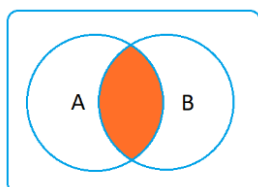
The growth of the Venn diagram dates back to 1880 when John Venn brought them to life in an article titled ‘On the Diagrammatic and Mechanical Representation of Propositions and Reasoning.’ It was in the Philosophical Magazine and Journal of Science. He is the one who originally generalized them, no wonder their naming, i.e., Venn Diagrams in 1918.

A Venn diagram (also called a set diagram or logic diagram) is a diagram that shows all possible logical relations between a finite collection of different sets. A Venn diagram consists of multiple overlapping closed curves, usually circles, each representing a set.

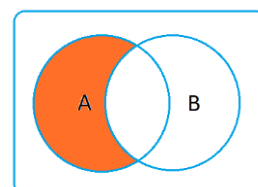
Set operations by Venn diagram



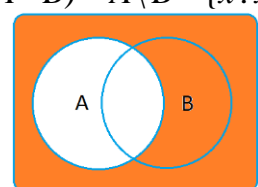
$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$



$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$



$$A \cap B' = (A - B) = A \setminus B = \{x : x \in A, x \notin B\}$$



$$A^c = \{x : x \in U, x \notin A\}$$

Low of Algebra of Sets

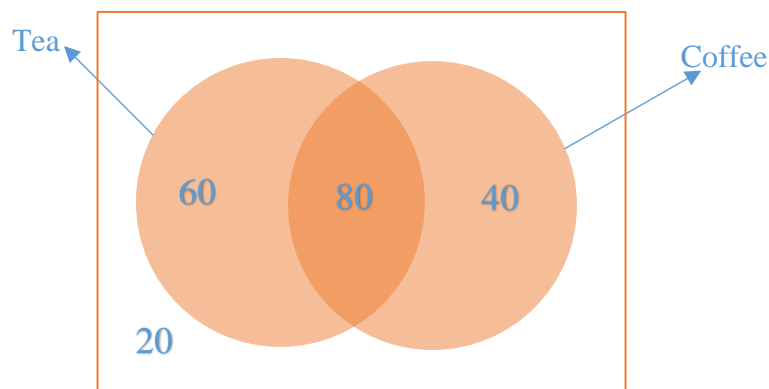
Idempotent Laws	$A \cup A = A \quad A \cap A = A$
Associative Laws	$(A \cup B) \cup C = A \cup (B \cup C)$ $(A \cap B) \cap C = A \cap (B \cap C)$
Commutative Laws	$A \cup B = B \cup A$ $A \cap B = B \cap A$
Distributive Laws	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
Complement Laws	$A \cup A^c = U$ $(A^c)^c = A$
De Morgan's Laws	$(A \cup B)^c = A^c \cap B^c$ $(A \cap B)^c = A^c \cup B^c$

Equation 22: Low of Algebra of Sets

EXERCISES

In a college, 200 students are randomly selected. 140 like tea, 120 like coffee and 80 like both tea and coffee.

1. How many students like only tea?
➤ 60
2. How many students like only coffee?
➤ 40
3. How many students like neither tea nor coffee?
➤ 20
4. How many students like only one of tea or coffee?
➤ 100
5. How many students like at least one of the beverages?
➤ 180



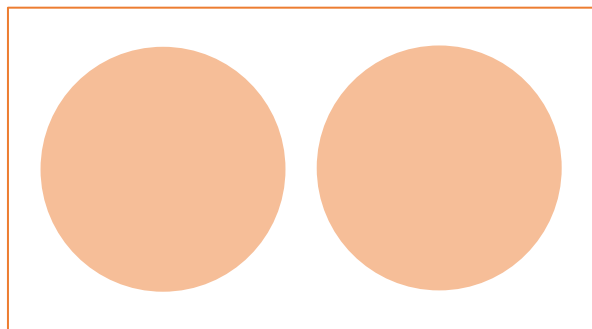
Additional Rule for Probability

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

Independent Probabilities

$$P(A \cap B) = P(A) \cdot P(B)$$

Mutually Exclusive Events



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

When two events cannot occur at the same time, then we say that the events are mutually exclusive.

For example: when we flip a coin then either heads can come or tails. Both heads and tails cannot be outcomes simultaneously.

Probability Tree Diagrams

A tree diagram is simply a way of representing a sequence of events.

Tree diagrams are particularly useful in probability since they record all possible outcomes in a clear and uncomplicated manner.

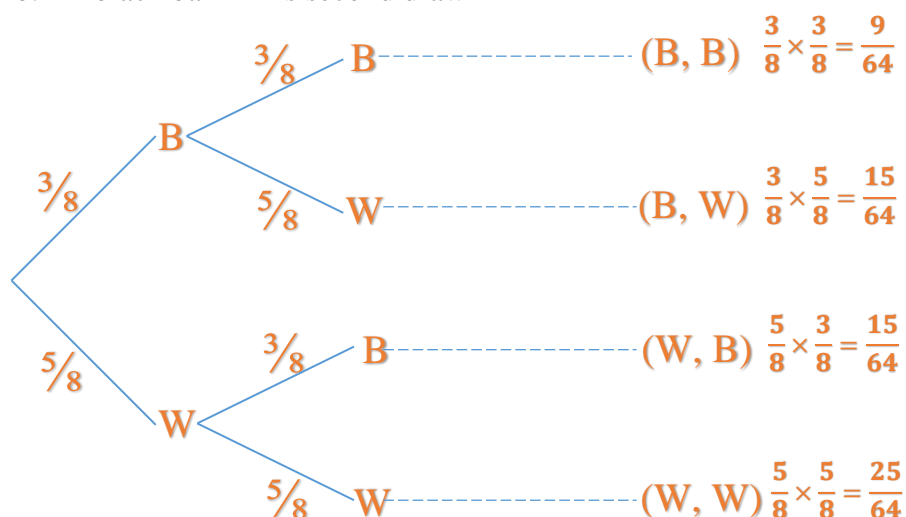
Example:

A bag contains 3 black balls and 5 white balls. Paul picks a ball at random from the bag and replaces it back in the bag. He mixes the balls in the bag and then picks another ball at random from the bag.

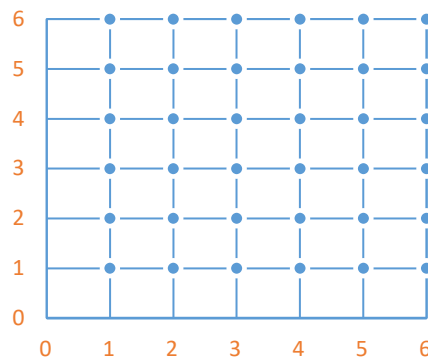
(1) Construct a probability tree of the problem.

(2) Calculate the probability that Paul picks:

- Two black balls
- A black ball in his second draw



Probability Space diagram



Example:

Two fair dice are thrown together. Find the probability that the sum of the resulting number is

- (1) odd
- (2) a prime number
- (3) even

Answer:

Construct the following probability diagram showing the sums:

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

- (1) $\frac{1}{2}$
- (2) $\frac{5}{12}$
- (3) $\frac{1}{2}$

Conditional Probability

The probability of the occurrence of an event B, given that even A has already occurred is called the conditional probability of B given A and denoted by $P(B|A)$.

when $P(A) \neq 0$

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

Equation 23: Conditional Probability

The conditional probability of A given by, B and denoted by $P(A|B)$.

when $P(B) \neq 0$

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

$$\Leftrightarrow P(A \cap B) = P(A|B)P(B)$$

If the events A and B are independent,

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

Equation 24: Independent events

Multiplication rule

Let A and B be two events. Then,

$$P(A \cap B) = P(B|A)P(A)$$

Equation 25: Multiplication Rule-01

If the two events are independent then

$$P(A \cap B) = P(A)P(B)$$

Equation 26: Multiplication Rule -02

Example:

Susan took two tests.

The probability of her passing both tests is $(A \cap B)$ 0.6.

The probability of her passing the first test is (B) 0.8.

What is the probability of her passing the second test given that she has passed the first test?

$$P(A \cap B) = 0.6, P(A) = 0.8, P(B|A) = ?$$

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{0.6}{0.8} = 0.75 \end{aligned}$$

Example:

1. What is the probability that the total of two dice will be greater than 9, given that the first die is a 5?

Let A = first die is 5

Let B = total of two dice is greater than 9

Given first die is five = $P(A) = \frac{6}{36} = \frac{1}{6}$

Possible outcomes for A and B: (5, 5), (5, 6)

$$P(A \cap B) = \frac{2}{36} = \frac{1}{18}$$

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

2. D and E are two events such that,

$$P(D) = 0.3, P(E) = 0.5, \text{ and } P(D|E) = 0.25$$

find,

- i. $P(D \cup E)$
- ii. $P(D \cap E)$
- iii. $P(D|E')$
- iv. $P(D'|E')$

Answer...

$$\begin{aligned} P(D|E) &= \frac{P(D \cap E)}{P(E)} \\ P(D \cap E) &= P(D|E) \cdot P(E) \\ &= 0.25 \times 0.5 \\ P(D \cap E) &= 0.125 \end{aligned}$$

$$\begin{aligned} P(D \cup E) &= P(D) + P(E) - P(D \cap E) \\ P(D \cup E) &= 0.3 + 0.5 - 0.125 = 0.675 \end{aligned}$$

$$\begin{aligned} P(D|E') &= \frac{P(D \cap E')}{P(E')} \\ &= \frac{P(D) - P(D \cap E)}{1 - P(E)} = \frac{0.3 - 0.125}{1 - 0.5} = 0.35 \end{aligned}$$

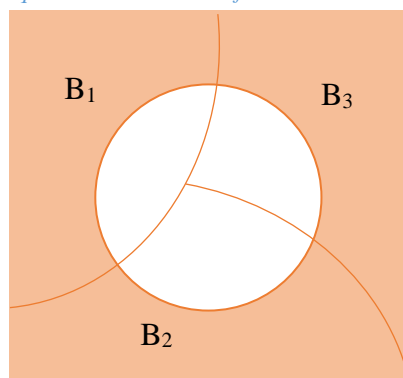
$$\begin{aligned} P(D'|E') &= \frac{P(D' \cap E')}{P(E')} \\ &= \frac{1 - P(D \cup E)}{1 - P(E)} = \frac{1 - 0.675}{1 - 0.5} = 0.65 \end{aligned}$$

The Law of Total Probability

For a given probability space (S), if B_1, B_2, \dots, B_n be a partition of S where, they are mutually exclusive and exhaustive events. $P(B_i) > 0$ for all $i = 1, 2, \dots, n$ then,

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

Equation 27: The Law of Total Probability



Bayes' Theorem

Let B_1, B_2, \dots, B_n be mutually exclusive and exhaustive events, and A be any event. Then for $i = 1, 2, \dots, n$,

$$\begin{aligned}P(B_i|A) &= \frac{P(B_i \cap A)}{P(A)} \\&= \frac{P(A|B_i)P(B_i)}{P(A)}\end{aligned}$$

Equation 28: Bayes' Theorem

Example:

In a factory, three machines, A, B and C make 30%, 45% and 25% of the products respectively. It is known from past experience that 2%, 3% and 2% of products made by each machine respectively are defective. Now, suppose that a finished product is randomly selected. What is the probability that it is defective?

B_1 = Product from Machine A

B_2 = Product from Machine B

B_3 = Product from Machine C

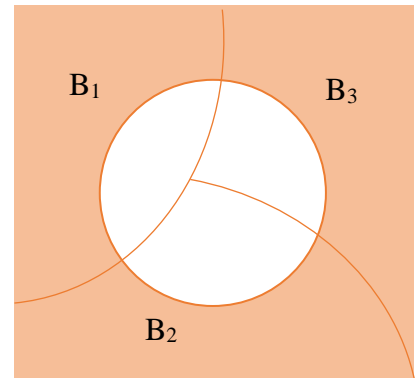
$P(D)$ = Probability of a product being Defective

$$P(D) = \sum_{i=1}^3 P(D | B_i)P(B_i)$$

$$P(D) = P(D | B_1)P(B_1) + P(D | B_2)P(B_2) + P(D | B_3)P(B_3)$$

$$P(D) = 0.02 \times 0.3 + 0.03 \times 0.45 + 0.02 \times 0.25$$

$$P(D) = 0.0245$$



Suppose that one item is selected at random and it is found to be defective, find the probability that the item was produced by machine.

$$P(B_1 | D) = \frac{P(D | B_1) \cdot P(B_1)}{P(D)} = \frac{0.02 \times 0.3}{0.0245} = 0.244$$

$$P(B_2 | D) = \frac{P(D | B_2) \cdot P(B_2)}{P(D)} = \frac{0.03 \times 0.45}{0.0245} = 0.551$$

$$P(B_3 | D) = \frac{P(D | B_3) \cdot P(B_3)}{P(D)} = \frac{0.02 \times 0.25}{0.0245} = 0.204$$

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

RANDOM VARIABLE

A Random Variable is a function, which assigns unique numerical values to all possible outcomes of a random experiment under fixed conditions. A random variable, usually written X , is a variable whose possible values are numerical outcomes of a random phenomenon.

A Random Variable is;

- Discrete if it has either finite or countably infinite values.
- Continuous if it takes values in a continuum.

For Example....

Consider two coins tossed simultaneously...

Sample Space: {HH, HT, TH, TT}

We are taking all the possibility to head may be obtained $(x) = \{0, 1, 2\}$

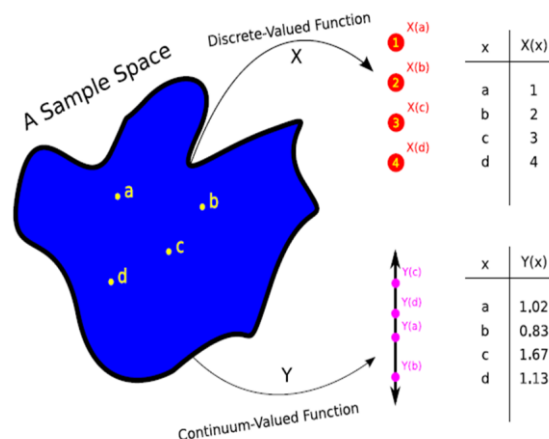
And we take the probability to each of event $[P(X=x)] = \{\frac{1}{4}, \frac{2}{4}, \frac{1}{4}\}$

So, in here "x" is called a Random Variable...

X	P(X=x)
0	$\frac{1}{4}$
1	$\frac{2}{4}$
2	$\frac{1}{4}$



DIFFERENT BETWEEN DISCRETE VS CONTINUOUS VARIABLE



Discrete Random Variable

A random variable is called a discrete random variable if its set of possible outcomes is countable.

Example:

Flip a coin and count the number of heads.

- Number of heads is represented by an integer value - a number between 0 and plus infinity.
- Therefore, the number of heads is a discrete random variable.

Number of calls per a minute in a phone exchange.

$$X = \{0, 1, 2, 3, 4, 5, \dots\}$$

Continuous Random Variable

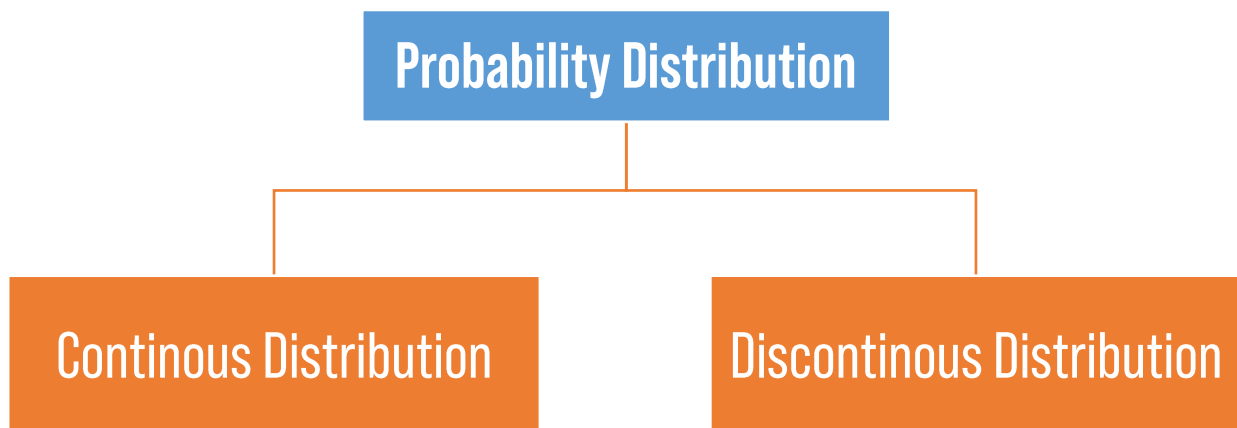
Continuous random variables, in contrast, can take on any value within a range of values.

Example:

- height of students in class
- weight of students in class
- time it takes to get to school
- distance traveled between classes

PROBABILITY DISTRIBUTION

A probability distribution is a table or an equation that links each possible value that a random variable can assume with its probability of occurrence.



Properties of probability density function

- $0 \leq P(X = x_i) = p_i \leq 1$
- $\sum_{i=1}^n P(X = x_i) = \sum_{i=1}^n p_i = 1$

Equation 29: Properties of probability density function

Example:

Using the above equation,

$$P(X = x) = \begin{cases} k & \text{if } x = 1, 2, 3, 4 \\ 0 & \text{others} \end{cases}$$

Find the value of k .

Explanation....

$P(X=x)$ is a probability mass function (PMF), which defines how probabilities are assigned to discrete values of X .

- **For $x=1,2,3,4$ the probability is k .**
That means X can take one of these four values, and each has the same probability k .
- **For any other value of x , the probability is 0.**
This means that X **cannot take** any other values outside $\{1,2,3,4\}$.

This type of probability distribution is common in discrete random variables.

k is a **constant probability value** that needs to be determined. It represents the probability of each of the possible values of X (i.e., 1, 2, 3, or 4). Since probability distributions must sum to 1, we can find k using this property.

Answer:

From Probability theory, we know that the sum of all probabilities for a discrete variable must be 1:

$$\sum_{i=1}^n P(X = x) = 1$$

Since $P(X = x) = k$ for $x = 1, 2, 3, 4$ we write:

$$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1$$

Since each of these probability is k :

$$k + k + k + k = 1$$

$$4k = 1$$

$$k = \frac{1}{4}$$

This means that each of $X = 1, 2, 3, 4$ has a probability of $\frac{1}{4}$, and all other values have probability 0.

Continuous Probability Distribution

If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution. A continuous probability distribution differs from a discrete probability distribution in several ways. The continuous probability distribution cannot be expressed in tabular form. An equation or formula is used to describe a continuous probability distribution.

In the following equation probability that X assumes a value between a and b is equal to the shaded area under the density function between the ordinates at $x=a$ and $x=b$. Probability Density Function is given by

$$f(x) = P(a \leq x \leq b) = \int_a^b f(x) dx \geq 0$$

Equation 30: Continuous Probability Distribution

The function $f(x)$ is a probability density function for the continuous random variable X , defined over the set of real numbers \mathbb{R} . The pdf $f(x)$ has two important properties

$$f(x) \geq 0, \text{ for all } x \in \mathbb{R}$$

$$\bullet \int_{-\infty}^{\infty} f(x) dx = 1$$

$$\bullet P(a < X < b) = \int_a^b f(x) dx$$

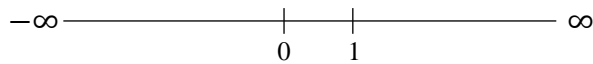
Example:

Consider the function...

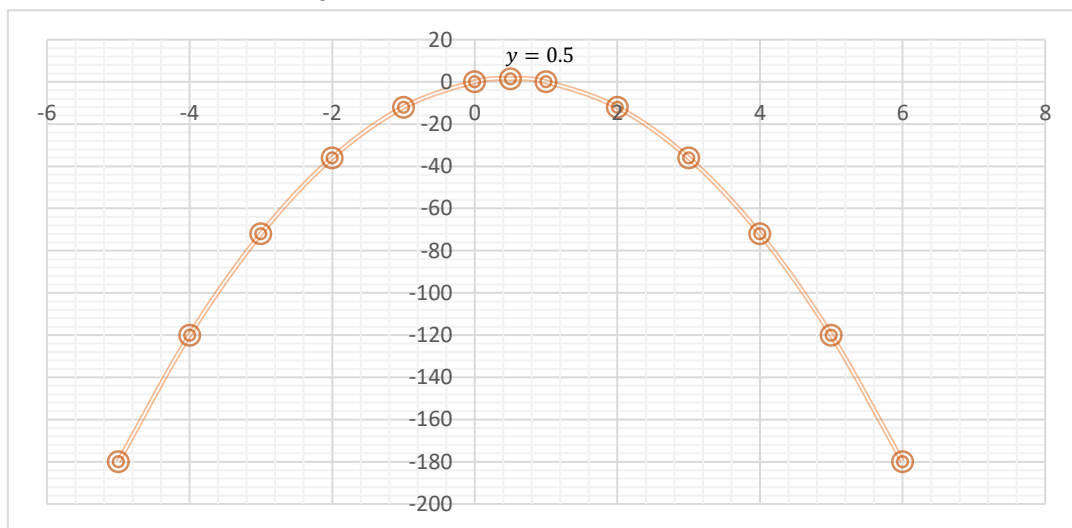
$$f(x) = \begin{cases} 6x(1-x) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

1. Check that $f(x)$ has the two required properties for a pdf, and sketch its graph.
2. Suppose that the continuous random variable X has the pdf $f(x)$. obtain the following probabilities without calculation:

- i) $P(x \leq -3)$
- ii) $P(0 \leq X \leq 1)$
- iii) $P(0.5 \leq X \leq 1)$



$$\begin{aligned} 1) & \int_{-\infty}^0 6x(1-x) dx + \int_0^1 6x(1-x) dx + \int_1^{\infty} 6x(1-x) dx \\ & 0 + 6 \left(\left[\frac{x^2}{2} \right]_0^1 - \left[\frac{x^3}{3} \right]_0^1 \right) + 0 \\ & 6 \times \frac{1}{6} = 1 \end{aligned}$$



- 2)
 - i) $\int_{-\infty}^{-3} 6x(1-x) dx = 0$
 - ii) $\int_0^1 6x(1-x) dx = 1$
 - iii) $\int_{0.5}^1 6x(1-x) dx = 0.5$

MATHEMATICAL EXPECTATION

MEAN OF A RANDOM VARIABLE

Let X be a random variable with probability distribution $f(x)$. The mean or expected value of X

$$\mu = E(X) = \sum_x xP(x) \text{ if } X \text{ is discrete}$$

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx \text{ if } X \text{ is continuous.}$$

Equation 31: Mathematical Expectation (Mean Value of the probability Distribution)

Example:

Assuming that the two fair coins were tossed, we find that the sample space for our experiment is

$$S = \{HH, HT, TH, TT\}.$$

Let X represent the number of heads in the sample. So, we can write $x=0,1,2$

Then,

X	0	1	2
$P(X=x)$	1/4	1/2	1/4

$$\mu = E(X) = \sum_x xP(x)$$

$$E(X) = (0)\left(\frac{1}{4}\right) + (1)\left(\frac{1}{2}\right) + (2)\left(\frac{1}{4}\right) = 1$$

This result means that a person who tosses 2 coins, on the average get 1 head.

Theorem:

Let X be a random variable with probability distribution $f(x)$. The expected value of the random variable $g(X)$ is

$$\mu_{g(X)} = E[g(X)] = \sum_x g(x)f(x) \text{ if } X \text{ is discrete.}$$

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx \text{ if } X \text{ is continuous.}$$

Equation 32: If Random variable becoming a function

Theorem:

Let X be a continuous random variable with mean μ_X . Then,

$$E(aX + b) = aE(X) + b = a\mu_X + b$$

Equation 33: Relationship between $E[g(x)]$ and $E(x)$

for any real numbers a, b .

Proof:

For a continuous random variable x , mean of a function of x say $g(X)$, given by,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

So, for $g(X) = aX + b$, we find that,

$$E(aX + b) = \int_{-\infty}^{\infty} (aX + b) f(x) dx$$

$$E(aX + b) = \int_{-\infty}^{\infty} aX \cdot f(x) dx + \int_{-\infty}^{\infty} bf(x) dx$$

$$E(aX + b) = a \int_{-\infty}^{\infty} X \cdot f(x) dx + b \int_{-\infty}^{\infty} f(x) dx$$

$$E(aX + b) = a.E(X) + b \text{ ----- (1)}$$

From (1)...

$$E(X + b) = E(X) + b$$

$$E(b) = b$$

$$E(aX) = a.E(X)$$

Example:

Let X be a random variable with density function

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the expected value of $g(X) = 4X + 3$.

Answer:

We can solve this within two methods...

Method (1):

$$E[g(x)] = \int_{-1}^2 g(x) \cdot f(x) dx$$

$$E[g(x)] = \int_{-1}^2 (4x + 3) \cdot \frac{x^2}{3} dx$$

$$E[g(x)] = \frac{4}{3} \int_{-1}^2 x^3 dx + \int_{-1}^2 x^2 dx$$

$$E[g(x)] = \frac{4}{3} \times \left[\frac{x^4}{4} \right]_{-1}^2 + \left[\frac{x^3}{3} \right]_{-1}^2$$

$$E[g(x)] = 5 + 3 = 8$$

Method (2):

Finding $E(x)$...

$$E(x) = \int_{-1}^2 x \cdot f(x) dx$$

$$E(x) = \int_{-1}^2 x \cdot \frac{x^2}{3} dx$$

$$E(x) = \int_{-1}^2 \frac{x^3}{3} dx$$

$$E(x) = \left[\frac{x^4}{12} \right]_{-1}^2$$

$$E(x) = \frac{5}{4} \text{ ----- (1)}$$

Finding $E[g(x)]$...

$$E[g(x)] = E[4x + 3]$$

$$= 4E(x) + 3 \text{(From (1))}$$

$$= 4 \times \frac{5}{4} + 3$$

$$E[g(x)] = 8$$

VARIANCE AND COVARIANCE OF RANDOM VARIABLES

Let X be a random variable with probability distribution $f(x)$ and mean μ . The variance of X is,

$$\sigma^2 = E[(X - \mu)^2] = \sum_{-\infty}^{\infty} (x - \mu)^2 f(x), \text{ if } X \text{ is discrete,}$$

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x), \text{ if } X \text{ is continuous.}$$

Equation 34: Variance in Random Variables

The positive square root of the variance, σ , is called the standard deviation of X .

Theorem:

The variance of a random variable X is

$$\sigma^2 = E(X^2) - (E(X))^2 = E(X^2) - \mu^2 \quad (E(x) = \mu)$$

Equation 35: Variance Equation (Simplified)

Theorem:

Let X be a random variable with variance σ^2 . Then,

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

Equation 36: Relationship between $\text{Var}(aX+b)$ and $\text{Var}(x)$

for any real numbers a and b .

Proof:

We defined a new random variable:

$$Y = aX + b \quad \text{--- (1)}$$

Where:

- X is a random variable with mean μ_x and variance σ_x^2 .
- a and b are constants.
- We want to find the variance of Y , denoted as $\text{Var}(Y)$.

Variance is defined as:

$$\text{Var}(Y) = E[(Y - \mu_Y)^2] \quad \text{--- (2)}$$

$$\mu_Y = E[Y] = E[aX + b]$$

$$= aE[X] + b$$

$$\mu_Y = a\mu_x + b \quad \text{--- (3)}$$

We can rewrite (2) from (1), (3) as,

$$\text{Var}(aX + b) = E[(aX + b - (a\mu_x + b))^2]$$

$$\text{Var}(aX + b) = E[(aX - a\mu_x)^2]$$

$$\text{Var}(aX + b) = a^2 E(X - \mu_x)^2$$

By definition:

$$E[(X - \mu_x)^2] = \text{Var}(X) = \sigma_x^2$$

So:

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\text{Var}(aX + b) = a^2 \sigma_x^2$$

SUMMARY:

$$E(aX + b) = a.E(X) + b$$

$$E(X + b) = E(X) + b$$

$$E(b) = b$$

$$E(aX) = a.E(X)$$

$$E(aX^2) = a.E(X^2)$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\text{Var}(X + b) = \text{Var}(X)$$

$$\text{Var}(b) = 0$$

$$\text{Var}(aX) = a^2 \text{Var}(X)$$

Equation 37: Summarization of Expectation and Variance

Example:

Determine the mean and variance of the random variable X having the following probability distribution.

X=x	1	2	3	4	5	6	7	8	9	10
P(X)	0.15	0.10	0.10	0.01	0.08	0.01	0.05	0.02	0.28	0.20

$$E(X) = \sum x.P(X)$$

$$E(X) = 1 \times 0.15 + 2 \times 0.1 + 3 \times 0.1 + 4 \times 0.01 + 5 \times 0.08 + 6 \times 0.01 + 7 \times 0.05 + 8 \times 0.02 + 9 \times 0.28 + 10 \times 0.2$$

$$E(X) = 6.18$$

$$\text{Var}(X) = E(X^2) - E(X)^2$$

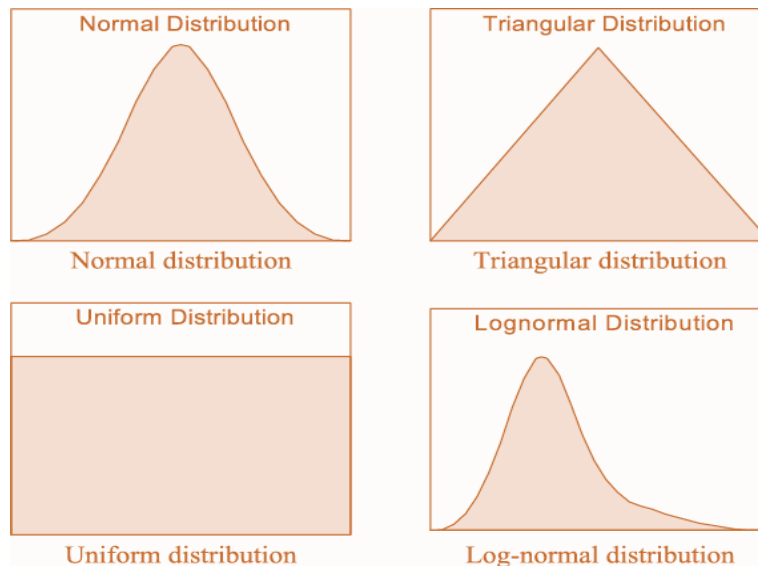
$$E(X^2) = \sum x^2 P(X)$$

$$E(X^2) = 50.38$$

$$\text{Var}(X) = 50.38 - 38.1924$$

$$\text{Var}(X) = 12.1876$$

PROBABILITY DISTRIBUTIONS



STATISTICAL EXPERIMENT

All statistical experiments have three things in common:

- The experiment can have more than one possible outcome.
- Each possible outcome can be specified in advance.
- The outcome of the experiment depends on chance.

Example:

- A coin toss has all the attributes of a statistical experiment.
- There is more than one possible outcome.
- We can specify each possible outcome in advance - heads or tails. And there is an element of chance.
- We cannot know the outcome until we actually flip the coin.

PROBABILITY DISTRIBUTION

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

Example:

- Consider a simple experiment in which we flip a coin two times.
- Suppose the random variable X is defined as the number of heads that result from two-coin flips.
- Then, the above table represents the probability distribution of the random variable X.

Number Of Heads	0	1	2
Probability	1/4	1/2	1/4

There are two probability distributions:

- Discrete probability Distribution
- Continuous probability Distribution

If a variable can take on any value between two specified values, it is called a continuous variable; otherwise, it is called a discrete variable.

Some examples will clarify the difference between discrete and continuous variables.

1. Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity.
 - However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.
2. Suppose the fire department that all fire fighters must weigh between 75 and 85 kg.
 - The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 75 and 80 kg.

Discrete probability Distribution

1. Discrete Uniform Distribution
2. Bernoulli Probability Distribution
3. Binomial Probability Distribution
4. Poisson Probability Distribution

Discrete Uniform Distribution

The simplest of all discrete probability distributions is one where the random variable assumes each of its values with an equal probability.

Such a probability distribution is called a discrete uniform distribution.

If the random variable X assumes the values $x_1, x_2, x_3 \dots x_n$, with equal probabilities, then the discrete uniform distribution is given by...

$$P(X = x) = \frac{1}{k}, \quad x = x_1, x_2, x_3 \dots x_n$$

Equation 38: Discrete Uniform distribution for Equal probabilities variables

$$\mu = \frac{1}{k} \sum_{i=1}^k x_i \quad \text{and} \quad \sigma^2 = \frac{1}{k} \sum_{i=1}^k (x_i - \mu)^2$$

Equation 39: Mean and Variance for Discrete Uniform Distribution

Binomial Distribution

A binomial experiment (also known as a Bernoulli trial) is a statistical experiment that has the following properties:

- The experiment consists of n repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
- The probability of success, denoted by P , is the same on every trial.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment:

You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

- The experiment consists of repeated trials. We flip a coin 2 times.
- Each trial can result in just two possible outcomes - heads or tails.
- The probability of success is constant 0.5 on every trial.
- The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Notation:

The following notation is helpful, when we talk about binomial probability.

- x: The number of successes that result from the binomial experiment.
- n: The number of trials in the binomial experiment.
- p: The probability of success on an individual trial.
- q: The probability of failure on an individual trial, (This is equal to 1 - P.)

Probability density function

$$X \sim \text{bin}(n, p)$$

$$P(X = k) = {}^nC_k p^k (1 - p)^{n-k}$$

Equation 40: Probability density function

- n: The total number of trials in the binomial experiment.
- k: number of successes trials $k=0, 1, 2, \dots, n$
- p: The probability of success on an individual trial.

Example:

A fair coin is tossed 6 times. Find the probability of getting

(a) exactly 4 heads

$$X \sim \text{bin}(6, 0.5)$$

$$P(X = k) = {}^nC_k p^k (1 - p)^{n-k}$$

$$P(X = 4) = {}^6C_4 p^4 (1 - p)^2$$

$$P(X = 4) = \frac{6!}{2! \times 4!} \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^2$$

$$P(X = 4) = 0.234375$$

(b) more than 3 heads

$$\begin{aligned}
 \sum_{i=4}^6 P(X=i) &= P(X=4) + P(X=5) + P(X=6) \\
 &= {}^6C_4 p^4 (1-p)^2 + {}^6C_5 p^5 (1-p)^1 + {}^6C_6 p^6 (1-p)^0 \\
 &= \left(\frac{1}{2}\right)^6 ({}^6C_4 + {}^6C_5 + {}^6C_6) \\
 &= 0.34375
 \end{aligned}$$

(c) more than or equal 4 heads

$$\begin{aligned}
 \sum_{i=4}^6 P(X=i) &= P(X=4) + P(X=5) + P(X=6) \\
 &= {}^6C_4 p^4 (1-p)^2 + {}^6C_5 p^5 (1-p)^1 + {}^6C_6 p^6 (1-p)^0 \\
 &= \left(\frac{1}{2}\right)^6 ({}^6C_4 + {}^6C_5 + {}^6C_6) \\
 &= 0.34375
 \end{aligned}$$

(d) more than or equal 1 head

$$\begin{aligned}
 \sum_{i=1}^6 P(X=i) &= P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6) \\
 &= {}^6C_1 p^1 (1-p)^5 + {}^6C_2 p^2 (1-p)^4 + {}^6C_3 p^3 (1-p)^3 + {}^6C_4 p^4 (1-p)^2 + {}^6C_5 p^5 (1-p)^1 + {}^6C_6 p^6 (1-p)^0 \\
 &= \left(\frac{1}{2}\right)^6 ({}^6C_1 + {}^6C_2 + {}^6C_3 + {}^6C_4 + {}^6C_5 + {}^6C_6) \\
 &= 0.984375
 \end{aligned}$$

The binomial distribution has the following properties:

$$\mu_X = nP \qquad \sigma_X^2 = nP(1-p)$$

Equation 41: Mean & Variance for the distribution

Example:

A student takes an exam of 18 multiple choice questions with 4 choices per question. Find the expected number of correct answers and its standard deviation.

$$\begin{aligned}
 X &\sim \text{bin}\left(18, \frac{1}{4}\right) & \sigma_X^2 &= nP(1-p) \\
 \mu_X &= nP & \sigma_X^2 &= 18 \times \frac{1}{4} \times \frac{3}{4} \\
 \mu_X &= 18 \times \frac{1}{4} = 4.5 & \sigma_X^2 &= 1.8
 \end{aligned}$$

Poisson Distribution

A Poisson experiment is a statistical experiment that has the following properties:

- A discrete random variable X is said to follow a Poisson distribution if it assumes only non-negative integer values.
- The random variable x denotes the number of occurrences over a given span.
- There is only one parameter λ , which is the average rate of occurrence.
- The occurrence of the event is not dependent on another occurrence of that event.

Poisson vs. Other Distributions:

- It is a discrete probability distribution (only counts whole numbers).
- It is different from a binomial distribution, which counts the number of successes in a fixed number of trials.

Example Situations:

- The number of calls received at a call center per hour.
- The number of typos found on a page of a book.
- The number of bus arrivals at a stop in 10 minutes.
- The number of emails received per day.

The probability distribution of the Poisson random variable X , representing the number of outcomes occurring in a given time interval or specified region denoted by t is...

$$P(X; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}, \quad x = 0, 1, 2, \dots$$

when $t = 1$; In here “ t ” means if we calculated mean(λ) for 5 minutes but we doing experiment for 10 minutes than t could be $t = 2$

$$X \sim \text{Poiss}(\lambda)$$

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

Equation 42: Probability Distribution for Poisson distribution

Theorem:

Both the mean and variance of the Poisson distribution are λt .

$$X \sim \text{Poiss}(\lambda t)$$

$$\text{Mean} = E(X) = \lambda t$$

$$\text{Variance} = \text{Var}(X) = \lambda t$$

Equation 43: Means and Variance for Poisson Distribution

Example:

The number of telephone calls made to a switch board during an afternoon can be distributed by Poisson distribution with a mean of eight calls per five minutes period.

Find the probability that in the next five minutes,

a. No calls

$$P(X; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$$
$$P(X = 0) = \frac{e^{-8} \times 8^0}{0!} = e^{-8}$$

b. Five

$$P(X; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$$
$$P(X = 5) = \frac{e^{-8} \times 8^5}{5!} = \frac{4096}{15e^8}$$

c. at least three

$$P(X; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$$
$$P(X \geq 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$
$$P(X \geq 3) = 1 - \left[\frac{e^{-8} \times 8^0}{0!} + \frac{e^{-8} \times 8^1}{1!} + \frac{e^{-8} \times 8^2}{2!} \right] = 0.986246$$

d. at most four calls is made.

$$P(X; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$$
$$P(X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$
$$P(X \leq 4) = \frac{e^{-8} \times 8^0}{0!} + \frac{e^{-8} \times 8^1}{1!} + \frac{e^{-8} \times 8^2}{2!} + \frac{e^{-8} \times 8^3}{3!} + \frac{e^{-8} \times 8^4}{4!} = 0.0996324$$

The Poisson Distribution as an Approximation for the Binomial Distribution

Theorem:

Let be a binomial random variable with probability distribution $b(x; n, p)$ When $n \rightarrow \infty, p \rightarrow 0$ and $np \rightarrow \mu$ remains constant,

$$X \sim \text{bin}(n, p)$$
$$\text{when } n \uparrow\uparrow \text{ and } p \downarrow\downarrow$$
$$\lambda = np$$
$$X \sim \text{Poiss}(\lambda)$$

Equation 44: Binomial to Poisson Distributions

Example:

On a particular production line, the probability that an item is defective is 0.01. Using suitable approximation, find the probability that, in a batch of 200 items,

- There are no defective items
- There are more than five defective items.
- There are exactly 175 defective items.

Continuous Probability Distributions

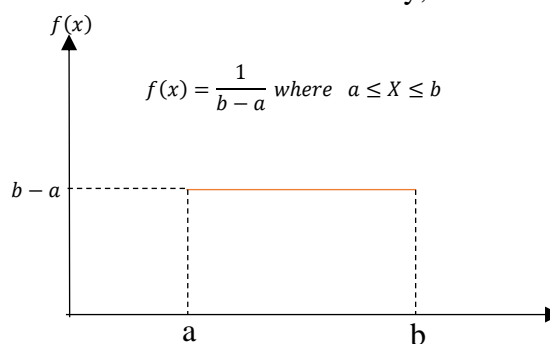
A continuous probability distribution differs from a discrete probability distribution in several ways. As a result, a continuous probability distribution cannot be expressed in tabular form. Instead, an equation or formula is used to describe a continuous probability distribution. It can be divided into two parts.

- Uniform Distribution
- Normal Distribution

Uniform Distribution

- One of the simplest continuous distributions in all of statistics is the continuous uniform distribution.
- This family of distributions is used to describe situations where the possible outcomes are all equally likely to occur.
- If X is a continuous random variable with a uniform distribution defined by,

$$f(x; A, B) = \begin{cases} \frac{1}{B - A} & A \leq X \leq B \\ 0 & \text{otherwise} \end{cases}$$



theorem

The mean and variance of the uniform distribution are,

$$X \sim uni(A, B)$$

$$Mean(\mu_x) = E(X) = \frac{1}{2}(B + A)$$

$$Variance(\sigma_x^2) = Var(X) = \frac{1}{12}(B - A)^2$$

Equation 45: mean & standard deviation for Uniform Distribution

Normal Distribution

The normal probability density function usually called the Normal distribution is one of the widely used probability models. Most phenomena such as

- Average marks of student
- Diameters of machine parts
- Life time of television bulb
- Weights of packages are normally distributed.

The normal distribution is also useful for approximating other distribution such as the binomial.

- The normal probability density function for a continuous random variable X is,

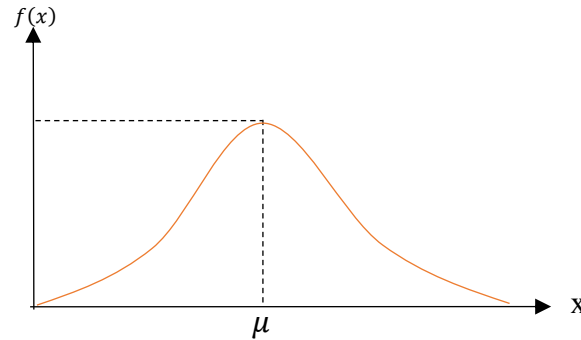
$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } -\infty < x < \infty$$

μ –mean of x ($E(X) = \mu$)

σ – Standard deviation of X ($\text{var}(X) = \sigma^2$)

Equation 46: Normal probability density function for a continuous random variable

- The curve defined by the above function is a bell – shaped symmetrical distribution.

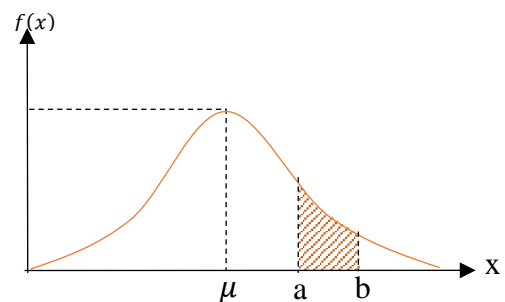


The normal distribution has the following properties:

- It is symmetric about the mean value.
- It is bell- shaped and has one mode.
- The total area under the normal distribution curve equal one.
- The two tails of the distribution approach the horizontal axis but not touch the axis.

The probability that a normally distributed random variable X assumes values between a and b is $P(a \leq X \leq B)$ which is equal to the proportion of total area under the curve between the limits a and b and this can be determined by integral calculus.

$$\begin{aligned} P(a \leq X \leq B) &= \int_a^b f(X) dx \\ &= \int_a^b \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \right] dx \end{aligned}$$



Standard Normal Distribution

- In practical situations it is not available to use the density function directly for finding area under the normal curve.
- Tables that allow us to compute areas under the normal probability density function are based on the standard normal distribution.
- The standard normal distribution has the same features as any normal distribution.
- The mean of the standard normal distribution is 0 and variance is one.

A random variable X any mean and standard deviation can be transformed to a standardized random variable Z by using the relation,

$$X \sim N(\mu, \sigma^2)$$

μ –mean of x

σ – Standard Deviation of X

$$X \sim N(\mu, \sigma^2)$$

$$\frac{X - \mu}{\sigma} \sim N(0,1)$$

$$Z = \frac{X - \mu}{\sigma} \Rightarrow Z \sim N(0,1)$$

Equation 47: Relationship between variable Z and variable X , μ and σ

- If X is normally distributed, Z is also normally distributed random variable with mean 0 and standard deviation 1.
- If Z is the standardized normal distributed random variable, the Z has the probability density function,

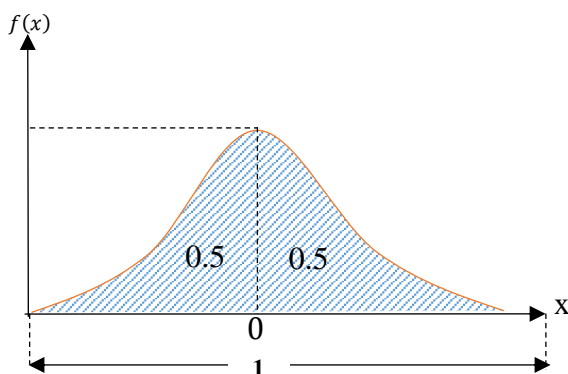
$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Z)^2} \text{ for } -\infty < Z < \infty$$

$$E(Z) = 0$$

$$\text{var}(Z) = 1$$

$$Z \sim N(0,1)$$

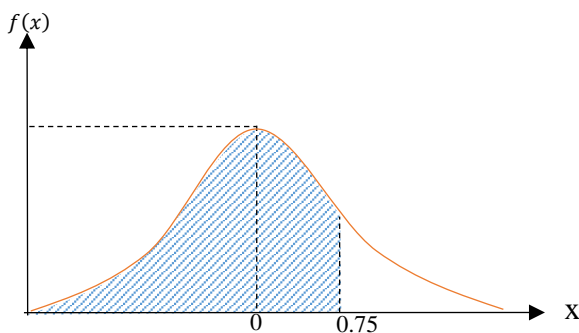
Therefore, the distribution is,



Example:

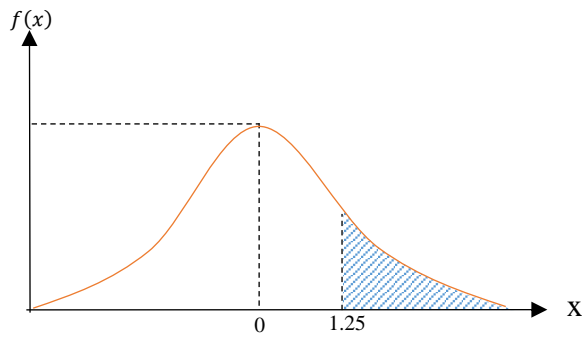
Using normal distribution table find;

$$P(Z < 0.75) = 0.7734$$



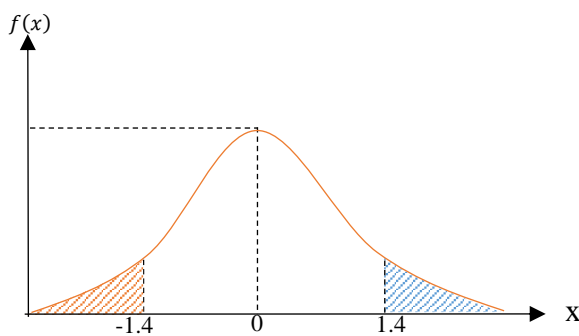
z	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734

$$P(z > 1.25) = 1 - P(z < 1.25) = 1 - 0.8944 = 0.1056$$



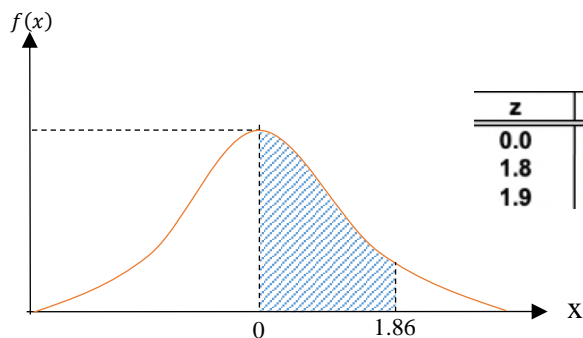
z	0.00	0.01	0.02	0.03	0.04	0.05
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944

$$\begin{aligned}
 P(z < -1.4) &= P(z > 1.4) \text{ (Because it's symmetric)} \\
 &= 1 - P(z < 1.4) \\
 &= 1 - 0.9192 = 0.0808
 \end{aligned}$$



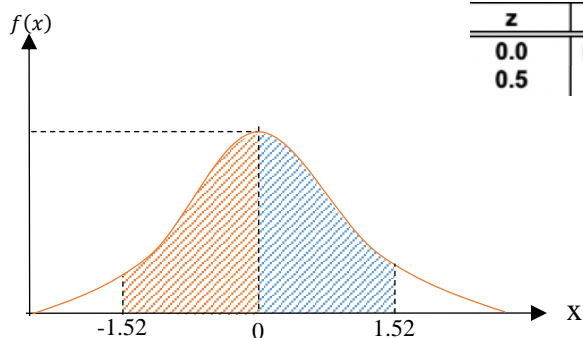
z	0.00	0.01	0.02	0.03	0.04	0.05
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265

$$P(0 \leq z \leq 1.86) = 0.9686 - 0.5000 = 0.4686$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750

$$P(-0.52 \leq Z \leq 0) = P(0 \leq Z \leq 0.52) = 0.6985 - 0.5000 = 0.1985$$



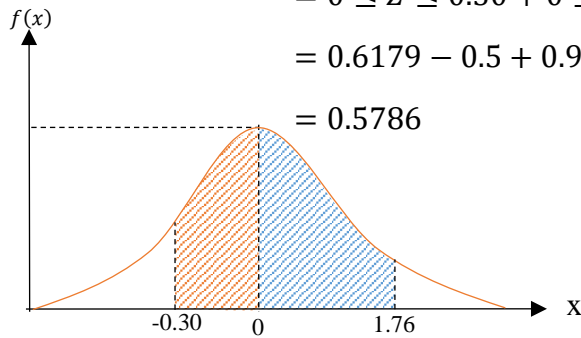
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123

$$P(-0.30 \leq Z \leq 1.76) = -0.30 \leq Z \leq 0 + 0 \leq Z \leq 1.76$$

$$= 0 \leq Z \leq 0.30 + 0 \leq Z \leq 1.76$$

$$= 0.6179 - 0.5 + 0.9607 - 0.5$$

$$= 0.5786$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608

The random variable

$$X \sim N(12, 2^2)$$

(If a question given like this you have to Calculate the Z value first)

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

Find,

$$P(X < 15)$$

$$\frac{X - \mu}{\sigma} < \frac{15 - 12}{2}$$

$$Z < 1.5$$

$$P(X < 15) = P(Z < 1.5)$$

$$P(X > 10)$$

$$\frac{X - \mu}{\sigma} > \frac{10 - 12}{2}$$

$$Z > -1 = Z < 1$$

$$P(X > 10) = P(Z < 1)$$

$$P(9 < X < 13)$$

$$\frac{9 - \mu}{\sigma} < Z < \frac{13 - \mu}{\sigma}$$

$$\frac{9 - 12}{2} < Z < \frac{13 - 12}{2}$$

$$-1.5 < Z < 0.5$$

$$P(9 < X < 13) = P(-1.5 < Z < 0.5)$$

Computing Probabilities for a Normal Distribution

Example 1:

lengths of a particular species of worm are normally distributed with mean 140cm and standard deviation 10cm. Find the probability that a worm is selected at random is,

1. less than 120cm long,
2. more than 148cm long,
3. between 148cm and 154cm long.

Answer:

$$X \sim N(140, 10^2)$$

1) $X < 120$

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

$$Z < \frac{120 - 140}{10}$$

$$Z < -2$$

$$P(Z < -2) = P(2 < Z)$$

$$= 1 - P(Z < 2)$$

$$= 1 - 0.9772 = 0.0228$$

2) $148 < X$

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

$$\frac{148 - 140}{10} < Z$$

$$0.8 < Z$$

$$P(0.8 < Z) = 1 - (Z < 0.8)$$

$$= 1 - 0.7881$$

$$= 0.2119$$

3) $148 < X < 154$

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

$$\frac{148 - 140}{10} < Z < \frac{154 - 140}{10}$$

$$0.8 < Z < 1.4$$

$$P(0.8 < Z < 1.4) = 0.9192 - 0.7881 = 0.1311$$

Example 3:

The weights of 1000 packages in a brand of serial are normally distributed with mean of 32 grams and standard deviation of 1.3 grams.

1. What is the probability that the weight of a package is between 32 grams and 34 grams.
2. Find the expected number of packages of weight between 32 grams and 34 grams.
3. Find the expected number of packages of weight more than 36 grams.

Answer:

$$X \sim N(32, 1.3^2)$$

$$1) 32 < X < 34$$

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

$$\frac{32 - 32}{1.3} < Z < \frac{34 - 32}{1.3}$$

$$0 < Z < 1.53846$$

$$\begin{aligned} P(32 < X < 34) &= P(0 < Z < 1.53846) \\ &= 0.9370 - 0.5000 \\ &= 0.4370 \end{aligned}$$

$$\begin{aligned} 2) \text{ Expected Number of Packages} &= 0.4370 \times 1000 \\ &= 437 \text{ Packages} \end{aligned}$$

$$3) 36 < X$$

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

$$\frac{36 - 32}{1.3} < Z$$

$$3.076923 < Z$$

$$\begin{aligned} P(3.076923 < Z) &= 1 - P(3.076923 < Z) \\ &= 1 - 0.9989 = 0.001 \end{aligned}$$

$$\begin{aligned} \text{Expected Number of Packages} &= 0.001 \times 1000 \\ &= 1 \text{ Package} \end{aligned}$$