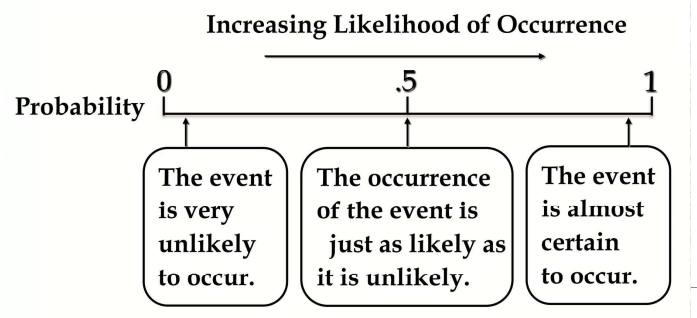
Statistics for IT

PROBABILITY

Week 1

What is **Probability**?

- A probability provides a quantitative description of the chances or likelihoods associated with various outcomes



Why learn **Probability**?

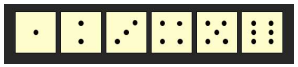
- Probability provides information about the likelihood that something will happen.
  - Meteorologists, use weather patterns to predict the probability of rain
  - In epidemiology, probability theory is used to understand the relationship between exposures and the risk of health effects.
- Understanding **probability gives you the ability to pre future events may turn out**.

Experiments and **Outcomes**

- Experiment – An experiment is any process that generate well defined outcomes
- Outcome – possible result of an experiment
  - Eg: If you toss 2 coins the four possible outcomes are HH,HT,TH,TT

Sample Space

- It is the set of all possible outcomes
- e.g. All 6 faces of a die:



- e.g. All 52 cards of card pack:

Probability of an Event

n(A) – number of elements in the set of the event A

n(S) – number of elements in the set of the sample space

P(A) – probability of event A

$$P(A) = \frac{n(A)}{n(S)}$$

Event

- Event - An event is an outcome of an experiment, usually denoted by a capital letter
- Examples
  - Experiment: Record an age
    - Event A: person is 30 years old
    - Event B: person is older than 65
  - Experiment: Toss a die
    - Event A: observe an odd number
    - Event B: observe a number greater than 2

Mutually Exclusive Events

- If two events cannot occur at the same time they are called mutually exclusive events
- Eg: When tossing a coin, the event of getting head and tail are mutually exclusive.
- For mutually exclusive events $P(A \cap B) = 0$

Independent Events

- If the occurrence of an event A does not affect the occurrence of event B then A and B are independent events
- Eg: simultaneously tossing two coins
- For independent events

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

Conditional Probability

- Conditional probability is a measure of the probability of an event occurring, given that another event has already occurred.
- The conditional probability of event B given the occurrence of event A is

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

Basic rules of probability

- $0 \leq P(A) \leq 1$
- $P(A') = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Statistics for IT

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

Week 2

Random Variable

- When the value of a variable is the outcome of a statistical experiment that variable is a random variable
- Notations
  - X – random variable X
  - P(X=x) The probability that the random variable X is equal to a particular value x
  - Eg: P(X=1) the probability that the random variable X is equal to 1

Statistical Experiment

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

Types of random variables

- Discrete variables – take only integer values
- Continuous variables – take any value within a range of values

Probability Distributions

- Probability distribution is a table, an equation or a graph that links each outcome of a statistical experiment with its probability of occurance

The mean of a discrete probability distribution

- It is also known as expected value
- If the experiment is repeated many times the average value of the random variable is defined as the expected value

$E(x) = \Sigma x\, p(x)$

- Eg: a fair die is tossed. Calculate the expected value

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

$E(x) = (1x\ 1/6) + (2x\ 1/6) + (3x\ 1/6) + (4x\ 1/6) + (5x\ 1/6) + (6x\ 1/6)$

$= 3.5$

Discrete Probability Distributions

- The probability distribution of a discrete random variable can always be represented by a table
  - The probability that X can take a specific value is p(x)
    $P(X=x) = p(x)$
  - p(x) is non negative for all real x
    $0 \leq p(x) \leq 1$
  - The sum of p(x) over all possible values of x is 1
    $\Sigma\, p(x) = 1$

The variance of a discrete probability distribution

- $V(x) = E(x^2) - [E(x)]^2$ where $E(x^2) = \Sigma x^2\, p(x)$
- Eg: a fair die is tossed. Calculate variance

$E(x) = (1x\ 1/6) + (2x\ 1/6) + (3x\ 1/6) + (4x\ 1/6) + (5x\ 1/6) + (6x\ 1/6)$

$= 3.5$

$E(x^2) = (1^2x\ 1/6) + (2^2x\ 1/6) + (3^2x\ 1/6) + (4^2x\ 1/6) + (5^2x\ 1/6) + (6^2x\ 1/6)$

$= 15.16$

$V(X) = 15.16 - 3.5^2$

$= 2.916$

Statistics for IT

BINOMIAL DISTRIBUTION

Week 3

Binomial random variable

- A binomial random variable is the number of successes (s) in 'n' repeated trials of a binomial experiment

Binomial Experiment

- It is a statistical experiment with following properties
  - A fixed number of trials (n)
  - Each trial should be success or failure
  - The trials are independent
  - The probability of success (p) at each trial is a constant

Binomial Distribution

- The probability distribution of a binomial random variable is called a binomial distribution

  X ~ Bin(n,p)

- The probability function of a binomial distribution is

  $$P(x) = {}^{n}C_{x}\ p^{x}q^{n-x}$$ where q=1-p

- Eg: A coin is tossed 10 times. Find the probability of getting exactly 3 heads

  $$P(x) = {}^{n}C_{x}\ p^{x}q^{n-x}$$

  $$P(3) = {}^{10}C_{3}\ 0.5^{3}0.5^{10-3}$$

  = 0.117

Mean & variance of a binoimial distribution

- $E(x) = np$
- $V(x)=npq$

Statistics for IT

POISSON DISTRIBUTION

Week 4

Binomial random variable

- A binomial random variable is the number of successes (s) in 'n' repeated trials of a binomial experiment

Binomial Experiment

- It is used for events that occurs randomly in a specified unit of space, distance or time
  - A fixed number of trials (n)
  - Each trial should be success or failure
  - The trials are independent
  - The probability of success (p) at each trial is a constant

Binomial Distribution

- The probability distribution of a binomial random variable is called a binomial distribution

  X ~ Bin(n,p)

- The probability function of a binomial distribution is

  $$P(x) = {}^{n}C_{x}\ p^{x}q^{n-x}$$ where q=1-p

- Eg: A coin is tossed 10 times. Find the probability of getting exactly 3 heads

  $$P(x) = {}^{n}C_{x}\ p^{x}q^{n-x}$$

  $$P(3) = {}^{10}C_{3}\ 0.5^{3}0.5^{10-3}$$

  = 0.117

Mean & variance of a binoimial distribution

- $E(x) = np$
- $V(x)=npq$

Statistics for IT

INTRODUCTION TO STATISTICS

Week 5

Types of Data

- **Qualitative Data:** They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated. For example, data on attributes such as intelligence, honesty, wisdom, cleanliness, and creativity collected using the students of your class a sample would be classified as qualitative.
- **Quantitative Data:** These can be measured and not simply observed. They can be numerically represented and calculations can be performed on them. For example, data on the number of students playing different sports from your can be classified as quantitative.

Data

- Data are measurements or observations that are collected as a source of information
- Eg: The number of people in Sri Lanka

  the countries where people were born

  the value of sales of a particular product

Sources of Data

- **Primary Data**

  These are the data that are ***collected for the first time*** by an investigator for a specific purpose. Primary data are 'pure' in the sense that no statistical operations have been performed on them and they are original. An example of primary data is the **Census of Sri Lanka**
- **Secondary Data**

  They are the data that are ***sourced from someplace*** that has originally collected it. This means that this kind of data has already been collected by some researchers or investigators in the past and is available either in published or unpublished form. This information is impure as statistical operations may have been performed on them already. .

Primary Data

Examples for primary data

- Customer surveys
- Market research
- Scientific experiments
- Traffic counts

Primary Data Collection Methods

- **Interviews**
  - It involves two groups of people, where the first group is the interviewer (the researcher(s) asking questions and collecting data) and the interviewee (the subject or respondent that is being asked questions).
  - Interviews can be carried out in 2 ways, namely; in-person interviews and telephonic interviews.
  - An in-person interview requires an interviewer or a group of interviewers to ask questions from the interviewee in a face-to-face fashion. It can be direct or indirect, structured or structure, focused or unfocused

- **Surveys & Questionnaires**
  - They are a group of questions typed or written down and sent to the sample of study to give responses.
  - After giving the required responses, the survey is given back to the researcher to record.
  - There are 2 main types of surveys used for data collection, namely; online and offline surveys.

- **Observation**
  - The observation method is mostly used in studies related to behavioral science.
  - There are different approaches to the observation method—structured or unstructured, controlled or uncontrolled, and participant, non-participant, or disguised approach

Secondary data collection methods

- data are available in various resources including
  - Government publications
  - Public records
  - Historical and statistical documents
  - Business documents
  - Technical and trade journals

Secondary data

Sampling

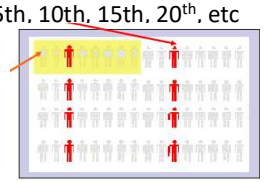
- Sampling is the process of selecting units from a population.
- Population: it is the set of all observations considered in the research
- Sample: *sample* is a subset of a *population*

Advantages of sampling

- Cost is lower
- Data collection is faster
- Improve accuracy & quality
- If the items are destroyed through the test then sampling is the only alternative

Types of probability sampling techniques

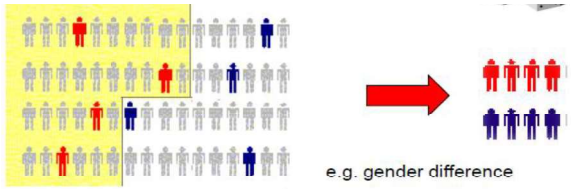
- **Random sampling**
  - In this method, each item in the population has the same probability of being selected as part of the sample as any other item. Random sampling can be done with or without replacement. eg: 100 are listed and a group of 20 may be selected from this list at random.
- **Systematic sampling**
  - In this method, every nth element from the list is selected as the sample. eg: from a list we would select every 5th, 10th, 15th, 20th, etc



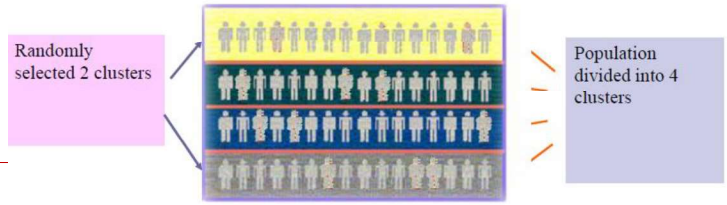
Sampling methods

- **Probability Sampling**
  - In probability sampling technique the *units* are selected from the *population* at *random* using *probabilistic methods*.
  - **reasons for using probability sampling**
    - Making statistical inferences
    - Achieving a representative sample
    - Minimising sampling bias
  - *bias* means that the units selected from the population for inclusion in your sample are *not representative* of that population

- **Stratified sampling**
  - A stratum is a subset of the population that share at least one common characteristic. Within each stratum, a simple random sample or systematic sample is selected. Examples of stratums might be males and females, or managers and non-managers.



- **Cluster Sampling**
  - The population is divided into groups called clusters & a set of clusters are selected randomly to include in the sample.
    - · **One-stage sampling**. All of the elements within selected clusters are included in the sample.
    - · **Two-stage sampling**. A subset of elements within selected clusters are randomly selected for inclusion in the sample



Types of non probability sampling techniques

- · **Quota sampling**
  - In *quota sampling*, the aim is to end up with a *sample* where the *strata* (*groups*) being studied (e.g. males vs. females students) are *proportional* to the *population* being studied. For instance, if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop.
- · **Convenience sampling**
  - A convenience sample is simply one where the *units* that are selected for inclusion in the *sample* are the *easiest to access*. For example if there are 10,000 students, if the *sample size* is 100 students, we may stand at the main entrance, & gather data from passing by students
- · **Purposive sampling**
  - we sample with a *purpose* in mind. First the respondents are verified to check whether they meet the criteria for being in the sample. Purposive sampling can be very useful for situations where you need to reach a targeted sample quickly and where sampling for proportionality is not the primary concern.In super market asking questions from certain people

Non Probability sampling

- In non probability sampling the *samples* are selected based on the *subjective judgement* of the researcher, rather than *random selection*
- reasons for using non probability sampling
  - the *procedures* used to *select units* for inclusion in a *sample* are much easier, quicker and cheaper when compared with *probability sampling*

- **Snowball sampling**
  - In snowball sampling, you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who they may know who also meet the criteria. Snowball sampling is particularly appropriate when the *population* you are interested in is *hidden* and/or *hard-to-reach*. These include *populations* such as drug addicts, homeless people, individuals with AIDS.

Statistics for IT

PRESENTATION OF DATA

Week 6

Tables

Presentation of Data

- Presentation of data makes it easy to understand about a dataset and help to make correct interpretations
- Two methods of presenting data are
  - Tabular form – presenting data in a simple table
  - Pictorial form – data are presented using diagrams, charts or graphs

Frequency table

- The frequency of a data value is the number of times the data value occurs
- Eg: the marks awarded for an assignment for 20 students are as follows. Present this information in a frequency table

6,4,7,10,5,6,7,8,7,8,8,9,7,5,6,6,9,4,7,8

Value	Tally	Frequency
4	//	2
5	//	2
6	////	4
7	~~////~~	5
8	////	4
9	//	2
10	/	1

Cumulative frequency table

- Cumulative frequency is the total of a frequency and all frequencies below it in a frequency distribution
- Eg:

Value	Frequency	Cumulative frequency
4	2	2
5	2	4
6	4	8
7	5	13
8	4	17
9	2	19
10	1	20

Value	Tally	Frequency
28-60	//	2
61-93	~~////~~ ~~////~~	10
94-126	~~////~~ ~~////~~ /	11
127-159	///	3
160-192	///	3
192-225	//	2

- between five to ten rows in a frequency table is suitable.

Group Frequency Table

- When the set of data values are spread out, it is difficult to set up a frequency table for every data value as there will be too many rows in the table. So we group the data into **class intervals**
- Eg: The number of calls per day for taxi service was recorded for the month of December. The results were as follows:

28	122	217	130	120	86	80	90	120	140
70	40	145	187	113	90	68	174	194	170
100	75	104	97	75	123	100	82	109	120
81									

Set up a grouped frequency table for this set of data values.

Relative Frequency and percentage frequency table

- Relative frequency & percentage frequency columns for the above table is as follows

Value	Frequency	Relative frequency	Percentage frequency
28-60	2	2/31=0.06	6
61-93	10	10/31=0.32	32
94-126	11	11/31=0.35	35
127-159	3	3/31=0.09	9
160-192	3	3/31=0.09	9
192-225	2	2/31=0.06	6

Definitions

Sales	No of Days
36-40	2
41-45	7
46-50	8
51-55	11
56-60	2

Consider the example above

- Class interval – a range into which data may be grouped
  eg: 34 – 40
- Class limit – 41 is the lower limit and 45 is the upper limit of the class interval 41-45
- Real limit/class boundary – 40.5 & 45.5 are the real limit of the class 41-45
  - Lower real limit = (lower limit of the class+upper limit of the previous class)/2
  - Upper real limit = (upper limit of the class+lower limit of the next class)/2
- Class mark – class mark of 41-45 class is (41+45)/2 = 43
- Class width – difference between the real limits. Class width of 41-45 class is 45.5 – 40.5 =5

Graphs

Histogram

Steps to draw a histogram

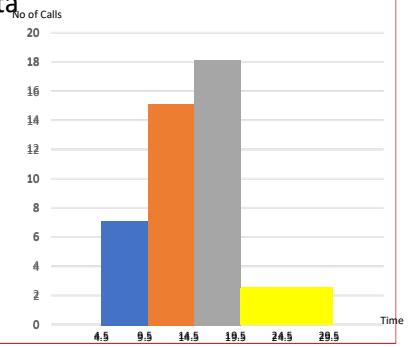
- Mark frequencies in the 'y' axis
- Mark class boundaries in the 'x' axis
- Determine the height of the rectangle
  - If the class widths are equal then take frequency as the height of the rectangle
  - If class widths are not equal

$$\text{height} = \frac{\text{frequency}}{\text{class width}} \times k$$

(k is a constant)

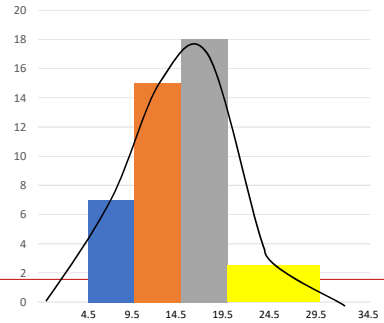
- Eg: draw a histogram for following data

time	No of calls
5-9	7
10-14	15
15-19	18
20-24	5



Frequency Curve

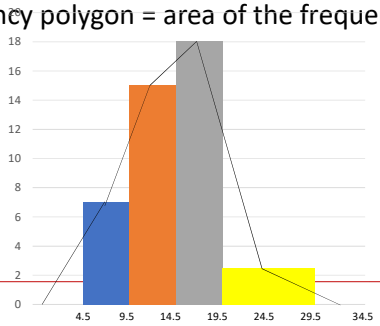
- If we fit a smooth curve to a frequency polygon, we get a frequency curve



Frequency Polygon

- Frequency polygon is a line graph drawn by joining all the mid points of top of the bars of a histogram

Area of the frequency polygon = area of the frequency histogram



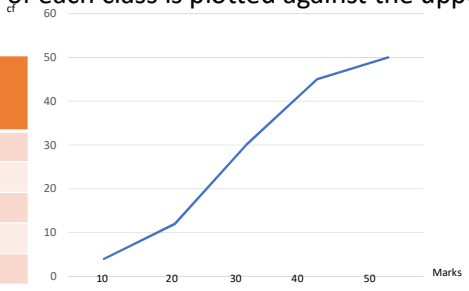
Cumulative frequency curve / Ogive

- Ogive is the graphical representation of a cumulative frequency distribution

Less than ogive

- Cumulative frequencies are in the ascending order
- The cumulative frequency of each class is plotted against the upper limit of the class interval
- Eg:

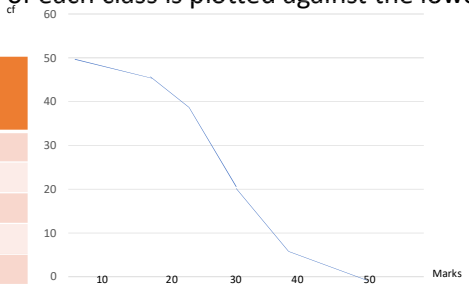
Marks	No of students	Less than cumulative frequency
0-10	4	4
10-20	8	12
20-30	18	30
30-40	15	45
40-50	5	50



Grater than ogive

- Cumulative frequencies are in the decending order
- The cumulative frequency of each class is plotted against the lower limit of the class interval
- Eg:

Marks	No of students	Grater than cumulative frequency
0-10	4	50
10-20	8	46
20-30	18	38
30-40	15	20
40-50	5	5



Statistics for IT

MEASURES OF CENTRAL TENDENCY

Week 7

Mean

- Mean is the average value of a data set
- Different types of means
  - Arithmetic mean
  - Weighted mean
  - Harmonic mean
  - Geometric mean

Central Tendency Measures

- Measures that indicate the central value of a distribution
- The 3 most common measures of central tendency are
  - Mean
  - Median
  - mode

Arithmetic Mean

- Mean of ungrouped data

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum X}{N}$$

- Eg: The arithmetic mean of the numbers 8, 3, 5, 12, and 10 is

$$\bar{X} = \frac{8 + 3 + 5 + 12 + 10}{5} = \frac{38}{5} = 7.6$$

- Mean of ungrouped frequency distribution

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \cdots + f_KX_K}{f_1 + f_2 + \cdots + f_K} = \frac{\sum fX}{\sum f}$$

If 5, 8, 6, and 2 occur with frequencies 3, 2, 4, and 1, respectively, the arithmetic mean is

$$\bar{X} = \frac{(3)(5) + (2)(8) + (4)(6) + (1)(2)}{3 + 2 + 4 + 1} = \frac{15 + 16 + 24 + 2}{10} = 5.7$$

- Eg: Using the given frequency distribution, find the mean. The data represent the number of miles run during one week for a sample of 20 runners.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$
5.5–10.5	1	8	8
10.5–15.5	2	13	26
15.5–20.5	3	18	54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		$\Sigma f \cdot X_m = 490$

$$\bar{X} = \frac{\Sigma f \cdot X_m}{n} = \frac{490}{20} = 24.5 \text{ miles}$$

- Mean of grouped frequency distribution

$$\bar{X} = \frac{\sum fX}{\sum f}$$

Step 1 Make a table as shown.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\bar{X} = \frac{\Sigma f \cdot X_m}{n}$$

[*Note:* The symbols $\Sigma f \cdot X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.]

Weighted mean

Find the **weighted mean** of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\Sigma wX}{\Sigma w}$$

where $w_1, w_2, \ldots, w_n$ are the weights and $X_1, X_2, \ldots, X_n$ are the values.

- Eg: A student received an A in English Composition (3 credits), a C in Introduction to Psychology (3 credits), a B in Biology (4 credits), and a D in Physical Education (2 credits). Assuming A= 4 grade points, B = 3 grade points, C = 2 grade points, D = 1 grade point, and F = 0 grade points, find the student's grade point average.

Solution

Course	Credits (w)	Grade (X)
English Composition I	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology I	4	B (3 points)
Physical Education	2	D (1 point)

$$\overline{X} = \frac{\Sigma wX}{\Sigma w} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} = 2.7$$

The grade point average is 2.7.

Median

- Median is the middle value of an ordered set of data

- Median of ungrouped data
  - if there are n numbers median is $\frac{(n+1)^{th}}{2}$ term
  - Eg: The number of rooms in the seven hotels in the town X is 713, 300, 618, 595, 311, 401, and 292. Find the median.

Solution

Step 1 Arrange the data in order.

292, 300, 311, 401, 595, 618, 713

Step 2 Select the middle value.

292, 300, 311, 401, 595, 618, 713

↑

Median

Hence, the median is 401 rooms.

  - The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median.684, 764, 656, 702, 856, 1133, 1132, 1303

Solution

656, 684, 702, 764, 856, 1132, 1133, 1303

↑

Median

Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$MD = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

- Median of ungrouped frequency distributions
- Eg:

x	f	Cf
0	4	4
1	3	7
2	2	9
3	4	13
4	1	14
5	1	15

(n+1)/2 th term =(15+1)/2 th term = 8^{th} term

Median is 2

- Median of grouped frequency distributions

$$M_d = L + \frac{\left(\frac{n}{2} - C\right)}{f} \times h$$

where

- L- lower real limit of the median class,
- C- cumulative frequency preceding the median class
- f - frequency of the median class
- h – width of the median class
- n – number of data

- Eg:

height	frequency	Cf
60-62	5	5
63-65	18	23
66-68	42	65
69-71	27	92
72-73	8	100

$M_d = 65.5 + \left(\frac{100/2 - 23}{42}\right) \times 3$

$= 67.42$

Finding the Median Class

- To determine the median class for grouped data:
- Construct a cumulative frequency distribution.
- Divide the total number of data values by 2.
- Determine which class will contain this value. For example, if n=50, 50/2 = 25, then determine which class will contain the 25th value - the median class.

Statistics for IT

MEASURES OF CENTRAL TENDENCY

Week 8

Mode

- Mode is the most repeated term in a group of data

Unimodal

- A data set that has only one value that occurs with the greatest frequency is said to be **unimodal.**
- Mode of 3,7,9,4,5,6,3,7,1,7 is 7

Bimodal

- If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be **bimodal.**
- There are 2 modes for the data set 2,4,2,4,5,4,7,8,9,8,8. they are 4 and 8. these kind of distributions are called bimodal distributions

Multimodal

- If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal.**

No mode

- When no data value occurs more than once, the data set is said to have ***no mode.***
- There is no mode for data set 3,5,8,7,9,4,1

- Mode of ungrouped frequency distribution
  - Eg:

x	f
2	8
3	2
5	16
4	1

Mode is 5

- Mode of grouped frequency distribution

$$M = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c$$

- L - lower boundary of the modal class
- Δ_1 - difference between the frequency of the modal class and the class preceding it
- Δ_2 - difference between the frequency of the modal class and the class after it
- C - class interval of the modal class

- Eg: find the mode

Time	Frequency
1-10	8
11-20	14
21-30	12
31-40	9
41-50	7

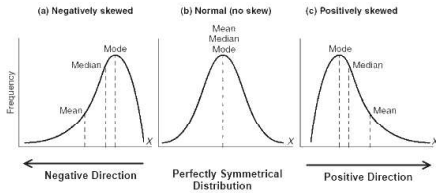
M = 10.5+(6)10

6+2

=18

Relationship between mean,median,mode

- Mean-mode = 3(mean-median)



Statistics for IT

MEASURES OF DISPERSION Cont.....

Week 10

Co-efficient of variation

- It is also known as the relative standard deviation
- The coefficient of variation (CV) is a relative measure of variability that indicates the size of a standard deviation in relation to its mean

$$CV = \frac{\text{standard deviation}}{\text{mean}} \times 100$$

- Higher values indicate that the standard deviation is relatively large compared to the mean.
- It helps to compare two data sets on the basis of the degree of variation.

Eg: Two plants C and D of a factory show the following results about the number of workers and the wages paid to them. Using coefficient of variation formulas, find in which plant, C or D is there greater variability in individual wages.

No. of workers	5000	6000
Average monthly wages	$2500	$2500
Standard deviation	9	10

CV = $(\sigma/\mu) \times 100$, $\mu \neq 0$
CV = $(9/2500) \times 100$
CV = 0.36%
CV for plant D
CV = $(\sigma/\mu) \times 100$
CV = $(10/2500) \times 100$
CV = 0.4%
Plant C has CV = 0.36 and plant D has CV = 0.4
Hence plant D has greater variability in individual wages.

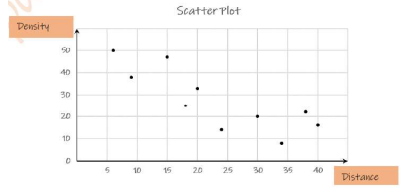
Statistics for IT

REGRESSION ANALYSIS

Week 12

Scatter diagram

- It is a mathematical diagram which uses cartesian coordinates to displaying values for two variables for a set of data



- Independent variable – independent variable is one which is not affected by the changes in other variables. It is plotted along the x axis
- Dependent variable – dependent variable is one whose values are determine by the values of the independent variable. It is plotted along y axis

Linear regression

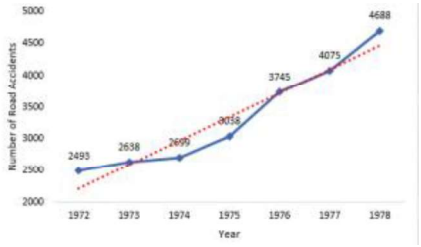
- Linear regression attempts to modal the relationship between two variables by fitting a line to observe data

Curve fitting

- Methods of curve fitting
  - Free hand method
  - Method of semi average
  - Moving average method
  - Least square method

Free hand method

- Draw a free hand smooth curve (or a straight line) through the points



Method of semi average

- The data is divided into two equal parts. In case of odd number of data, two equal parts can be made simply by omitting the middle value.
- The average of each part is calculated, thus we get two points.
- Each point is plotted at the mid-point of each half.
- Join the two points by a straight line.
- The straight line can be extended on either side.

- Eg: Fit a trend line by the method of semi-averages for the given data

Year	1990	1991	1992	1993	1994	1995	1996	1997
Sales	15	11	20	10	15	25	35	30

- Solution:

Year	Production	Average
1990	15	$\frac{15+11+20+10}{4}=14$
1991	11	
1992	20	
1993	10	
1994	15	
1995	25	
1996	35	$\frac{15+25+35+30}{4}=26.25$
1997	30	

Table 9.2

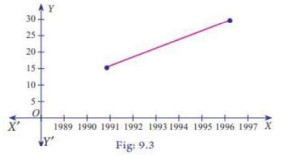


Fig: 9.3

Statistics for IT

REGRESSION ANALYSIS Contd...

Week 13

- Eg: from the data below calculate 3 month moving averages then plot the data and draw the trend line

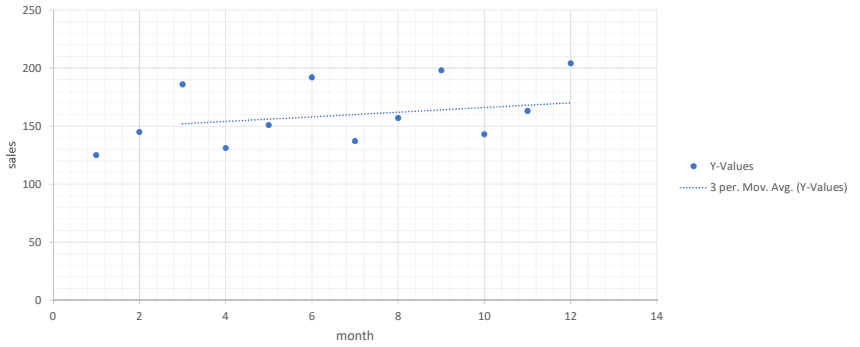
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sales $000	125	145	186	131	151	192	137	157	198	143	163	204

- solution

Month	Sales ($000)	Three-month moving total ($000)	Three-month moving average ($000)
January	125		
February	145	456=(125+145+186)	(456 ÷ 3) = 152
March	186	462=(145+186+131)	(462 ÷ 3) = 154
April	131	468=(186+131+151)	(468 ÷ 3) = 156
May	151	474	158
June	192	480	160
July	137	486	162
August	157	492	164
September	198	498	166
October	143	504	168
November	163	510	170
December	204		

Moving average method

- A moving average is a series of averages, calculated from historic data.
- Moving averages can be calculated for any number of time periods, for example a three-month moving average, a seven-day moving average, or a four-quarter moving average.



Least square method

- If the least squares regression line y on x is $y = a + bx$, the values of a and b are found by solving the simultaneous equations.

$$\Sigma\, y = na + b\Sigma x$$

$$\Sigma\, xy = a\Sigma\, x + b\Sigma\, x2$$

- The regression line can be used for estimation, prediction or forecasting.

- Eg: Using least square method, find the regression line for the following information.

x	y
2	4
3	6
7	10
8	12
10	13

- Solution

x	y	xy	x^2
2	4	8	4
3	6	18	9
7	10	70	49
8	12	96	64
10	13	130	100
30	45	322	226
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$

$\sum y = na + b\sum x \quad \Rightarrow 45 = 5a + 30b$

$\Rightarrow 5a + 30b = 45$-----(1)

$\sum xy = a\sum x + b\sum x^2 \quad \Rightarrow 322 = 30a + 226b$

$\Rightarrow 30a + 226b = 322$-----(2)

$a = \frac{51}{23}$ and $b = \frac{26}{23}$.

Therefore, the regression line is $y = a + bx$.

$y = \frac{51}{23} + \frac{26}{23}x$

Statistics for IT

APPLICATION OF STATISTICS

Week 14

introduction

- Statistics is indispensable for decision-making in various sectors. he goal of statistics is to gain understanding from the data. It has a wide range of applications in sectors such as
  - Health
  - Business and finance
  - Social science
  - Industry
  - IT
  - Banking & insurance
  - Environmental science

Statistical applications in health sector

- Epidemiology

  is the study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine.
- Clinical research

  Clinical research is a branch of healthcare science that determines the safety and effectiveness (efficacy) of medications, devices, diagnostic products and treatment regimens intended for human use.
- Quantitative psychology is the science of statistically explaining and changing mental processes and behaviors in humans.

Business & finance

- Business analytics

  is a rapidly developing business process that applies statistical methods to data sets (often very large) to develop new insights and understanding of business performance & opportunities
- Econometrics

  is a branch of economics that applies statistical methods to the empirical study of economic theories and relationships.
- Actuarial science

  is the discipline that applies mathematical and statistical methods to assess risk in the insurance and finance industries

Environmental science

- Environmental statistics

  is the application of statistical methods to environmental science. Weather, climate, air and water quality are included, as are studies of plant and animal populations.

Population ecology

is a sub-field of ecology that deals with the dynamics of species populations and how these populations interact with the environment.

Social science

- Social statistics

  is the use of statistical measurement systems to study human behavior in a social environment.

- Psychometrics

  is the theory and technique of educational and psychological measurement of knowledge, abilities, attitudes, and personality traits.

- Demography

  is the statistical study of all populations.