1. **Analytics**

Page no. 1:

**Analytics** is the careful study and examination of data to find useful patterns, insights, and trends. These insights help people make better decisions and improve how things are done in business, science, and many other areas. It uses tools from statistics, math, and computer science to turn data into helpful information.

**Understanding Data**

Analytics starts by getting to know the data. This means looking at all the information we have. Some of this data is **structured**, like spreadsheets or databases. Other data is **unstructured**, such as written text, pictures, or videos. The data can come from many places like sales, customer chats, machines, websites, or social media.

**Data Preparation**

Before we can study the data, we need to **clean it up** and **get it ready**. This step includes fixing messy data, removing repeated information, filling in missing parts, and making sure everything is in the same format. Sometimes, we also have to combine data from different sources into one place so it's easier to work with.

**Descriptive Analytics: (Describe, Summarize, Charts, relationships between variables,**

**Helps stakeholders to know what happened in past , helpful for further analysis)**

**One Liner Answer :** Describing the data

Descriptive analytics is about looking at past data to understand what has happened. It includes creating summary statistics, using charts and graphs to show the data, and checking how different things are related. This type of analysis helps people understand past events and gives background for deeper analysis.

**Diagnostic Analytics: (** Why this happened, identify patterns, hypothsis testing, corelation analysis,

casual inference testing (estimation/ conclusion)

Diagnostic analytics tries to explain why something happened. It looks for patterns and relationships in the data to find the reasons behind certain results or trends. This kind of analysis often uses tools like hypothesis testing, checking for correlations, and methods to find cause-and-effect links.

**Predictive Analytics: (predictions : forecasting, churn prediction, risk assessments, demands forecasting)**

**One line: in this we study about future predictions based on past data.**

Predictive analytics uses past data to guess what might happen in the future. It involves building models using statistics or machine learning to find patterns in the data. These models can help with things like predicting sales, customer behavior, risk levels, or future demand.

**Prescriptive Analytics: (recommending what to do further)**

Prescriptive analytics takes it a step further by suggesting what actions to take. It uses tools like optimization and simulation to test different scenarios and figure out the best solution, based on specific goals or limits. This helps businesses make smart decisions, reduce risks, and get better results.

**Continuous Improvement:**  
Analytics is not a one-time task, it’s an iterative process that needs regular updates and improvements. As new data comes in and business situations change, analytics models and strategies should be adjusted to stay accurate and useful. Organizations should create feedback loops to use insights from analytics in their decision-making and keep improving over time.

Page no .50

1. **Descriptive Statistical measures**

They give **brief information** and they **summarize** given **dataset.**

**Descriptive Statistics:** helps describing the data, gives short summary of data

1. Measures of central tendency
2. Measures of variability

**Mean, median, mode :** measures of central tendency ( summary statistics )

1. **Measures of Variability :**

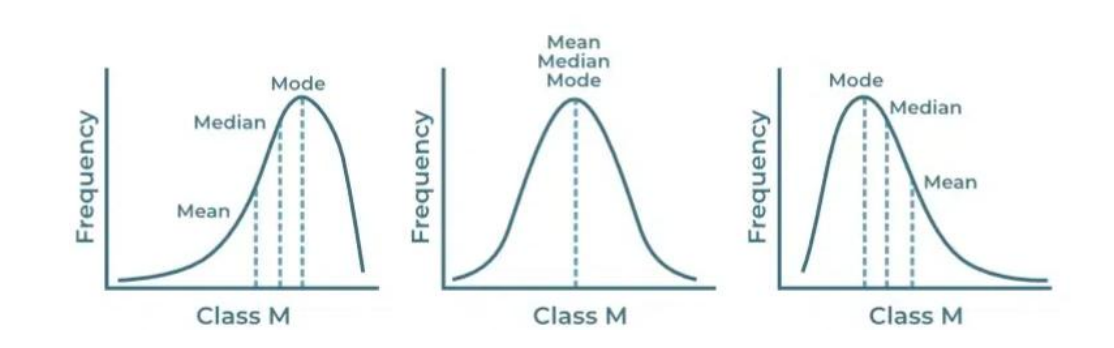
**A:** 20,50,80,10,25

**B:** 100,0,100,10,0

A is consistent, this is measure of variability ( how much spread the data is)

1. **Measures of central tendency :**

Central Tendency: one values which is present at center of data (average/middle value)

****

**Asymmetric data (left skewed) symmetric data asymmetric data (right skewed)**

**Types of Averages:**

1) Arithmetic Mean (A.M.)

2) Weighted Arithmetic Mean

3) Median

4) Mode

5) Geometric Mean (G.M.)

6) Harmonic Mean (H.M.)

**Requisites of a good average:**

1) Easy to understand

2) Simple to compute

3) Based on all the items.

4) Not unduly affected by extreme observations. ( outliers )

5) Rigidly defined. (Well defined format) Rigid: fixed

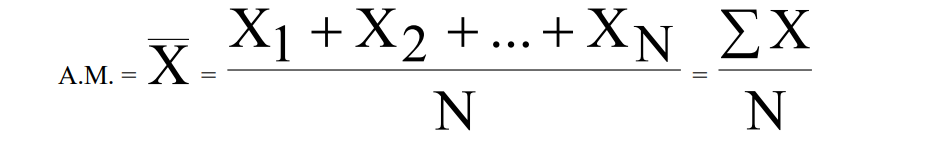
6) Capable for further algebraic treatment.

7) Easy to interpret

**Arithmetic Mean(A.M.):**

**Arithmetic mean for Raw Data:**

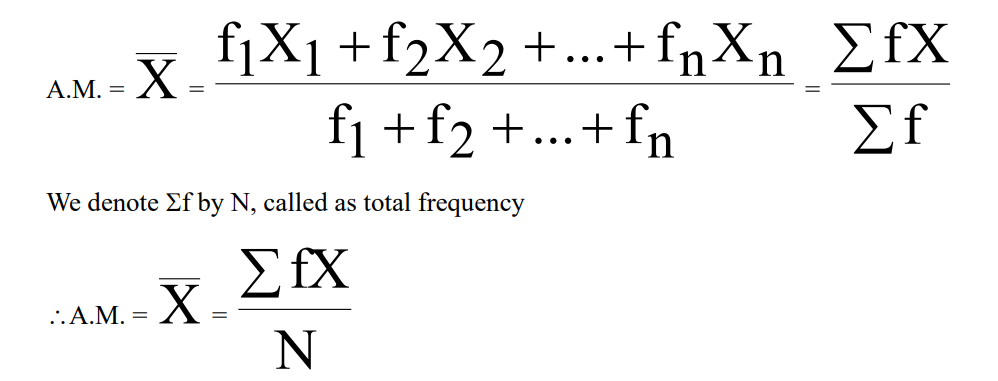
The A.M. of N observations X1, X2, …, XN is denoted by **X**

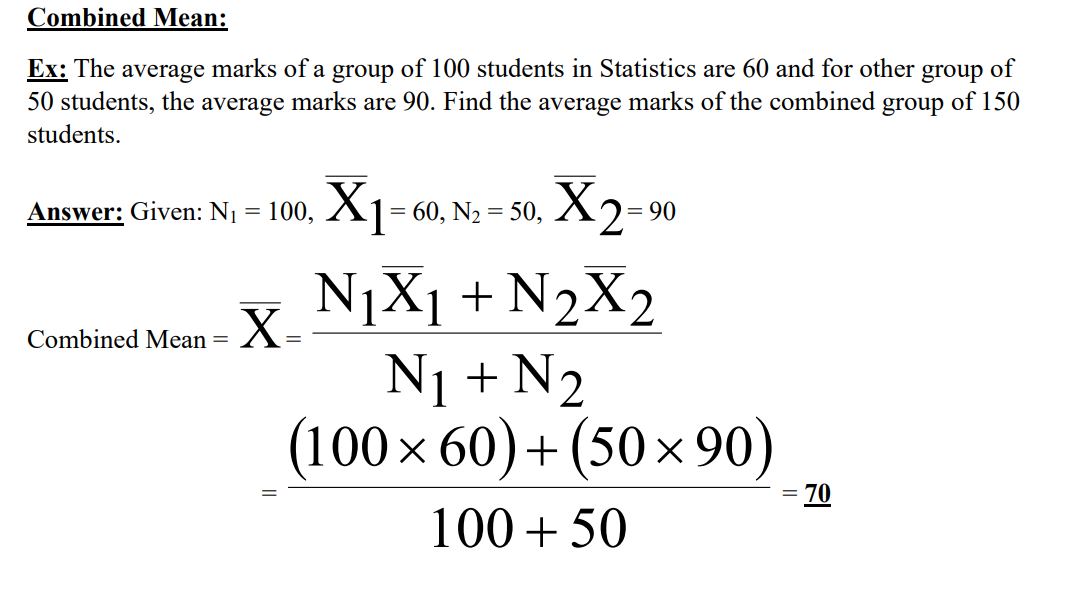
****

**Arithmetic Mean for a frequency distribution:**

Consider a data of n observations X1, X2, …, Xn occurring with respective frequencies

**f1, f2, …, fn. Then the A.M. is denoted by**

****

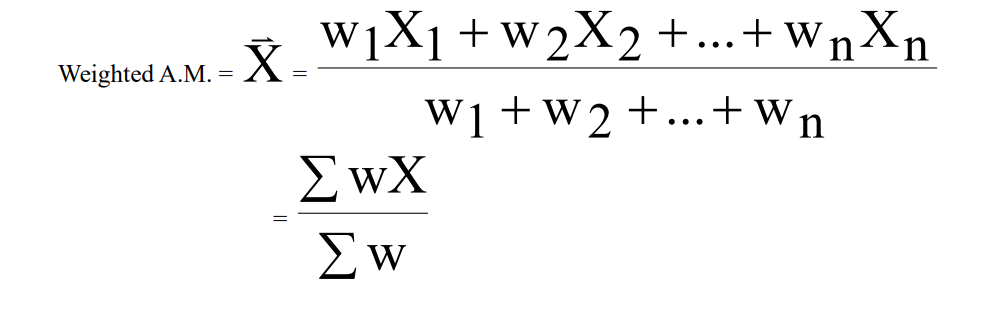


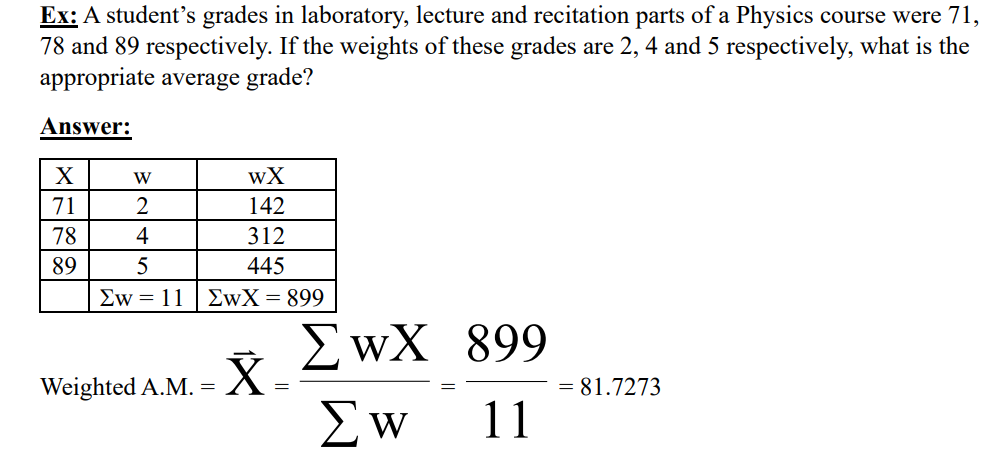
**Weighted Arithmetic Mean:**

weight => importance

Sometimes we associate certain weighing factor (or weights) with the numbers X1, X2, …, Xn.

Suppose the weights are w1, w2, …, wn. The weighted arithmetic mean is denoted By X





**Merits**:

1. It is rigidly defined.

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

3.It is based on all the observations. (considers all values, if one value is changed mean can change)

4.It is capable of further algebraic treatment.

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

average.

**Demerits**:

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

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

values.

3. It is affected by extreme values

4.It cannot be calculated for frequency distribution having open end class- intervals e.g.

class-intervals like below 10 or above 50 etc.

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

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

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

Even though A.M. has various demerits, it is considered to be the best all averages as it

satisfies most of the requisites of a good average. A.M. is called the Ideal Average.

Page no.2

**Data analytics Life Cycle:**

* A structured approach to extracting insights and value from data.
* process from defining the problem to implementing solutions.

1. **Problem Definition:**

• Identify the business problem or opportunity that analytics can address.

• Define clear objectives and key performance indicators (KPIs) to measure success.

• Ensure alignment with organizational goals and stakeholder needs.

1. **Data Collection:**

• Identify relevant data sources both internal and external to the organization.

• Gather data from databases, spreadsheets, files, APIs, sensors, social media, etc.

• Ensure data quality, completeness, and relevance for analysis.

1. **Data Preparation:**

• Cleanse the data by removing duplicates, correcting errors, and handling missing or

inconsistent values.

• Transform the data into a suitable format for analysis (e.g., normalization,

aggregation, or feature engineering).

• Integrate data from multiple sources if necessary

1. **Exploratory Data Analysis (EDA):**

• Explore the dataset to understand its structure, distribution, and relationships.

• Visualize data using charts, graphs, and statistical summaries.

• Identify patterns, trends, outliers, and potential insights.

1. **Feature Engineering:**

• Select, create, or transform features that are relevant and predictive for the analysis.

• Apply techniques such as dimensionality reduction, encoding categorical variables, or

deriving new features.

1. **Modelling:**

• Select appropriate analytical techniques or algorithms based on the problem and data

characteristics.

• Split the data into training, validation, and testing sets.

• Train machine learning or statistical models using the training data.

• Tune hyperparameters and evaluate model performance using validation data.

• Validate the model's performance on unseen data using the testing set.

1. **Interpretation and Evaluation:**

• Interpret the model results in the context of the problem and business objectives.

• Evaluate the model's performance using relevant metrics (e.g., accuracy, precision,

recall, or AUC).

• Assess the impact of the analytics solution on the business problem and its alignment

with KPIs.

1. **Deployment:**

• Deploy the analytics solution into production or operational systems.

• Integrate the model into decision-making processes or business workflows.

• Monitor the model's performance in real-world scenarios and collect feedback for

continuous improvement.

1. **Monitoring and Maintenance:**

• Establish monitoring mechanisms to track the performance and behavior of the

deployed model.

• Monitor data quality, model drift, and other relevant metrics over time.

• Retrain or update the model periodically with new data to ensure relevance and

accuracy.

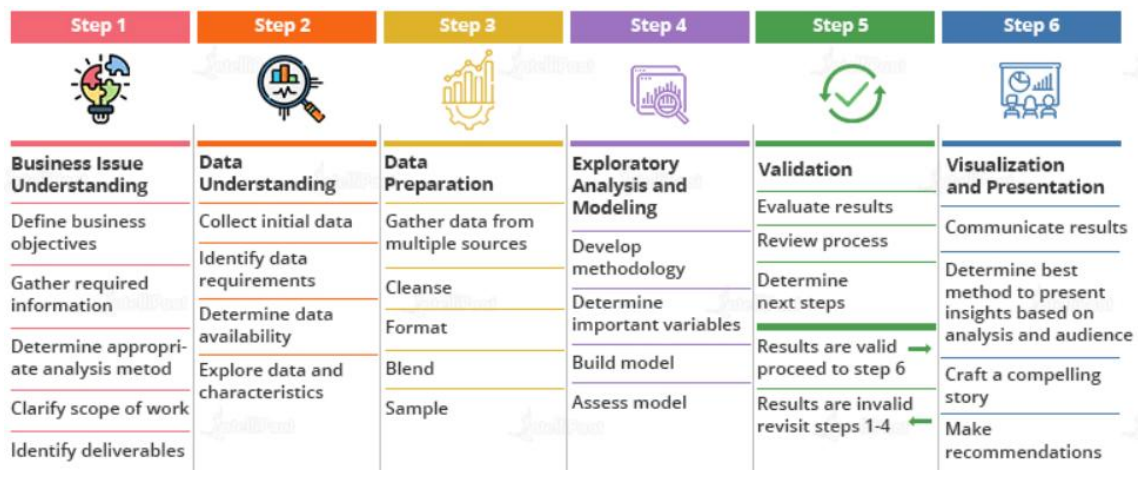
1. **Iterative Improvement:**

• Continuously refine and improve the analytics solution based on feedback, changing

business requirements, and new data.

• Iterate through the lifecycle stages as needed to address evolving challenges and

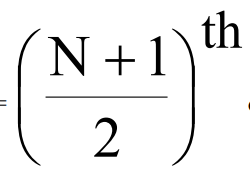
opportunities.

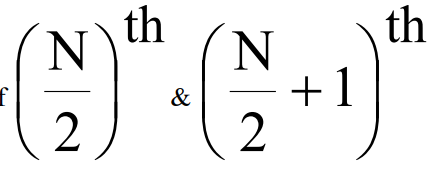


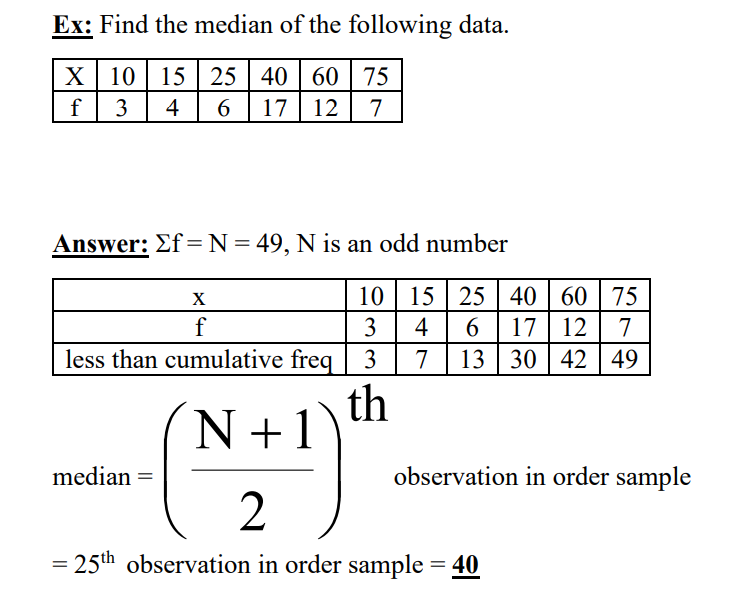
Page no.56

**Median:**

* Median is central value of dataset.
* Dataset should be sorted
* Can be 1 value from dataset or can be mean of 2 values from dataset
* N means no. of values

If N is odd then **median** will be :  **term in dataset.**

If N is even then **median** will be **mean of**  **term in dataset.**



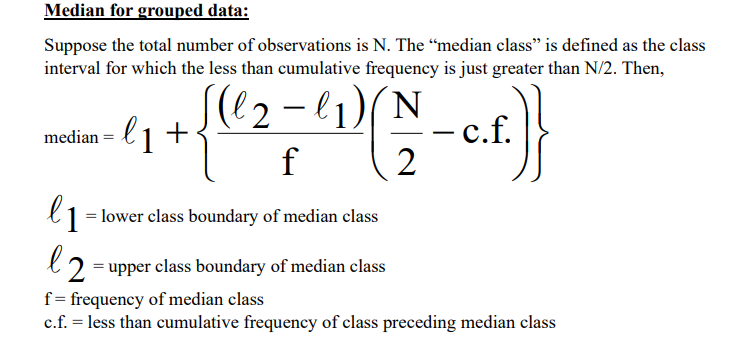
Less than cumulative freq means :

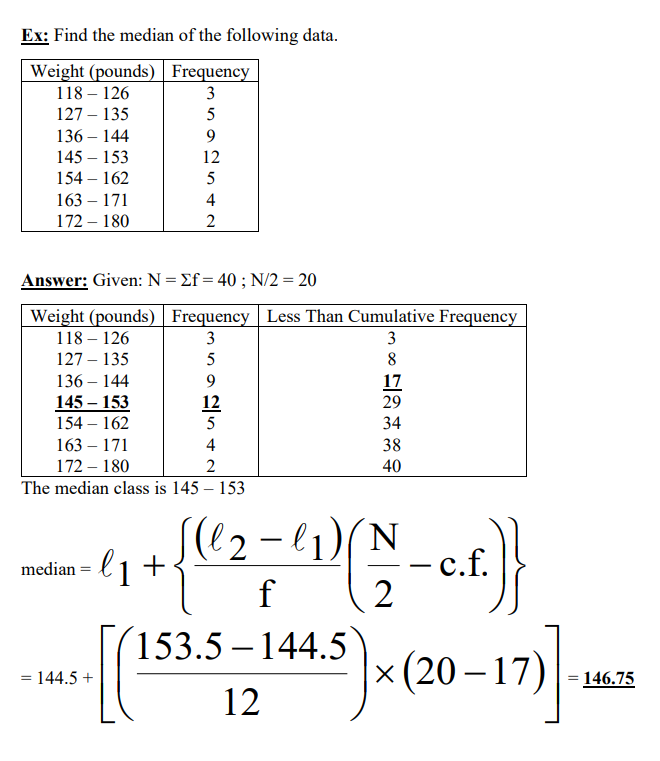
In this example the frequency of values less than 10 is 3,

frequency of values less than 15 = 3+4 = 7

because, 4 values are less than 15 but also 3 values are less than 10, so we have to count them too.

That’s cumulative





* We use c.f. as cumulative frequency less than or equal to N/2
* We calculate median and mode of continuous classes only.
* Here classes are not continuous means, 118 – 126, 127 – 135, and so on,

therefore we take **118 – 126** as **117.5 – 126.5**,

**127 – 135** as **126.5 – 135.5** and so on..

So the classes (x) will be like :

117.5 – 126.5

126.5 – 135.5

135.5 – 144.5 ……….

* If we don’t make them continuous, the answer may vary a bit, but won’t change much.

**MERITS AND DEMERITS OF MEDIAN**

**Merits:**

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

2.It is quite rigidly defined.

3.It can be computed for a distribution with open-end classes. (ex. A class can be like: frequencies Below 10 are 5, frequencies above 50 are 7, likewise)

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

5.It can be determined graphically.

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

extremes are unknown.

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

8.It can be calculated even for qualitative data.

**Demerits:**

1.When the number of observations is large, the pre-requisite of arranging observations in

ascending/descending order of magnitude is a difficult process.

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

3. It is not capable of further mathematical treatment.

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

changes