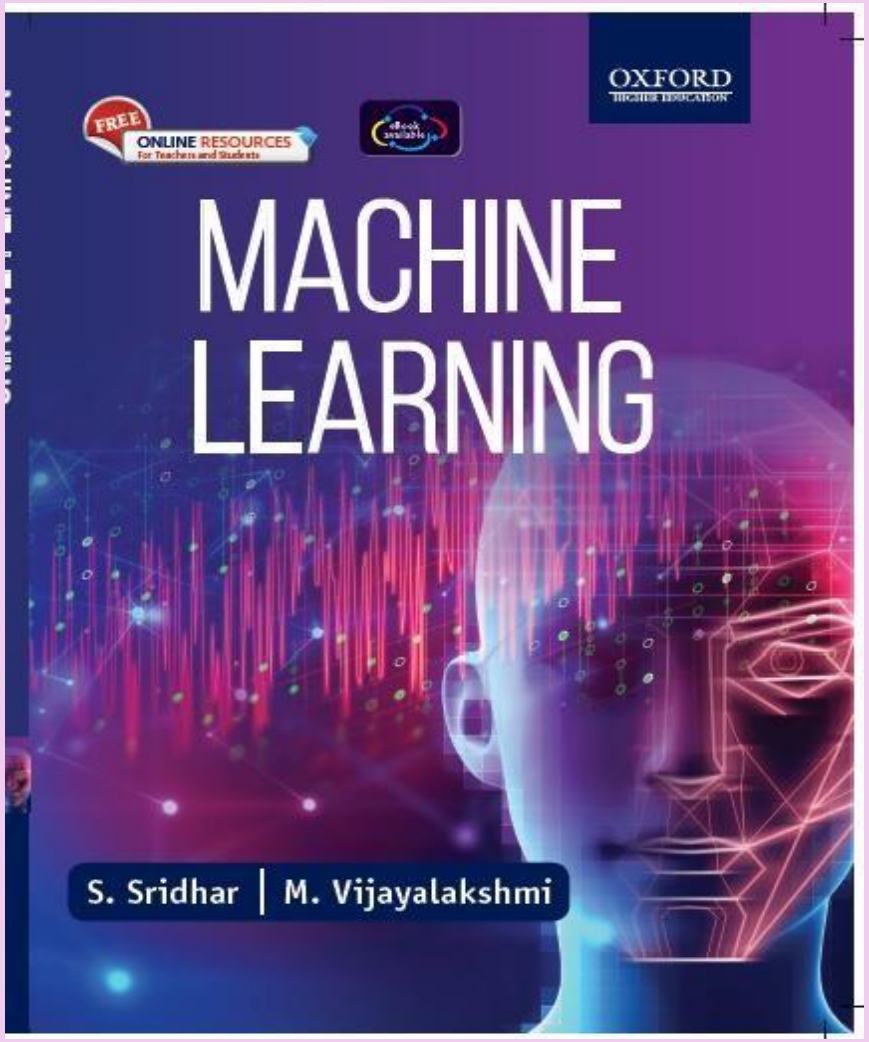


OXFORD
UNIVERSITY PRESS

Machine Learning

S. Sridhar and M. Vijayalakshmi

Chapter 2

Understanding of Data

What is Data?

- DATA ARE FACTS
- FACTS ARE IN THE FORM OF NUMBERS, AUDIO, VIDEO, AND IMAGE
- NEED TO ANALYZE DATA FOR TAKING DECISIONS.
- data sources like flat files, databases, or data warehouses.
- Operational data : normal business procedures and processes. For example, daily sales data
- Non-operational data used for decision making.

Elements /Characteristics of Big Data

1. small-scale computer is called ‘small data’: volume is less and can be stored and processed.
2. Big data, on the other hand, is a larger data whose volume is much larger than ‘small data’

Elements /Characteristics of Big Data

1. Volume –GB,TB,PB,HB(1 Million TB)
2. Velocity – The fast arrival speed of data and its increase in data volume is noted as velocity
3. Variety

- Form – There are many forms of data. Data types range from text, graph, audio, video, to maps.
- Function – various sources like human conversations, transaction records, and old archive data.
- Source of data –open/public data, social media data and multimodal data.

4. Veracity of data –conformity to the facts, truthfulness, believability, and confidence in data
5. Validity – Validity is the accuracy of the data for taking decisions or for any other goals that are needed by the given problem.
6. Value – Value is the characteristic of big data that indicates the value of the information that is extracted from the data and its influence on the decisions that are taken based on it.

The data quality of the numeric attributes is determined by factors like precision, bias, and accuracy

- **Precision**: closeness of repeated measurements. standard deviation is used to measure the precision.
- **Bias** : systematic result due to erroneous assumptions of the algorithms or procedures.
- **Accuracy** is the degree of measurement of errors to the true value of the quantity.

Data Sources

Types of Data

- STRUCTURED DATA
- SEMI-STRUCTURED DATA
- UNSTRUCTURED DATA

Structured Data

1. Record data:

- **The measurements in matrix. Rows- entities, cases, or records.**
- **The columns -attributes, features, or fields.**
- **Label is the term that is used to describe the individual observations.**

2.Graphics data

- **Relationships among objects. hyperlink**

3. Data matrix

- **It is a variation of the record type ,numeric attributes. Matrix operations can be applied on these data.**

4.Ordered data –

- **Temporal data** –with time.
- **Sequence data** – It is like sequential data but does not have time stamps. This data involves the sequence of words or letters.
- **Spatial data** – It has attributes such as positions or areas. For example, maps are spatial data where the points are related by location.

Unstructured Data

AN UNSTRUCTURED DATA CAN BE ANY ONE OF THE FOLLOWING –

- VIDEO, IMAGE, PROGRAMS
- BLOG DATA
- 80% OF ORGANIZATION DATA

Semi-Structured Data

A SEMI-STRUCTURED DATA CAN BE ANY ONE OF THE FOLLOWING –

- XML/JSON OBJECTS
- RSS FEEDS
- HIERARCHICAL RECORDS

Data Storage

❖Flat Files

I. Simplest and most commonly available data source.
II. Data is stored in plain ASCII or EBCDIC
III. Minor changes of data in flat files affect the results of the data mining algorithms.
IV. Suitable only for storing small dataset and not desirable if the dataset becomes larger.

• **CSV files – CSV stands for comma-separated value files** where the values are separated by commas. These are used by spreadsheet and database applications. The first row may have attributes and the rest of the rows represent the data.

• **TSV files – TSV stands for Tab separated values files** where values are separated by Tab. Both CSV and TSV files are generic in nature and can be shared. There are many tools like Google Sheets and Microsoft Excel to process these files.

Data Storage

❖ **Database System**

I. Database files and a database management system
II. original data and metadata.
III. Database administrator, query processing, and transaction manager.

1.**A transactional database:** Each record is atransaction
2. **Time-series database :** log files
3.**Spatial databases:** vector(points,lines)

Data Storage

- **OTHER TYPES**

World Wide Web (WWW) It provides a diverse, worldwide online information source. The objective of data mining algorithms is to mine interesting patterns of information present in WWW.

XML (eXtensible Markup Language) It is both human and machine interpretable data format that can be used to represent data that needs to be shared across the platforms.

Data Stream It is dynamic data, which flows in and out of the observing environment. Typical characteristics of data stream are huge volume of data, dynamic, fixed order movement, and real-time constraints.

RSS (Really Simple Syndication) It is a format for sharing instant feeds across services.

JSON (JavaScript Object Notation) It is another useful data interchange format that is often used for many machine learning algorithms.

BIG DATA ANALYTICS AND TYPES OF ANALYTICS

Data analytics refers to the process of data collection, preprocessing and analysis.

1. Descriptive analytics:

- Describing the main features of the data.

2. Diagnostic analytics:

- question – 'Why?'.(casual analysis)
- the cause and effect of the events.

3. Predictive analytics:

- What will happen in future given this data?

4. Prescriptive analytics

- finding the best course of action for the business organizations. Prescriptive analytics goes beyond prediction and helps in decision making by giving a set of actions

BIG DATA ANALYSIS FRAMEWORK

Is a layered architecture

1. Date connection layer
2. Data management layer
3. Data analytics later
4. Presentation layer

Big Data processing cycle

Big Data processing cycle involves data management that consists of the following steps.

1. Data collection
2. Data preprocessing
3. Applications of machine learning algorithm
4. Interpretation of results and visualization of machine learning algorithm

1.Data Collection

Time is spent for collection of good quality data. ‘Good data’ is one that has the following properties:

1. **Timeliness** –
2. **Relevancy –**.
3. **Knowledge about the data** –

1.Data Collection

open/public data, social media data and multimodal data.

1. **Open or public data source** –
2. **Social media –**
3. **Multimodal data -**

2.Data Preprocessing

In real world, the available data is 'dirty'.

- Incomplete data • Inaccurate data
- Outlier data • Data with missing values
- Data with inconsistent values • Duplicate data

✓ Data preprocessing improves the quality of the data mining techniques.

✓ Raw data must be preprocessed to give accurate results.

✓ The process of detection and removal of errors in data → data cleaning.

✓ Making the data processable for ML algorithms→ Data wrangling

Table 2.1: Illustration of 'Bad' Data

Patient ID	Name	Age	Date of Birth (DoB)	Fever	Salary
1.	John	21		Low	–1500
2.	Andre	36		High	Yes
3.	David	5	10/10/1980	Low	" "
4.	Raju	136		High	Yes

- Salary = ' ' is **incomplete data**.
- DoB of patients, John, Andre, and Raju, is the **missing data.**
- The age of David is '5' but his DoB - 10/10/1980 -→ **inconsistent data.**

Outliers :characteristics that are different from other data and have very unusual values. It might be a typographical error. It is often required to distinguish between noise and outlier data.

Missing Data Analysis

1. **Ignore the tuple**
2. **Fill in the values manually**
3. **A global constant can be used to fill in the missing attributes:'** Unknown' or be 'Infinity'.
4. **The attribute value may be filled by the attribute value.**
5. **Use the attribute mean for all samples belonging to the same class.**
6. **Use the most possible value to fill in the missing value.**

Removal of Noisy or Outlier Data

1. Noise is a random error or variance in a measured value.
2. Removed using binning -the given data values are sorted and distributed into equal frequency bins. The bins are also called as buckets.
3. The binning method then uses the neighbor values to smooth the noisy data.

- **'smoothing by means'** mean of the bin removes the values of the bins
- '**smoothing by bin medians**' where the bin median replaces the bin values
- **'smoothing by bin boundaries**' the closest bin boundary.
- **The maximum and minimum values** are called bin boundaries.

Example

Example 2.1: Consider the following set: S = {12, 14, 19, 22, 24, 26, 28, 31, 34}. Apply various binning techniques and show the result.

Solution: By equal-frequency bin method, the data should be distributed across bins. Let us assume the bins of size 3, then the above data is distributed across the bins as shown below:

Bin 1 : 12 , 14, 19

Bin 2 : 22, 24, 26

Bin 3 : 28, 31, 32

By smoothing bins method, the bins are replaced by the bin means.

Bin 1 : 15, 15, 15

Bin 2 : 24, 24, 24

Bin 3 : 30.3, 30.3, 30.3

Using smoothing by bin boundaries method, the bins' values would be like:

Bin 1 : 12, 12, 19

Bin 2 : 22, 22, 26

Bin 3 : 28, 32, 32

Data Integration and Data Transformations

1. Merge data from multiple sources into a single data source. Lead to redundant data.
2. Goal of data integration is to detect and remove redundancies that arise from integration.
3. Normalization, the attribute values are scaled to fit in a range (say 0-1) to improve the performance of the data mining algorithm.
4. **Min-Max**
5. **z-Score**

Min-Max Procedure

1. Each variable V is normalized by its difference with the minimum value divided by the range to a new range, say 0–1.

$$min\text{-}max = \frac{V - min}{max - min} \times (new\ max - new\ min) + new\ min \qquad (2.1)$$

Min and max are the minimum and maximum of the given data, new max and new min are the minimum and maximum of the target range, say 0 and 1.

Consider the set: V = {88, 90, 92, 94}. Apply Min-Max procedure and map the marks to a new range 0–1.

Solution: The minimum of the list V is 88 and maximum is 94. The new min and new max are 0 and 1, respectively.

For marks 88,

$$min\text{-}max = \frac{88 - 88}{94 - 88} \times (1 - 0) + 0 = 0$$

Similarly, other marks can be computed as follows:

For marks 90,

$$min\text{-}max = \frac{90 - 88}{94 - 88} \times (1 - 0) + 0 = 0.33$$

For marks 92,

$$min\text{-}max = \frac{92 - 88}{94 - 88} \times (1 - 0) + 0 = \frac{4}{6} = 0.66$$

For marks 94,

$$min\text{-}max = \frac{94 - 88}{94 - 88} \times (1 - 0) + 0 = \frac{6}{6} = 1$$

Marks {88, 90, 92, 94} are mapped to the new range {0, 0.33, 0.66, 1}. Thus, the Min-Max normalization range is between 0 and 1.

z-Score Normalization

Difference between the field value and mean value, and by scaling this difference by standard deviation of the attribute.

$$V* = V - \mu/\sigma$$

Here, s is the standard deviation of the list V and m is the mean of the list V.

Example 2.3: Consider the mark list V = {10, 20, 30}, convert the marks to z-score

Solution: The mean and Sample Standard deviation (s) values of the list V are 20 and 10, respectively.

$$\text{z-score of } 10 = \frac{10-20}{10} = -\frac{10}{10} = -1$$

$$\text{z-score of } 20 = \frac{20-20}{10} = \frac{0}{10} = 0$$

$$\text{z-score of } 30 = \frac{30-20}{10} = \frac{10}{10} = 1$$

Hence, the z-score of the marks 10, 20, 30 are -1, 0 and 1, respectively.

2.4 DESCRIPTIVE STATISTICS

1. It is used to summarize and describe data.

Dataset and Data Types

1. A dataset can be assumed to be a collection of data objects.
2. The data objects may be records, points, vectors, patterns, events, cases, samples or observations.
3. These records contain many attributes.
4. An attribute can be defined as the property or characteristics of an object.

2.4 DESCRIPTIVE STATISTICS

1. It is used to summarize and describe data.

Dataset and Data Types

1. Every attribute should be associated with a value.
2. This process is called measurement.
3. The type of attribute determines the data types, often referred to as measurement scale types.

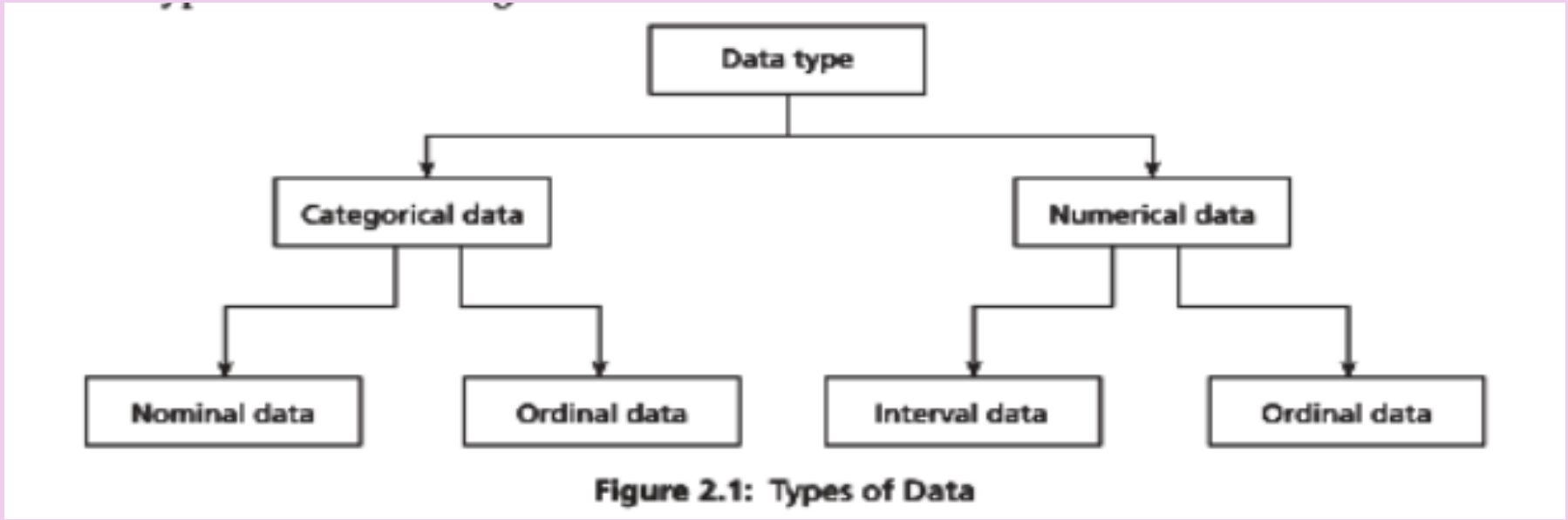


Figure 2.1: Types of Data

1. Categorical or qualitative data
2. Numerical or quantitative data

Table 2.2: Sample Patient Table

Patient ID	Name	Age	Blood Test	Fever	Disease
1.	John	21	Negative	Low	No
2.	Andre	36	Positive	High	Yes

Categorical or Qualitative Data

- **Nominal Data** –patient ID.

1. Data that can be categorized does not have numerical value.
2. Nominal data type provides only information but has no ordering among data.
3. Only operations like (=, ≠) are meaningful for these data.

- **Ordinal Data** –

1. It provides enough information and has natural order.
2. Fever = {Low, Medium, High} is an ordinal data. Certainly, low is less than medium and medium is less than high, irrespective of the value.
3. Any transformation can be applied to these data to get a new value.

Numeric or Qualitative Data

- **Interval Data –**

1. Interval data is a numeric data for which the differences between values are meaningful.
2. For example, there is a difference between 30 degree and 40 degree.
3. Only the permissible operations are + and -.

- **Ratio Data –**

1. For ratio data, both differences and ratio are meaningful.

Another way of classifying the data is to classify it as:

1.Discrete Data This kind of data is recorded as integers.

2.Continuous Data It can be fitted into a range and includes decimal point. For example, age is a continuous data. Though age appears to be discrete data, one may be 12.5 years old and it makes sense. Patient height and weight are all continuous data.

Third way of classifying the data is based on the number of variables used in the dataset.

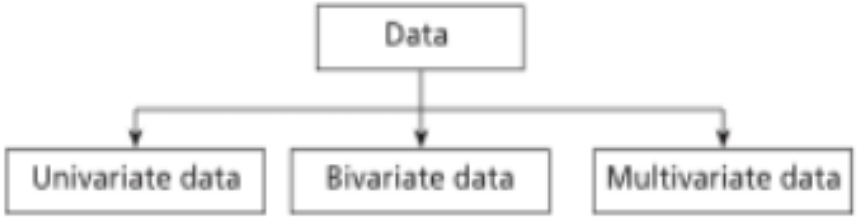


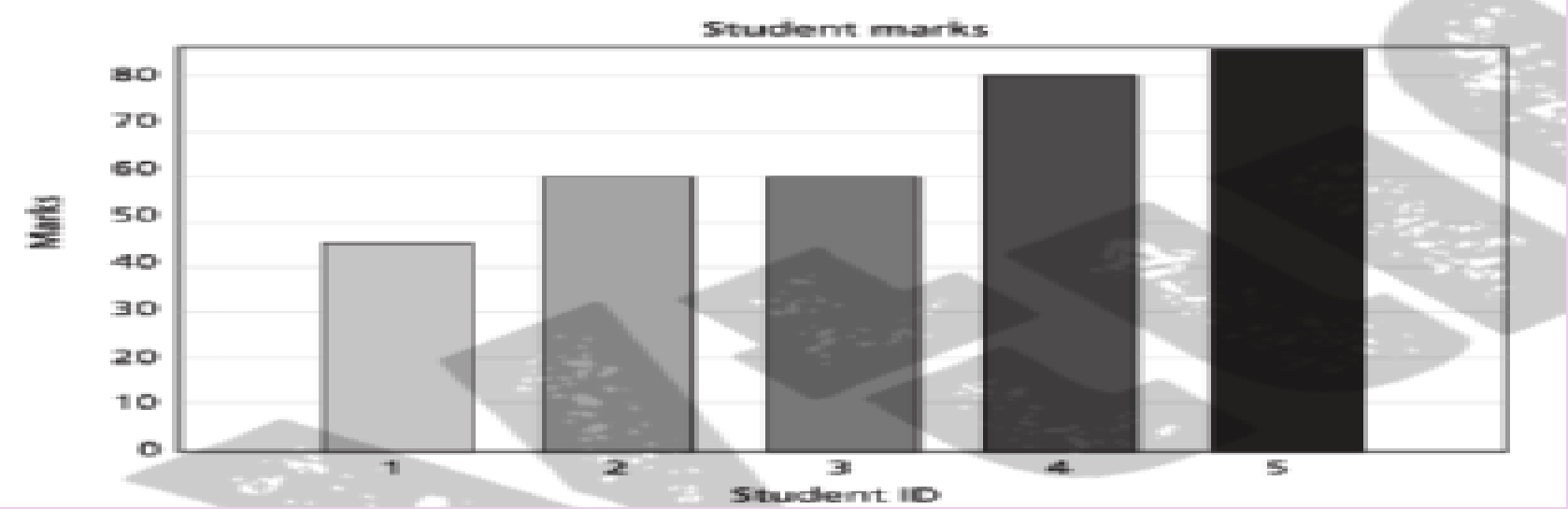
Figure 2.2: Types of Data Based on Variables

UNIVARIATE DATA ANALYSIS AND VISUALIZATION

1. The dataset has only one variable. A variable can be called as a category.
2. Univariate does not deal with cause or relationships.
3. The aim of univariate analysis is to describe data and find patterns.
4. It involves finding the frequency distributions, central tendency measures, dispersion or variation, and shape of the data.

2.5.1 Data Visualization

Bar Chart A Bar chart (or Bar graph) is used to display the frequency distribution for variables. Bar charts are used to illustrate discrete data. The charts can also help to explain the counts of nominal data. It also helps in comparing the frequency of different groups. The bar chart for students' marks {45, 60, 60, 80, 85} with Student ID = {1, 2, 3, 4, 5} is shown below



Pie Chart These are equally helpful in illustrating the univariate data. The percentage frequency distribution of students' marks {22, 22, 40, 40, 70, 70, 70, 85, 90, 90} is below.

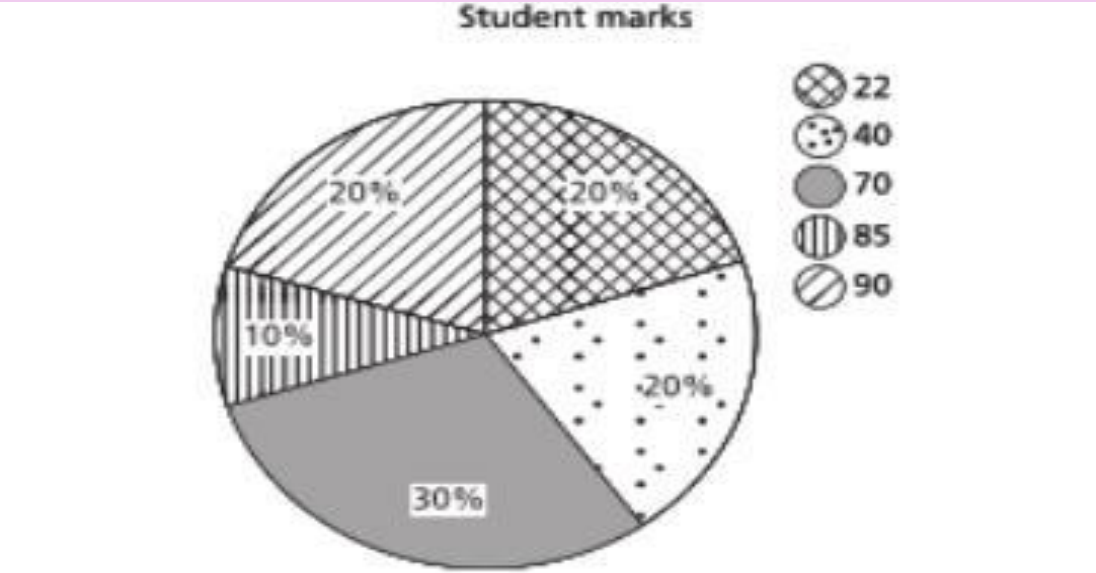


Figure 2.4: Pie Chart

It can be observed that the number of students with 22 marks are 2. The total number of students are 10. So, 2/10 × 100 = 20% space in a pie of 100% is allotted for marks 22 .

Histogram showing frequency distributions. The histogram for students' marks {45, 60, 60, 80, 85} in the group range of 0-25, 26-50, 51-75, 76-100 is given below in Figure 2.5. One can visually inspect from Figure 2.5 that the number of students in the range 76-100 is 2.

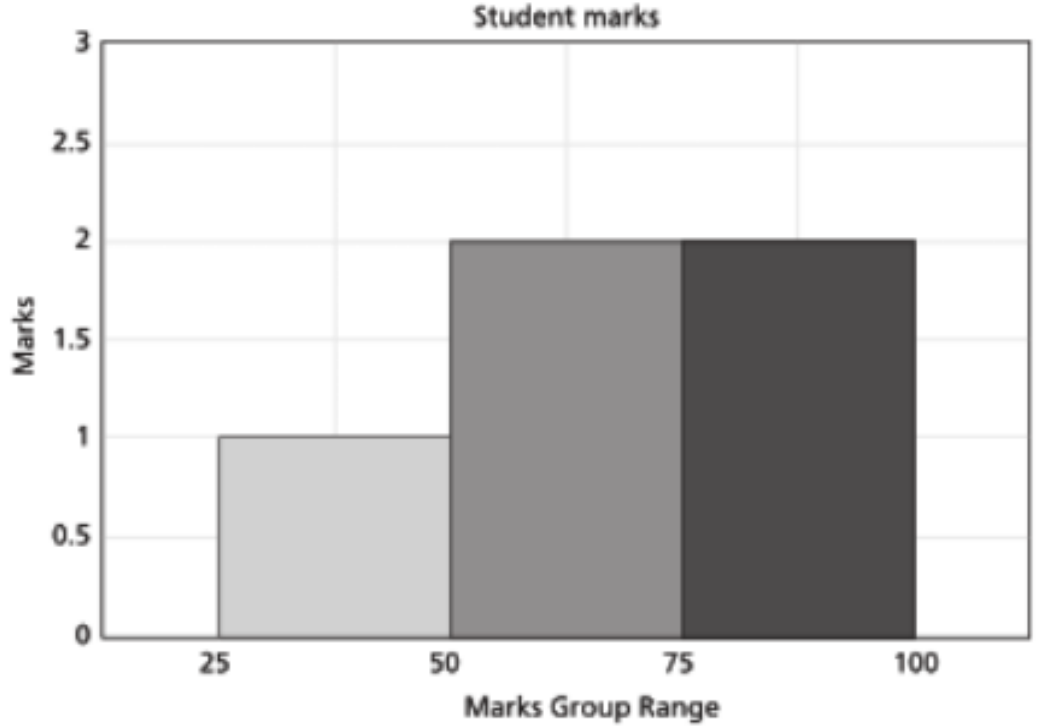


Figure 2.5: Sample Histogram of English Marks

Dot Plots less clustered as compared to bar charts. Dot plot of English marks for five students with ID as {1, 2, 3, 4, 5} and marks {45, 60, 60, 80, 85} is given. The advantage is that by visual inspection one can find out who got more marks.

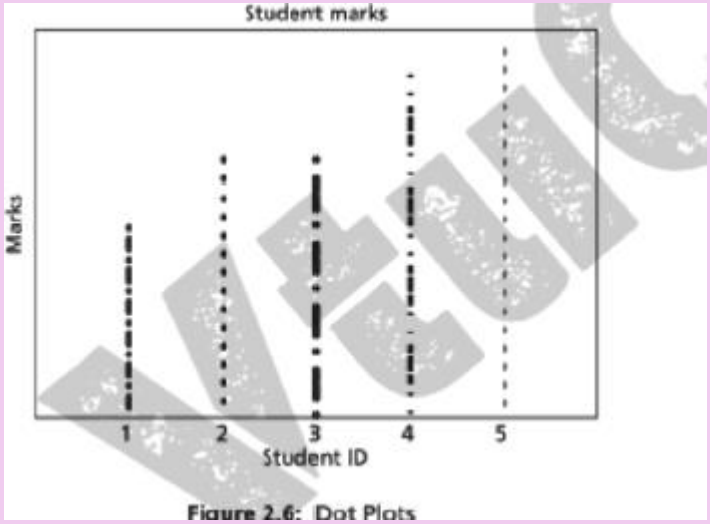


Figure 2.6: Dot Plots

2.5.2 Central Tendency

1. Mean – Arithmetic average (or mean) is a measure of central tendency that represents the 'center' of the dataset. Mathematically, the average of all the values in the sample (population) is denoted as x. Let x1, x2, … , xN be a set of 'N' values or observations, then the arithmetic mean is given as:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (2.3)$$

- Weighted mean –weighted mean gives different importance to all items as the item importance varies.

Geometric mean – Let x1, x2, … , xN be a set of 'N' values or observations. Geometric mean is the Nth root of the product of N items. The formula for computing geometric mean is given as follows:

$$\text{Geometric mean} = \left(\prod_{i=1}^{n} x_i\right)^{\overline{N}} = \sqrt[N]{x_1 \times x_2 \times \cdots \times x_N} \qquad (2.4)$$

Here, n is the number of items and xi are values. In larger cases, computing geometric mean is difficult. Hence, it is usually calculated as:

$$\text{Anti-log of } \frac{\log(x_1) + \log(x_2) + \cdots + \log(x_N)}{N} \qquad (2.5)$$

$$= \text{anti-log } \frac{\sum_{i=1}^{n} \log(x_i)}{N} \qquad (2.6)$$

2. Median – The middle value in the distribution is called median. If the total number of items in the distribution is odd, then the middle value is called median. A median class is that class where $(N/2)^{th}$ item is present. In the continuous case, the median is given by the formula

$$\text{Median} = L_1 + \frac{\frac{N}{2} - cf}{f} \times i \qquad (2.7)$$

Median class is that class where N/2th item is present. Here, i is the class interval of the median class and L1 is the lower limit of median class, f is the frequency of the median class, and cf is the cumulative frequency of all classes preceding median.

3. Mode – Mode is the value that occurs more frequently in the dataset. the value that has the highest frequency is called mode.

2.5.3 Dispersion

The spreadout of a dataset around the central tendency (mean, median or mode) is called dispersion. Dispersion is represented by various ways such as range, variance, standard deviation, and standard error. These are second order measures.

1.Range -Difference between the maximum and minimum of values of the given list of data.

2.Standard Deviation Standard deviation is the average distance from the mean of the dataset to each point.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} \qquad (2.8)$$

The formula for sample standard deviation is given by: Here, N is the size of the population, xi is observation or value from the population and m is the population mean. Often, N – 1 is used instead of N in the denominator of Eq. (2.8).

Quartiles and Inter Quartile Range

1. Percentiles are about data that are less than the coordinates by some percentage of the total value.

2.kth percentile is the property that the k% of the data lies at or below Xi.

3.For example, median is 50th percentile and can be denoted as Q0.50. The 25th percentile is called first quartile (Q1) and the 75th percentile is called third quartile (Q3).

4. Inter Quartile Range (IQR) is the difference between Q3 and Q1.

5.Outliers are normally the values falling apart at least by the amount 1.5 × IQR above the third quartile or below the first quartile.

Interquartile is defined by Q0.75 – Q0.25.

Example 2.4: For patients' age list {12, 14, 19, 22, 24, 26, 28, 31, 34}, find the IQR.

Solution: The median is in the fifth position. In this case, 24 is the median. The first quartile is median of the scores below the mean i.e., {12, 14, 19, 22}. Hence, it's the median of the list below 24. In this case, the median is the average of the second and third values, that is, Q0.25 = 16.5.

Similarly, the third quartile is the median of the values above the median, that is {26, 28, 31, 34}.

So, Q0.75 is the average of the seventh and eighth score. In this case, it is 28 + 31/2 = 59/2 = 29.5.

Hence, the IQR using Eq. (2.10) is:

= Q0.75 – Q0.25

= 29.5-16.5 = 13

Five-point Summary and Box Plots The median, quartiles Q1 and Q3, and minimum and maximum written in the order < Minimum, Q1, Median, Q3, Maximum > is known as five-point summary

Example 2.5: Find the 5-point summary of the list {13, 11, 2, 3, 4, 8, 9}.
Solution: The minimum is 2 and the maximum is 13. The Q1, Q2 and Q3 are 3, 8 and 11, respectively. Hence, 5-point summary is {2, 3, 8, 11, 13}, that is, {minimum, Q1, median, Q3, maximum}. Box plots are useful for describing 5-point summary. The Box plot for the set is given in Figure 2.7.

Figure 2.7: Box Plot for English Marks

2.5.4 Shape

Skewness and Kurtosis (called moments) indicate the symmetry/asymmetry and peak location of the dataset.

Skewness

The measures of direction and degree of symmetry are called measures of third order. skewness should be zero as in ideal normal distribution. More often, the given dataset may not have perfect symmetry (consider the following Figure 2.8).

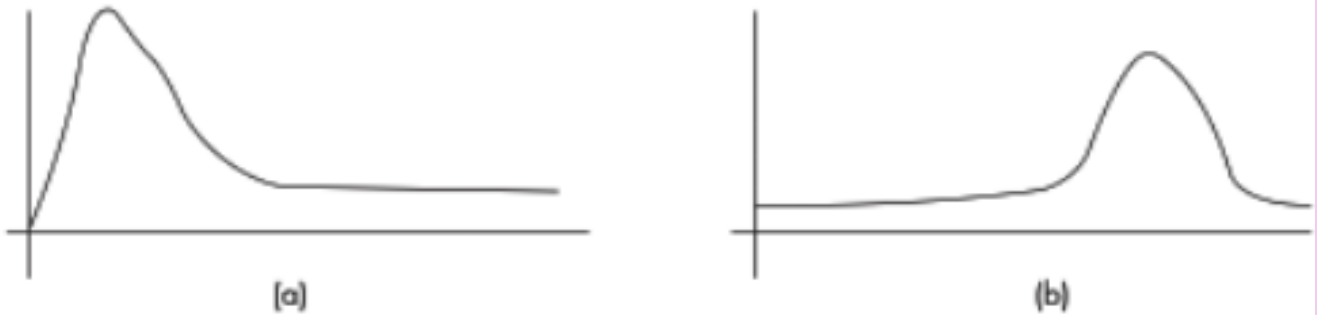


Figure 2.8: (a) Positive Skewed and (b) Negative Skewed Data

The dataset may also either have very high values or extremely low values. If the dataset has far higher values, then it is said to be skewed to the right. On the other hand, if the dataset has far more low values then it is said to be skewed towards left. If the tail is longer on the left-hand side and hump on the right-hand side, it is called positive skew. Otherwise, it is called negative skew.

Generally, for negatively skewed distribution, the median is more than the mean. The relationship between skew and the relative size of the mean and median can be summarized by a convenient numerical skew index known as Pearson 2 skewness coefficient.

$$\frac{3 \times (\mu - median)}{\sigma} \tag{2.12}$$

Also, the following measure is more commonly used to measure skewness. Let $X_1, X_2, \cdots, X_N$ be a set of 'N' values or observations then the skewness can be given as:

$$\frac{1}{N} \times \sum_{i=1}^{N} \frac{(x_i - \mu)^3}{\sigma^3} \tag{2.13}$$

Here, μ is the population mean and σ is the population standard deviation of the univariate data. Sometimes, for bias correction instead of N, $N - 1$ is used.

Kurtosis

Kurtosis also indicates the peaks of data. If the data is high peak, then it indicates higher kurtosis and vice versa.

Kurtosis is the measure of whether the data is heavy tailed or light tailed relative to normal distribution. It can be observed that normal distribution has bell-shaped curve with no long tails. Low kurtosis tends to have light tails. The implication is that there is no outlier data. Let $x_1, x_2, \cdots, x_N$ be a set of 'N' values or observations. Then, kurtosis is measured using the formula given below:

$$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^4 / N}{\sigma^4} \tag{2.14}$$

It can be observed that $N - 1$ is used instead of N in the numerator of Eq. (2.14) for bias correction. Here, $\bar{x}$ and σ are the mean and standard deviation of the univariate data, respectively.

MAD and CV

Mean Absolute Deviation (MAD)

MAD is another dispersion measure and is robust to outliers. Normally, the outlier point is detected by computing the deviation from median and by dividing it by MAD. Here, the absolute deviation between the data and mean is taken. Thus, the absolute deviation is given as:

$$|x - \mu| \tag{2.15}$$

The sum of the absolute deviations is given as $\Sigma|x - \mu|$

Therefore, the mean absolute deviation is given as: $\frac{\Sigma|x - \mu|}{N}$ (2.16)

Coefficient of Variation (CV)

Coefficient of variation is used to compare datasets with different units. CV is the ratio of standard deviation and mean, and %CV is the percentage of coefficient of variations.