Statistics: Science of collection, presentation, analysis, and reasonable interpretation of collected information. It presents a rigorous scientific method for gaining insight into data. It is useful to make inference and predict relations of variables.

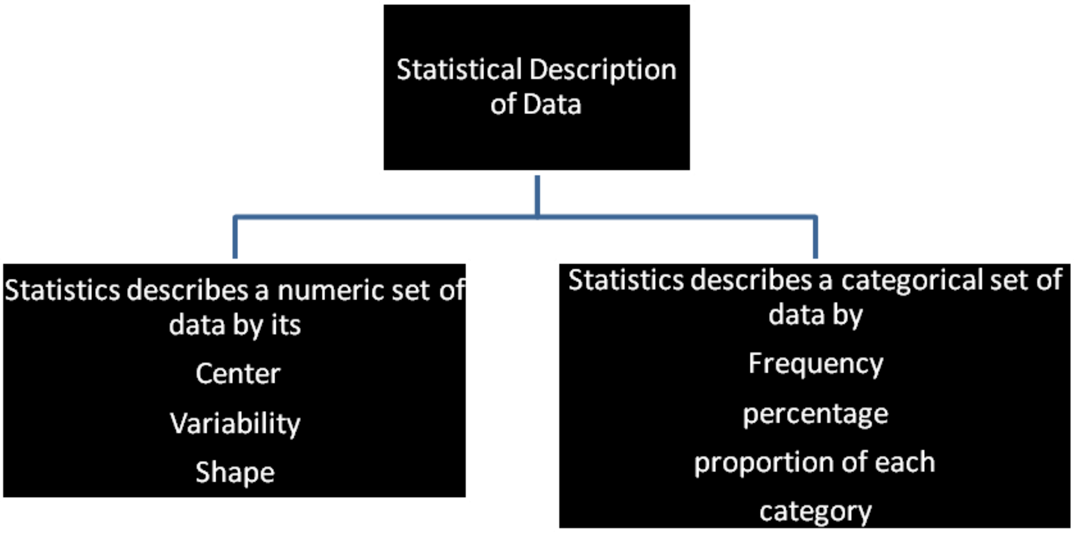
Population: is the aggregate or totality of statistical data forming a subject of investigation. It is a collection of objects, observations or measurements either actual or conceptual.

Sample: A small sub-collection of population. A Sample is a proportion of population, which is examined with a view to estimate the characteristics of the population.

Objectives of sampling

1. Gathering the maximum information about the population with the minimum effort, cost and time.
2. To obtain the best possible values of the parameters under specific conditions.
3. Sampling determines the reliability of these estimates.
4. A generalization from sample to population is called statistical inference.

Statistical model: A mathematical model that symbolizes a set of statistical assumptions about the generation of sample data.



Statistical Description of Data

Statistics describes a numeric set of data by its

Center

Variability

Shape

Statistics describes a categorical set of data by

Frequency

percentage

proportion of each category

Parameter: A parameter is a statistical measure based on the units of a population.
e. g. populations mean, population variance, correlation coefficient etc.

Statistic: It is a statistical measure based on all the observations selected in a sample, independent of population parameter. e. g. sample mean, sample variance, correlation coefficient etc.

Sampling Distribution: Consider all possible samples of size n which can be drawn from a given population at random. For each sample, we can compute mean. The frequency distribution of these means is known as sampling distribution of means.

Similarly, the frequency distribution of standard deviations is known as sampling distribution of the standard deviation.

Standard Error: The standard deviation of the sampling distribution s defined as standard error.

Precision: Reciprocal of standard error is defined as precision.

Estimate: Judgment or opinion of the approximate size or amount.

Statistical Estimation: A part of inference where a population parameter is estimated from the corresponding sample statistics.

Estimate: A statement about an unknown population parameter.

Estimator: Method or rule to determine an unknown population parameter.

Variable: any characteristic of an individual or entity. A variable can take different values for different individuals. Variables can be *categorical* or *quantitative*.

Types of measurement

Discrete: Quantitative data are called discrete if the sample space contains a finite or countably infinite number of values. E.g. how many days did you watch a movie during the last 7 days?

Continuous: Quantitative data are called continuous if the sample space contains an interval or continuous span of real numbers. e. g. Weight, height, temperature

– Height: 1.72 meters, 1.7233330 meters

- **Nominal**
  - Categorical variables with
  - No inherent order or ranking sequence

Value may be a numerical, but without numerical value

e. g. measurement such as female vs. male. The only operation that can be applied to Nominal variables is enumeration.

- **Ordinal**

Variables with an inherent rank or order

e.g. mild, moderate, severe. Can be compared for equality, or greater than or less than, but not *how much* greater or less.

- **Interval**

Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely fixed.

Calendar dates and temperatures on the Fahrenheit scale are examples.

Addition and subtraction can be performed. Multiplication and division are not meaningful operations.

- **Ratio** –

Variables with all properties of Interval plus an absolute, non-arbitrary zero point,

e.g. age, weight, temperature (Kelvin).

Addition, subtraction, multiplication, and division are all meaningful operations.

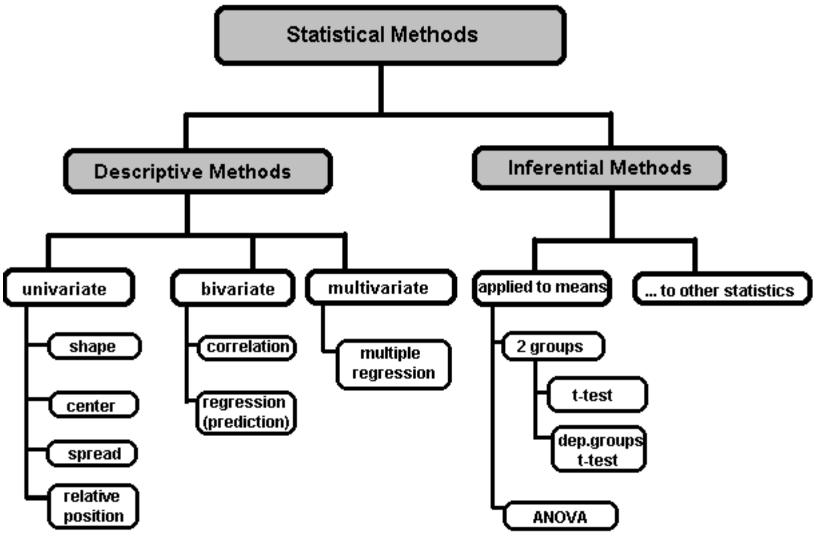
Qualitative vs. Quantitative variables –

Qualitative variables: values are texts e.g. Female, male, boy, girl

Also called as string variables.

Quantitative variables: numeric variables.

A Taxonomy of Statistics



Frequency Distribution:

Consider a data set of 26 children of ages 1-6 years

Ungrouped Data

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

Grouped Data

Age Group	1-2	3-4	5-6
Frequency	8	12	6

Data Presentation

Graphical Presentation:

We look for the overall pattern and for striking deviations from that pattern. Over all pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an *outlier*.

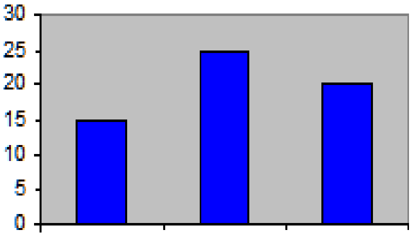
Bar diagram and Pie charts are used for categorical variables.

Histogram, stem and leaf and Box-plot are used for numerical variable.

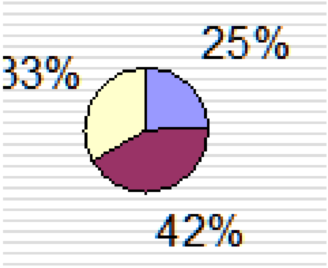
Data Presentation –Categorical Variable

Bar Diagram: Lists the categories and presents the percent or count of individuals who fall in each category.

Pie Chart: Lists the categories and presents the percent or count of individuals who fall in each category.



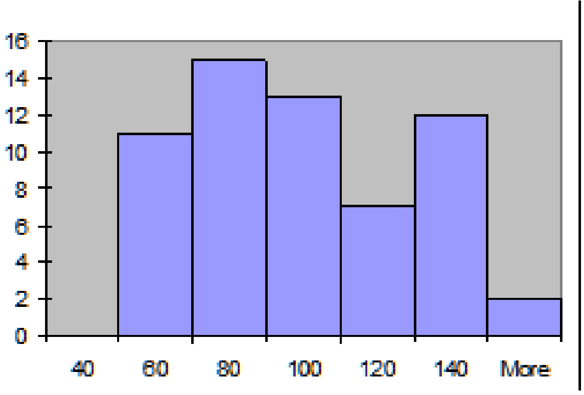
Bar Chart



Pie Chart

Graphical Presentation –Numerical Variable

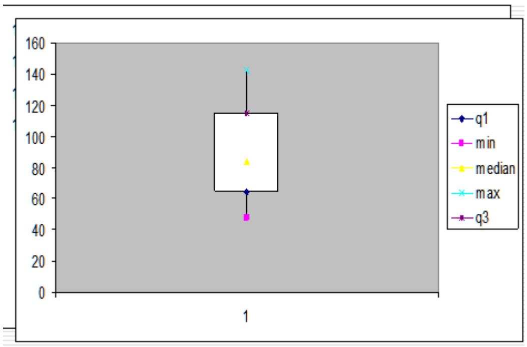
Histogram: Overall pattern can be described by its shape, center, and spread.



Histogram

Box Plot: A box plot is a graph of the five number summary. The central box spans the quartiles. A line within the box marks the median.

Lines extending above and below the box mark the smallest and the largest observations (i.e., the range). Outlying samples may be additionally plotted outside the range.



The five number summary of a box plot consists of

- the smallest (Minimum) observation,
- the first quartile (Q1)
- the median(Q2)
- the third quartile
- the largest (Maximum) observation written in order from smallest to largest.

Distribution - of a variable tells us what values the variable takes and how often it takes these values.

- Unimodal - having a single peak
- Bimodal - having two distinct peaks
- Symmetric - left and right half are mirror images.

Measures of Central Tendency

- Center measurement is a summary measure of the overall level of a dataset

Mean, Median, Mode, Geometric mean and Harmonic Mean

Let $x_1, x_{2,} ... x_n$ are n observations of a variable x.

Then the mean of this variable,

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

For frequency data

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + ... + f_n x_n}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Note that if the data is grouped data then each x_i is the class representative of that class

i.e., $x_i = \dfrac{x_i^l + x_i^h}{2}$, where

x_i^l :lower bound for i$^{\text{th}}$ class

x_i^h :upper bound for i$^{\text{th}}$ class

Median: The middle value in an ordered sequence of observations.

Mode: The value that is observed most frequently. The mode is undefined for sequences in which no observation is repeated.

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions.

Mean – Mode = 3 (Mean – Median)

Median for Individual series

In individual series, where data is given in the raw form,arrange the data in ascending or descending order.

- If the value of **N** is odd then simply the value of **(N+1)/2** th item is median for the data.
- If the value of **N** is even, then **Median = [(N+1)/2 item + (N/2 + 1)th item]/2**

Median from Discrete Data

When the data follows a discrete set of values,

use $\left(\frac{N+1}{2}\right)^{th}$ item for finding the median.

First form a cumulative frequency distribution.

The median is that value which corresponds to

the cumulative frequency in which $\left(\frac{N+1}{2}\right)^{th}$ item lies.

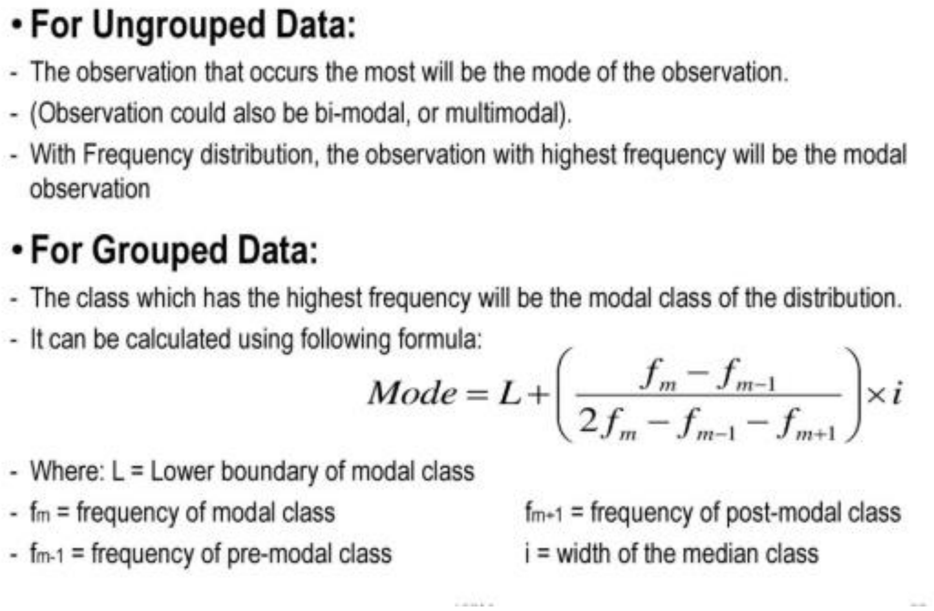
Grouped Data

Step 1: Construct the cumulative frequency distribution.
Step 2: Decide the class that contain the median.
Class Median is the first class with the value of cumulative frequency equal at least n/2.
Step 3: Find the median by using the following formula:

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m}\right) i$$

Where:
n = the **total frequency**
F = the **cumulative frequency** ***before*** class median
f_m = the **frequency** of the class median
i = the class width
L_m = the **lower boundary** of the class median

Mode



- **For Ungrouped Data:**
- The observation that occurs the most will be the mode of the observation.
- (Observation could also be bi-modal, or multimodal).
- With Frequency distribution, the observation with highest frequency will be the modal observation

- **For Grouped Data:**
- The class which has the highest frequency will be the modal class of the distribution.
- It can be calculated using following formula:

$$Mode = L + \left(\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}}\right) \times i$$

- Where: L = Lower boundary of modal class
- f_m = frequency of modal class
- f_{m-1} = frequency of pre-modal class

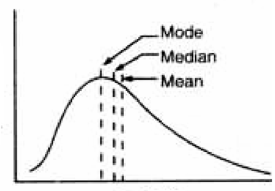
f_{m+1} = frequency of post-modal class
i = width of the median class

Shape of Data measured by

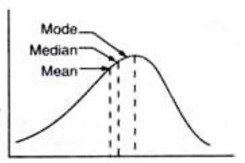
- **Skewness: Lack of symmetry**
- **Kurtosis: Degree of peakedness**

Skewness: measures asymmetry of data

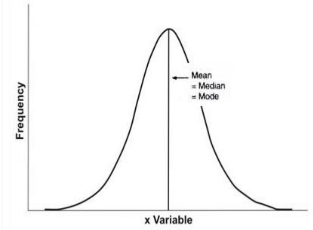
- Positive or right skewed: Longer right tail
- Negative or left skewed: Longer left tail



Positively Skewed Frequency Distribution

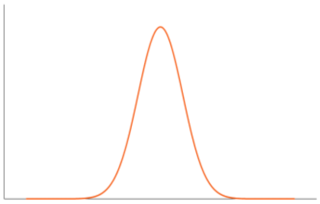


Negatively Skewed Frequency Distribution

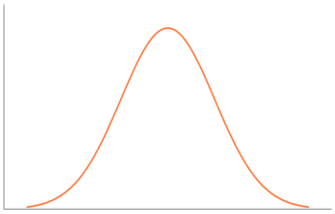


Mean > Median > Mode **Mean < Median < Mode** **Mean = Median = Mode**

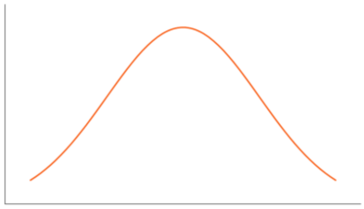
Kurtosis measures peakeness of the distribution



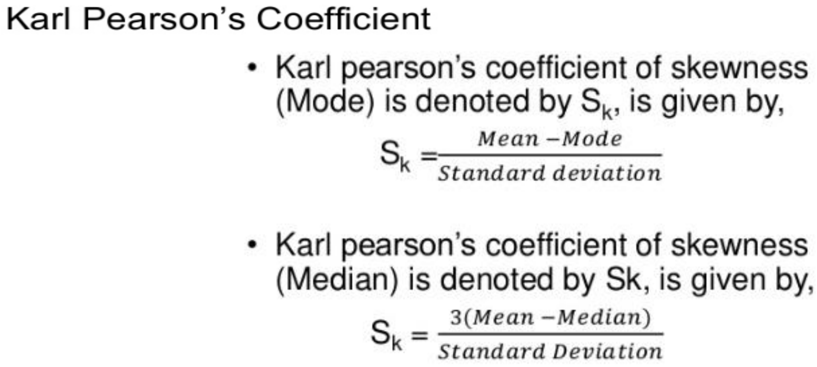
Leptokurtic positive excess kurtosis



Mesokurtic Zero kurtosis



Platykurtic negative excess kurtosis



Karl Pearson's Coefficient

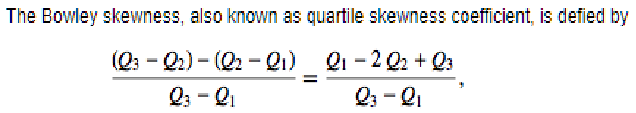
- Karl pearson's coefficient of skewness (Mode) is denoted by S_k, is given by,

$$S_k = \frac{Mean - Mode}{Standard\ deviation}$$

- Karl pearson's coefficient of skewness (Median) is denoted by Sk, is given by,

$$S_k = \frac{3(Mean - Median)}{Standard\ Deviation}$$

Bowley's Coefficient of Skweness :



The Bowley skewness, also known as quartile skewness coefficient, is defied by

$$\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 - 2\,Q_2 + Q_3}{Q_3 - Q_1},$$

Note that

Coefficient of Skewness $S_k > 0 \Rightarrow$ data is positively skewed.

Coefficient of Skewness $S_k < 0 \Rightarrow$ data is negatively skewed.

Moments about Mean

r^{th} moment μ_r about mean $\bar{x}$ is given by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^r \text{, where}$$

$$N = \sum_{i=1}^{n} f_i = \text{total number of observations}$$

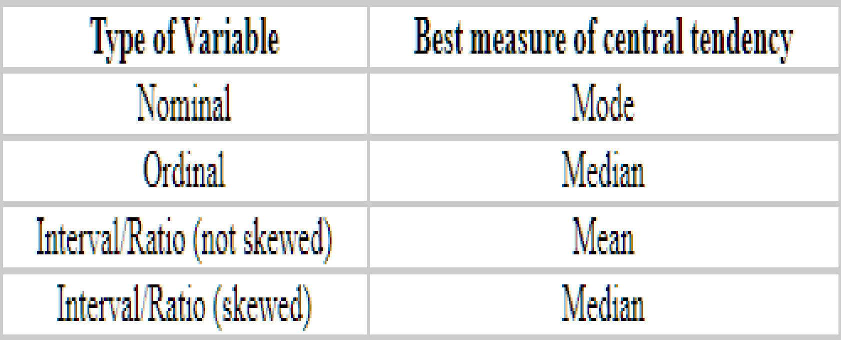
The coefficient of Kurtosis or Kurtosis is given by $\beta_2 = \frac{\mu_4}{\mu_2^2}$, *where*

μ_4 : 4th moment about mean, μ_2 :second moment or variance.

Note that :

1. If $\beta_2 = 3$, the data is normal or mesokurtic
2. If $\beta_2 > 3$, the data is peaked or leptokurtic
3. If $\beta_2 < 3$, the data is flat topped or platykurtic

Type of Variable Vs Measures of Central Tendency



Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Methods of Variability Measurement

Variability (or dispersion) measures the amount of scatter in a dataset.

Commonly used methods: *range*, *variance*, *standard deviation*, *inter quartile range*, *coefficient of variation etc*.

Range: The difference between the largest and the smallest observations.

Population Variance:

Let $x_1, x_{2,}...x_n$ are n observations of a variable x.

Then the variance of this variable, $Var(X) = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}$

For frequency data $Var(X) = \frac{\sum_{i=1}^{n} f_i\left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{n} f_i}$

Standard Deviation: Square root of the variance. $\sigma_X = +\sqrt{Var(X)}$

Sample Variance: If a sample of size N is drawn from a population, then the variance is given by

For ungrouped data $Var(X) = s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{N-1}$.

For grouped data $Var(X) = s^2 = \frac{\sum_{i=1}^{n} f_i\left(x_i - \overline{x}\right)^2}{N-1}, N = \sum_i f_i$.

Quartiles: Data can be divided into four regions that cover the total range of observed values.

Cut points for these regions are known as **quartiles**.

Quartiles of a data is the ((n+1)/4)th observation of the data, where q is the desired quartile and n is the number of observations of data.

The first quartile (Q1) is the first 25% of the data.

The second quartile (Q2) is between the 25th and 50th percentage points in the data.

The upper bound of Q2 is the median.

The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

Inter-quartile Range: Difference between Q3 and Q1.

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles.

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles.

25th percentile is the Q1, 50th percentile is the Median (Q2) and the 75th percentile of the data is Q3.

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent. $Coefficient\ of\ Variation = \frac{\sigma_X}{\bar{x}} \times 100$.

Softwares to perform statistical analysis and visualization of data.

SAS (System for Statistical Analysis), S-plus, R, Matlab, Minitab, BMDP, Stata, SPSS, StatXact, Statistica, LISREL, JMP, GLIM, HIL, MS Excel etc.

Random Variable

Basic Preliminaries: -

- **Random Experiment:** Experiment whose outcomes are not predictable.
- **Sample Point:** Each and every outcome of random experiment.
- **Sample Space:** Totality or aggregate of sample points. It is denoted by symbol S or Ω
- **Event:** A subset of sample space.
- **Impossible Event:** Event which does not contain any sample point.
- **Certain Event:** Event which contains all sample points**.**
- **Mutually Exclusive events:** The happening of any one event excludes the happening of the other event.
- **Exhaustive Event:** The events $\{A_1, A_2, \cdots, A_n\}$ are said to be exhaustive if $\bigcup_{i=1}^{n} A_i = S$ **.**
- **Independent Events:** The occurrence or non-occurrence of one does not affect the occurrence or non-occurrence of the other.

# Probability

If a random experiment results in 'n' mutually exclusive and equally likely outcomes of which 'm' are favourable to event A then probability of event A, denoted as $p(A)$ and is defined as $p(A) = \frac{m}{n}$.

Standard Results ::

1. $0 \le p(A) \le 1$
2. $p\left(A^C\right) = p\left(\overline{A}\right) = 1 - p(A)$
3. $p(\phi) = 0,\ p(S) = 1$
4. If $A \subseteq B$ then $p(A) \le p(B)$
5. $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
6. If A and B are **mutually exclusive** events then $p(A \cap B) = 0$
7. In particular, A and B are mutually exclusive $p(A \cup B) = p(A) + p(B)$
8. **Conditional Probability:** $p(A/B) = \frac{p(A \cap B)}{p(B)}$, $p(B/A) = \frac{p(A \cap B)}{p(A)}$**.**

   Thus $p(A \cap B) = p(A)\,p(B/A)$ **,** $p(A \cap B) = p(B)\,p(A/B)$

For n events $\mathrm{A}_1, \mathrm{A}_2, \cdots, \mathrm{A}_\mathrm{n}$ the probability of intersection event is

$$p(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_n) = p(A_1)\,p(A_2 / A_1)\,p(A_3 / A_1 \cap A_2) \cdots p(A_n / A_1 \cap A_2 \cap \cdots A_{n-1})$$

Sufficient condition for above result to hold is $p(A_1 \cap A_2 \cap \cdots A_{n-1}) > 0$.

9. If A and B are **independent** events then $p(A\cap B)=p(A)p(B)$.

10. $\{A_1,A_2,\cdots,A_n\}$ is a set of mutually exclusive events then $\sum_{i=1}^{n} p(A_i)=p(S)=1$.

11. **Binomial Probability ::** An experiment is performed n number of times repeatedly. A is an event known as success with probability p. If event A occurs r times among n trials then $P(r\,successes)={}^{n}C_r p^r q^{n-r}$, where p is the probability of success and $q=1-p$ the probability of failure.

Random Variables

Let Ω be a sample space associated with given random experiment. A function defined from Ω to a number set is known as **Random Variable**. Thus a random variable is a function whose domain set is a sample space of some random experiment.

The collection of values actually taken by a random variable is known as **Range** of the random variable. Depend on the range; random variables get classified into the following two types.

- **Discrete Random Variables::** Random variable whose range set is a discrete set or the set of natural numbers is known as discrete random variable.

e.g. Number of accidents on Mumbai-Pune expressway, Number of Neutrons emitted by a radioactive isotope, Number of students in a class.

- **Continuous Random Variables::** Random variable whose range set is an interval is known as continuous random variable.

e.g. Lifetime of an electrical or electronic components, Height of students, Amount of cosmic radiation, Time between arrivals of two flights at certain airport, Time taken for surgery at a certain hospital.

Note : Random variables arising through the counting processes are usually discrete in nature while those which arises through measuring processes are continuous in nature.

Examples of Random Variables

1. Consider the experiment of tossing of two fair coins simultaneously. The sample space is

$\Omega=\{HH,HT,TH,TT\}$. Let $X_1=$count of heads in a single toss. By this every sample point get assign by a real number viz $X_1(HH)=2, X_1(HT)=X_1(TH)=1, X_1(HH)=0$.

Thus X_1 is a random variable defined on Ω. Here Range of $X_1=\{0,1,2\}$. Since the range of X_1 is a subset of discrete set, X_1 is a discrete random variable.

Similarly, $X_2=$count of tails in a single toss is also a discrete random variable defined on Ω.

2. Consider the experiment of tossing of two fair dice simultaneously.

Sample space $\Omega=\{(i,j)/1\le i,j\le 6\}$.

i) $X_1=$ Sum of the numbers appear on the two faces, i.e., $X_1((i,j))=i+j$.

Range of $X_1=\{2,3,4,5,6,7,8,9,10,11,12\}$.

ii) $X_2=$ Product of the numbers appear on the two faces, i.e., $X_2((i,j))=ij$

Range of $X_2=\{1,2,3,\cdots,36\}$.

iii) $X_3=$ Minimum or Maximum of the two numbers appear on the two faces, i.e., $X_3((i,j))=\min(i,j)\,or\max(i,j)$. Range of $X_3=\{1,2,3,4,5,6\}$.

Note : Range of all X_1,X_2 and X_3 is a subset of set of Natural number set(Discrete set)

Hence X_1,X_2 and X_3 are discrete random variables (drv).

iv) $X_4=$ Quotient of the two numbers appear on the two faces, i.e., $X_4((i,j))=i/j$. Range of $X_4=\left\{\frac{i}{j}/1\le i,j\le 6\right\}$. Here X_4 is taking integer as well as rational values. Therefore range of X_4 is an interval $\left(\frac{1}{6}\ 6\right)$. Hence X_4 is a continuous random variable (crv).

v) $X_5 =$ Quotient of the sum of two numbers appear on the two faces, i.e.,$X_5(i,j) = \frac{1}{i+j}$. Range of $X_5 = \left\{\frac{1}{i+j}/1 \le i,j \le 6\right\}$. Here X_5 is taking integer as well as rational values. Therefore range of $X_5$is an interval $\left(\frac{1}{12}, \frac{1}{2}\right)$. Hence X_5 is a continuous random variable (crv).

We first discussed discrete random variables (drv).

Probability Mass Function (p. m. f.)

The probability mass function of discrete random variable X with range set $\{x_1, x_2, \cdots, x_n\}$ defined on a sample space Ω is the assignment $p_i = p(x_i) = p(X = x_i)$ such that

i) $p(x_i) \ge 0 \ \forall \ i = 1, 2, 3, \cdots, n.$ ii) $\sum_{i=1}^{n} p(x_i) = 1.$

The table containing the value of X along with the probabilities given by probability mass function is called probability distribution of the random variable X .

Examples

1. Consider the experiment of tossing of 3 coins simultaneously.
$\Omega = \{HHH, HHT, HTH, THH, HTT, TTH, THT, TTT\}$.
Let X be count of heads. Then c. d. f of X can be tabulated as follows.

$X = x_i$	0	1	2	3
$p(X = x_i)$	$1/8$	$3/8$	$3/8$	$1/8$

2. Consider the experiment of tossing of 2 fair dice simultaneously.
$\Omega = \{(i, j)/1 \le i, j \le 6\}$. Let X be count of sum of the numbers appear on the faces.

$X = x_i$	2	3	4	5	6	7	8	9	10	11	12
$p(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

3. A point is chosen at random in a circle of radius r. Let X be the distance of the point from the centre of the circle. Then the range of X is the closed interval with end points 0 and r , i.e., range is $[0, r]$. Here X is a continuous random variable.

Distribution Function or Cumulative Distribution Function (c. d. f)

Let Xbe a discrete random variable with range set $\{x_0, x_1, .., x_n\}$.The distribution function of X denoted as F_X, is the probability of the event$\{X \le a\}$, i. e.,

$$F_X(a) = p(X \le a) = \sum_{x_i \le a} p(X = x_i)$$

Let X be a discrete random variable taking values $\{x_1, x_2, \cdots, x_n\}$ with the p.m.f

$X = x_i$	x_1	x_2	x_3	$\cdots\cdots$	x_n
$p(X = x_i)$	p_1	p_2	p_3	$\cdots\cdots$	p_n

Then the cumulative distribution function (cdf) of X is

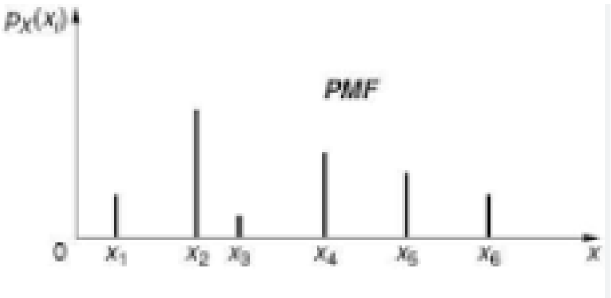
$X = x_i$	x_1	x_2	x_3	$\cdots\cdots$	x_n
$p(X = x_i)$	p_1	p_2	p_3	$\cdots\cdots$	p_n

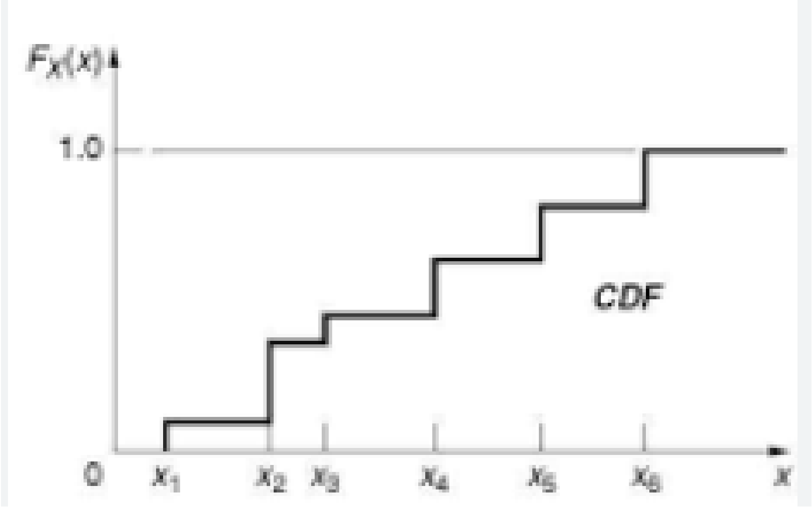
$F_X(a)$	p_1	p_1+p_2	$p_1+p_2+p_3$	$\cdots\cdots$	$\sum_{i=1}^{n} p_i = 1$

Note that: i) Probability Mass Function (pmf) is a function of discrete variable. There are two ways for graphical representation of pmf. The bar chart and the histogram. The sum of the lengths of the bars in the bar chart is 1 whereas the sum of the areas of the rectangles in the histogram is 1.

ii)$F_X$is a function of real continuous variable a. Graph of this function is step or staircase.

Graphs of p.m.f and c.d.f :





Illustrative Examples

Q 1) If X is a random variable the difference between heads and tails obtained when a fair coin is tossed 3 times. What are the possible values of X and its probability mass function? Also write the distribution function of X .

Sol$^{\mathbf{n}}$. $\Omega = \{HHH, HHT, HTH, THH, HTT, TTH, THT, TTT\}$. X can values from -3 to 3, i.e., Range of $X = \{-3,-1,1,3\}$. The pmf is

$X = x_i$	-3	-1	1	3
$p(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
$F_X(x)$	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{7}{8}$	$\frac{8}{8} = 1$

Q 2) A fair dice is rolled twice. Find the possible values of random variable X and its associated probability mass function, where X is the maximum of the values appearing in 2 rolls.

Sol$^{\mathbf{n}}$. $\Omega = \{(i, j)/1 \le i, j \le 6\}$. X can take values from 1 to 6, i.e., Range of $X=\{1,2,3,4,5,6\}$. The pmf is

$X=x_i$	1	2	3	4	5	6
$p(X=x_i)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Q 3) A random variable X takes values $-3, -1, 2, 5$ with respective probabilities $\frac{2k-3}{10}, \frac{k+1}{10}, \frac{k-1}{10}$ and $\frac{k-2}{10}$. Determine the distribution of X.

Sol^{n}. Theassignments are probabilities, equating the to one

$\frac{2k-3}{10}+\frac{k+1}{10}+\frac{k-1}{10}+\frac{k-2}{10}=1 \Rightarrow k=3$. Hence the distribution of X is

X	-3	-1	2	5
$p(X=x)$	$3/10$	$4/10$	$2/10$	$1/10$

Q 4) A random variable X has probability mass function (pmf) shown in the following tabular form. Find the value of unknown k. Hence write pmf and cdfof X. Draw graphs of pmf and cdf . Also find i) $p(1\le X<3)$ ii) $p(1<X\le 3)$

iii) $p(X<1)$ iv) $p(X>5)$

X	1	2	3	4	5
$p(X=x)$	$k/36$	$3k/36$	$4k/36$	$10k/36$	$18k/36$

Sol^{n}. Since the above assignment is probability distribution, sum of the probabilities is one implies $k=1$. Thus the pmf and cdf of X are

X	1	2	3	4	5
$p(X=x)$	$1/36$	$3/36$	$4/36$	$10/36$	$18/36$
$F_X(x)$	$1/36$	$4/36$	$8/36$	$18/36$	$36/36=1$

i) $p(1\le X<3)=p(X=1)+p(X=2)=\frac{1}{36}+\frac{3}{36}=\frac{4}{36}=\frac{1}{9}$.

ii) $p(1<X\le 3)=p(X=2)+p(X=3)=\frac{3}{36}+\frac{4}{36}=\frac{7}{36}$.

1. **Mathematical Expectation / Theoretical Mean** (analogous to Centre of Gravity)

Theoretical mean or expectation of X denoted as $E(X)$ or μ is defined as

$$E(X)=\sum_{i=1}^{n} x_i p_i .$$

Expectation value of X provides a central point of the distribution.

Note: Expected value of a random variable may not be actually taken by the variable.

2. **Variance**

Variance of X denoted as $Var(X)$.

$Var(X)=\sum_{i=1}^{n}\left(x_i-E(X)\right)^2 p_i$. This can be simplified as

$Var(X)=E(X^2)-\left(E(X)\right)^2$, where $E(X^2)=\sum_{i=1}^{n} x_i^2 p_i$

3. **Standard Deviation** $sd=+\sqrt{Var(X)}$

Results : Let X *and* Y be two random variables. Let a *and* b be any non zero constants.

i) $E(aX+b)=aE(X)+b$

ii) $E(X+Y)=E(X)+E(Y)$

iii) $Var(aX+b)=a^2 Var(X)$

iv) $sd(aX+b)=|a|\,sd(X)$

Illustrative Examples

Q 1) A random variable X has the following p.m.f.

X	0	1	2	3	4	5	6
$p(X=x_i)$	k	$3k$	$5k$	$7k$	$9k$	$11k$	$13k$

Find (i) k (ii) $p(X \geq 2)$ (iii) $p(0 < X < 5)$

(iv) What is the minimum value of C for which $p(X \leq c) > 0.5$

(v) What is distribution function of X ?

Sol^{n}. (i) $\sum_i p(X=x_i)=1 \Rightarrow k=\frac{1}{49}$.

(ii) $p(X \geq 2)=1-p(X<2)=1-p(X=0)-p(X=1)=\frac{45}{49}$

(iii) $p(0<X<5)=\sum_{i=1}^{4} p(X=x_i)=\frac{15}{49}$.

(iv) $p(X \leq 3)=\frac{16}{49}<0.5$ and $p(X \leq 4)=\frac{25}{49}=0.51$

$\therefore$ Minimum C is 4.

(v) p.m.f.

$X=x_i$	0	1	2	3	4	5	6
$p(X=x_i)$	$\frac{1}{49}$	$\frac{3}{49}$	$\frac{5}{49}$	$\frac{7}{49}$	$\frac{9}{49}$	$\frac{11}{49}$	$\frac{13}{49}$

c.d.f. $F(a)=p(X \leq a)$

$X=x_i$	0	1	2	3	4	5	6
$F(a)$	$\frac{1}{49}$	$\frac{4}{49}$	$\frac{9}{49}$	$\frac{16}{49}$	$\frac{25}{49}$	$\frac{36}{49}$	1

Q 2) Determine k such that the following functions are p.m.f.s

Sol^{n}. 1. $P(x)=k\,x, \quad x=1, 2, 3, \dots, 10$

$$k(1+2+3+4+5+6+7+8+9+10)=1 \quad \therefore k=\frac{1}{55}$$

2. $P(x)=k\frac{2^x}{x!}, \quad x=0, 1, 2, 3$

$$k\left(1+2+2+\frac{4}{3}\right)=1 \qquad \therefore k=\frac{3}{19}$$

3. $P(x)=k(2x^2+3x+1), x=0, 1, 2, 3$

$$k(1+6+15+28)=1 \qquad \therefore k=\frac{1}{50}$$

Q 3) Verify whether the assignment $p(X=n)=2^{-n}, n=1,2,3,\cdots$ is a probability mass function for random variable X .

Sol^{n}. To be a probability mass function

i) $p(X=n)>0 \,\forall\, n$ ii) $\sum_{n=1}^{\infty} p(X=n)=1$

$p(X=n)=\frac{1}{2^n}$ is in exponential function, condition i) holds.

Further $\sum_{n=1}^{\infty} p(X=n) = \sum_{n=1}^{\infty}\frac{1}{2^n} = \sum_{n=0}^{\infty}\left(\frac{1}{2}\right)^n - 1 = \frac{1}{1-(1/2)} - 1 = 1$

Hence condition ii) also holds. Therefore the assignment is a probability mass function.

Q 4) A box contains 8 items of which 2 are defective. A person draws 3 items from the box. Determine the expected number of defective items he has drawn.

Sol^n. Let X be the number of defective items drawn by a person. The pmf of X is

X	0	1	2
$p(X=x)$	$\frac{20}{56}$	$\frac{30}{56}$	$\frac{6}{56}$

The expected number of defects $= E(X) = \sum x\, p(X=x) = \frac{3}{4}$.

Q 5) A random variable X take four values with probabilities viz $\frac{1+3x}{4}$, $\frac{1-x}{4}$, $\frac{1+2x}{4}$ and $\frac{1-4x}{4}$. Find the condition on X that these values represent the probability mass function.

Sol^n. Let $p_1 = \frac{1+3x}{4}$, $p_2 = \frac{1-x}{4}$, $p_3 = \frac{1+2x}{4}$ and $p_4 = \frac{1-4x}{4}$.

As this is pmf , $0 \le p_i \le 1$ $for\, i = 1,2,3,4$ and $p_1 + p_2 + p_3 + p_4 = 1$.

All these probabilities sum up to 1. Hence consider first condition.

$$0 \le \frac{1+3x}{4} \le 1 \Rightarrow -\frac{1}{3} \le x \le 1 \ , \quad 0 \le \frac{1-x}{4} \le 1 \Rightarrow -3 \le x \le 1$$

$$0 \le \frac{1+2x}{4} \le 1 \Rightarrow -\frac{1}{2} \le x \le \frac{3}{2}, \quad 0 \le \frac{1-4x}{4} \le 1 \Rightarrow -\frac{3}{4} \le x \le \frac{1}{4}.$$

Therefore $-\frac{1}{3} \le X \le \frac{1}{4}$.

Q 6) A random variable has mean 2 and standard deviation $\frac{1}{2}$. Find

i) $E(2X-1)$ ii) $Var(X+2)$ iii) $sd\left(\frac{3X-1}{-4}\right)$

Sol^n. Given $E(X) = 2\, and\, sd(X) = \frac{1}{2} \Rightarrow Var(X) = \frac{1}{4}$

i) $E(2X-1) = 2E(X) - 1 = 2*2 - 1 = 3$.

ii) $Var(X+2) = Var(X) = \frac{1}{4}$.

iii) $sd\left(\frac{3X-1}{-4}\right) = \left|\frac{3}{-4}\right| sd(X) = \frac{3}{4}*\frac{1}{2} = \frac{3}{8}$.

Q 7) A sample space of size 3 is selected at random from a box containing 12 items of which 3 are defective. Let X denote the number of defective items in the sample. Write the probability mass function and distribution function of X. Find the expected number of defective items.

Sol^n. X be the number of defective items in the sample.

X	0	1	2	3
$p(X=x)$	84/220	108/220	27/220	1/220
$F_X(x)$	84/220	192/220	219/220	220/220=1

The expected number of defective items in a sample is $E(X) = \frac{165}{220} = 0.75.$

Q 8) A player tosses two fair coins. The player wins $2 if two heads occur and $1 if one head occur. On the other hand, the player losses $3 if no heads occur. Find the expected gain of the player. Is the game fair?

Sol$^{\mathbf{n}}$. The sample space is $\{HH, HT, TH, TT\}$. Since coins are fair,

$$p(HH) = p(HT) = p(TH) = p(TT) = \frac{1}{4}.$$

Let X be the player's gain. Then X takes values $-3, 1$ and 2 with

$$p(-3) = \frac{1}{4}, \quad p(1) = \frac{2}{4} \text{ and } p(2) = \frac{1}{4}.$$

Expected gain $E(X) = -3\times\frac{1}{4} + 1\times\frac{2}{4} + 2\times\frac{1}{4} = \frac{1}{4} = 0.25$.

Thus the expected gain of the player is $\$0.25$. Further $E(X) > 0$, the game is Favourable to the player.

- **Continuous Random Variable**

Probability density function (p.d.f.)

X a continuous random variable. Function $f(x)$ defined for all real $x \in (-\infty, \infty)$ is called **probability density function** (p.d.f.) if for any set B of real numbers, we get probability $p(X \in B) = \int_B f(x)dx$. Thus, $p(a \le X \le b) = \int_a^b f(x)dx$ .Also, the point probability is zero, i.e., $p(X = a) = 0$.

Axioms

i) $0 \le p(a \le X \le b) \le 1,$ or, $0 \le \int_a^b f(x)dx \le 1$ Note that $p(X \in B) = \int_B f(x)dx$.

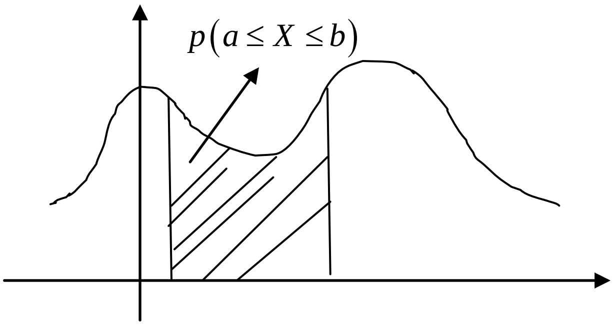
ii) $\int_{-\infty}^{\infty} f(x)dx = 1$.

Let $f(x)$ be the probability density function (pdf) of a continuous random variable X . The curve represented by the equation $y = f(x)$ is called the probability curve of X .

Note That :i) The probability curve lies completely above X - axis.

ii) The total area between the X - axis and the probability curve is unity.

iii) $p(a \le X \le b)$ represents the area under the probability curve bounded by the X - axis and the co-ordinates $X = a$ and $X = b$.



- **Cumulative Distribution Function (c.d.f.)**

X be a continuous random variable.

The cumulative distribution function denoted as $F(a)$ and is expressed as

$$F(a)=\int_{-\infty}^{a} f(x)dx \quad \text{or} \quad \frac{d}{da}f(a)=f(a).$$

Properties of c.d.f.

1) $F(a)$ is monotonically increasing. i.e. $a<b \Rightarrow F(a)\le F(b)$

2) $\lim_{a\to\infty} F(a)=1$

3) $\lim_{a\to-\infty} F(a)=0$

4) $p(a<X<b)=F(b)-F(a)$

Note that : For a continuous random variable X ,

$$p(a<X<b)=p(a\le X<b)=p(a<X\le b)=p(a\le X\le b)$$

5) $p(X>a)=1-p(X\le a)=1-F(a)$

- **Expected value, Variance and Standard Deviation**

$$\mu=E(X)=\int_{-\infty}^{\infty} x\, f(x)dx \quad ,\text{var}(X)=\int_{-\infty}^{\infty}(x-E(X))^2 f(x)dx,\ sd(X)=+\sqrt{Var(X)}$$

Illustrative Examples

Q 1) X cont. random variable with p.d.f

$$F(x)=\begin{cases}0 & \text{if} \quad x<0\\ \frac{x}{2} & \text{if} \quad 0\le x\le 2\\ 1 & \text{if} \quad 2<x\end{cases}$$

Find: i) $p\left(\frac{1}{2}<X<\frac{3}{2}\right)$ (ii) $p(1\le X\le 2)$

Sol$^{\mathbf{n}}$. i) $p(a\le X\le b)=F(b)-F(a)$

$$p\left(\frac{1}{2}<X<\frac{3}{2}\right)=F\left(\frac{3}{2}-\frac{1}{2}\right)=\frac{3}{4}-\frac{1}{4}=\frac{1}{2}.$$

ii) $p(1\le X\le 2)=F(2)-F(1)=1-\frac{1}{2}=\frac{1}{2}.$

Q 2) The time in years X required to complete a software project has a p.d.f.

$$f(x)=\begin{cases}kx(1-x), & 0\le x\le 1\\ 0 & \text{otherwise}\end{cases}$$

Compute the probability that the project will be completed in less than four months.

Solⁿ. $\int_{-\infty}^{\infty} f(x)dx = 1 \quad \Rightarrow \quad k = 6$

The probability that the project will be completed in < 4 months, i.e. less than $\frac{1}{3}$ years.

$$p(X \le 4) = \int_0^{1/3} f(x)\,dx = 0.259\,.$$

Q 3) X continuous random variable with probability density function.

$$f(x) = \begin{cases} \frac{x}{2}, & 0 \le x \le 2 \\ 0, & \text{otherwise} \end{cases}$$

Find (i) $p(1 \le X \le 1.5)$ (ii) $E(X)$ (iii) $\text{Var}(X)$ (iv) c.d.f.

Solⁿ. This is a ramp distribution.

(i) $p(1 \le X \le 1.5) = \int_1^{1.5} f(x)\,dx = \frac{5}{16}$

(ii) $E(X) = \int_0^2 \frac{x^2}{2} dx = \frac{4}{3}$

(iii) $E(X^2) = 2 \quad \therefore \text{Var}(X) = \frac{2}{9}$

(iv) $F(a) = \int_{-\infty}^{a} f(x)dx$

$-\infty < x < 0 \qquad \int_{-\infty}^{0} 0dx = 0$

$0 \le x \le 2 \qquad \int_0^a \frac{x}{2} dx = \frac{a^2}{4}$

$2 \le x < \infty \qquad F(a) = 1$

Q 4) X continuous random variable with probability density function

$$f(x) = \begin{cases} kx, & 0 \le x \le 2 \\ 2k, & 2 \le x < 4 \\ -kx + 6k, & 4 \le x < 6 \end{cases}$$

Find k and mean.

Solⁿ. $\int_0^6 f(x)\,dx = 1 \Rightarrow \int_0^2 kx\,dx + \int_2^4 2k\,dx + \int_4^6 (-kx + 6k)\,dx = 1$

Evaluating the integral and simplifying, we have $8k = 1 \Rightarrow k = \frac{1}{8}$.

$$E(X) = \int_0^6 xf(x)\,dx = \int_0^2 kx^2\,dx + \int_2^4 2kx\,dx + \int_4^6 (-kx^2 + 6kx)\,dx$$

$$= k\left(\frac{x^3}{3}\right)_0^2 + 2k\left(\frac{x^2}{2}\right)_2^4 - k\left(\frac{x^3}{3}\right)_4^6 + 6k\left(\frac{x^2}{2}\right)_4^6$$

Simplifying $E(X) = 3$.

Q 5) i) Verify $f(x) = 6x(1-x), 0 \le x \le 1$ is a probability density function for the diameter of a cable.

ii) Find the condition on b such that $p(X < b) = p(X > b)$.

Sol$^{\mathbf{n}}$. i) At $x = 0,\ f(x) = 0.$ For $0 < x < 1$, $0 < 1 - x < 1$and $0 < 6x < 6$. Therefore $0 < 6x(1-x) < 6$. Hence $f(x)$ is positive for $0 \le x < 1$.

Further $\int f(x)dx = \int_0^1 6x(1-x)dx = 6\left(\frac{1}{2} - \frac{1}{3}\right) = 1$.

This proves $f(x)$ is pdf for $0 \le x < 1$.

ii) $p(X < b) = p(X > b) \Rightarrow p(X < b) = 1 - p(X \le b)$.

Random variable is continuous, $p(X < b) = p(X \le b)$. Therefore

$$2p(X < b) = 1 \Rightarrow p(X < b) = \frac{1}{2} \Rightarrow \int_0^b f(x)dx = \frac{1}{2}.$$

Solving above equation we have $4b^3 - 6b^2 + 1 = 0.$

Q 6) A continuous random variable X has pdf $f(x) = 3x^2,\ 0 \le x < 1$. Find a such that $p(X > a) = 0.05$.

Sol$^{\mathbf{n}}$. $p(X > a) = 0.05 \Rightarrow \int_a^1 f(x)dx = 0.05$, i.e., $\int_a^1 3x^2 dx = 0.05 \Rightarrow 1 - a^3 = 0.05 \Rightarrow a = (0.95)^{1/3}$.