UNIT-3

Data Theory & Taxonomy of Data: Data as a whole: Understanding of Data as a whole for distinguishing and relating various types of data and Categorization of Data: Structured, Unstructured Data, Quantitative & Qualitative Data.

Views of Data: Understanding Data as an interdisciplinary framework for learning methodologies: covering statistics, neural networks, and fuzzy logic Measurement & Scaling Concepts: Measurement of variables and commonly used Statistical Tools: Number of procedures for measurement of the variables, Categorization procedures, Scale construction procedures and Techniques of data processing for qualitative as well as quantitative data.

Various types of Scales: Nominal, Ordinal, Interval & Ratio Scales

Data as a whole: Understanding of Data as a whole fordistinguishing and relating various types of data and Categorization of Data: Structured, Unstructured Data, Quantitative & Qualitative Data. + Various types of Scales: Nominal, Ordinal, Interval & Ratio Scales

- ***A data object represents an entity—in a sales database, the objects may be customers, store items, and sales; in a medical database, the objects may be patients; in a universitydatabase, the objects may be students, professors, and courses. Data objects aretypically described by attributes.***
- ***Data objects can also be referred to as samples, examples, instances, data points, or objects.*** If the data objects are stored in a database, they are data tuples. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.
- ***What Is an Attribute? An attribute is a data field, representing a characteristic or feature of a data object. The nouns attribute, dimension, feature, and variable are oftenused interchangeably*** in the literature. The term dimension is commonly used in data warehousing. ***Machine learning literature tends to use the term feature, while statisticians prefer the term variable.***
- Data mining and database professionals commonly use the term attribute, and we do here as well. Attributes describing a customer object can include, for example, customer ID, name, and address. Observed values for a given attribute are known as observations.
- ***A set of attributes used to describe a given object is called an attribute vector*** (or featurevector). The distribution of data involving one attribute (or variable) is called univariate.
- A bivariate distribution involves two attributes, the type of an attribute is determined by the set of possible values—nominal, binary, ordinal, or numeric—the attribute can have.
- ***Nominal Attributes Nominal means "relating to names." The values of a nominal attribute are symbols or names of things.*** Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values do not have any meaningful order***. In computer science, the values are also known as enumerations.***
- Suppose that hair color and marital status are two attributes describing person objects. In our application, possible values for hair color are black, brown, blond, red, auburn, gray, and white. The attribute marital status can take on the values single, married, divorced, and widowed. Both hair color and marital status are nominal attributes.
- ***Another example of a nominal attribute is occupation, with the values teacher, dentist, programmer, farmer, and so on. Although we said that the values of a nominal attributeare symbols or "names of things," it is possible to represent such symbols or "names" with numbers. With hair color, for instance, we can assign a code of 0 for black, 1 for brown, and so on.***
- Another example is customor ID, with possible values that are all numeric. However, in such cases, the numbers are not intended to be used quantitatively. That is, mathematical operations on values of nominal attributes are not meaningful. It mkes no sense to subtract one customer ID number from another, unlike, say, subtracting an age value from another (where age is a numeric attribute).

- Even though a nominal attribute may have integers as values, it is not considered a numeric attribute because the integers are not meant to be used quantitatively. Because nominal attribute values do not have any meaningful order about them and are not quantitative, it makes no sense to find the mean (average) value or median (middle) value for such an attribute, given a set of objects. One thing that is of interest, however, is the attribute's most commonly occurring value. This value, known as the mode, is one of the measures of central tendency.
- ***A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as Boolean if the two states correspond to true and false.***
- Binary attributes. Given the attribute smoker describing a patient object, 1 indicates that the patient smokes, while 0 indicates that the patient does not.
- ***The patient undergoes a medical test that has two possible outcomes. The attribute medical test is binary, where a value of 1 means the result of the test for the patient is positive, while 0 means the result is negative.***
- ***A binary attribute is symmetric if both of its states are equally valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1 One such example could be the attribute gender having the states male and female.***
- ***A binary attribute is asymmetric if the outcomes of the states are not equally important,such as the positive and negative outcomes of a medical test for HIV. By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g., HIV positive) and the other by 0 (e.g., HIV negative).***
- ***Ordinal Attributes: An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive valuesis not known.***
- Suppose that drink size corresponds to the size of drinks available at a fast-food restaurant. This nominal attribute has three possible values: small, medium, and large. The values have a meaningful sequence (which corresponds to increasing drink size); however, we cannot tell from the values how much bigger, say, a medium is than a large.
- ***Other examples of ordinal attributes include grade (e.g., A+, A, A−, B+, and so on) and professional rank.*** Professional ranks can be enumerated in a sequential order: for example, assistant, associate, and full for professors, and private, private first class, specialist, corporal, and sergeant for army ranks.
- ***Ordinal attributes are useful for registering subjective assessments of qualities that cannot be measured objectively; thus, ordinal attributes are often used in surveys for ratings.*** In one survey, participants were asked to rate how satisfied they were as customers.
- Customer satisfaction had the following ordinal categories: 0: very dissatisfied, 1: somewhat dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied. Ordinal attributes may also be obtained from the discretization of numeric quantities by splitting the value range into a finite number ofordered categories.

- The central tendency of an ordinal attribute can be represented by its mode and its median (the middle value in an ordered sequence), but the mean cannot be defined.
- ***Note that nominal, binary, and ordinal attributes are qualitative.*** That is, they describe a feature of an object without giving an actual size or quantity. The values of such qualitative attributes are typically words representing categories. If integers are used, they represent computer codes for the categories, as opposed to measurable quantities (e.g., 0 for small drink size, 1 for medium, and 2 for large).
- We look at numeric attributes, which provide quantitative measurements of an object. Numeric Attributes ***A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaledor ratio-scaled.***
- ***Interval-Scaled Attributes Interval-scaled attributes are measured on a scale of equal- size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us tocompare and quantify the difference between values.***
- A temperature attribute is interval-scaled. Suppose that we have the outdoor temperature value for a number of different days, where each day is an object. By ordering the values, we obtain a ranking of the objects with respect to temperature. In addition, we can quantify the difference between values.
- For example, a temperature of 20◦C is five degrees higher than a temperature of 15◦C. Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart.
- Temperatures in Celsius and Fahrenheit do not have a true zero-point, that is, neither 0 ◦C nor 0◦F indicates “no temperature.” (On the Celsius scale, for example, the unit of measurement is 1/100 of the difference between the melting temperature and the boiling temperature of water in atmospheric pressure.)
- Although we can compute the difference between temperature values, we cannot talk of one temperature value as being a multiple of another. Without a true zero, we cannot say, for instance, that 10◦C is twice as warm as 5◦C. That is, we cannot speak of the values in terms of ratios.
- Similarly, there is no true zero-point for calendar dates. (The year 0 does not correspond to the beginning of time.) This brings us to ratio-scaled attributes, for which a true zero-point exits. Because interval-scaled attributes are numeric, we can compute their mean value, in addition to the median and mode measures of central tendency.
- Ratio-Scaled Attributes: ***A ratio-scaled attribute is a numeric attribute with an inherent zero-point.*** That is, if a measurement is ratio-scaled, we can speak of a value as being amultiple (or ratio) of another value.
- In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.
- Ratio-scaled attributes. ***Unlike temperatures in Celsius and Fahrenheit, the Kelvin (K) temperature scale has what is considered a true zero-point (0◦K = −273.15◦C): It is thepoint at which the particles that comprise matter have zero kinetic energy.***
- ***Other examples of ratio-scaled attributes include count attributes such as years of experience (e.g., the objects are employees) and number of words (e.g., the objects***

aredocuments).

- Additional examples include attributes to measure weight, height, latitude and longitude, Getting to Know Your Data coordinates (e.g., when clustering houses), and monetary quantities (e.g., you are 100 times richer with $100 than with $1). 2.1.6
- Discrete versus Continuous Attributes In our presentation, we have organized attributes into nominal, binary, ordinal, and numeric types. There are many ways to organize attribute types. The types are not mutually exclusive. Classification algorithms developed from the field of machine learning often talk of attributes as being either discrete or continuous. Each type may be processed differently.
- ***A discrete attribute has a finite or countably infinite set of values, which may or may not be represented as integers. The attributes hair color, smoker, medical test, and drinksize each have a finite number of values, and so are discrete. Note that discrete attributes may have numeric values, such as 0 and 1 for binary attributes or, the values 0 to 110 for the attribute age.***
- An attribute is countably infinite if the set of possible values is infinite but the values can be put in a one-to-one correspondence with natural numbers.
- For example, the attribute customer ID is countably infinite. The number of customers can grow to infinity, but in reality, the actual set of values is countable (where the values can be put in one-to-one correspondence with the set of integers). Zip codes are another example. If an attribute is not discrete, it is continuous.
- The terms numeric attribute and continuous attribute are often used interchangeably in the literature. (This can be confusing because, in the classic sense, continuous values are real numbers, whereas numeric values can be either integers or real numbers.) In practice, real values are represented using a finite number of digits. ***Continuous attributes are typically represented as floating-point variables.***

Data is classified into different categories based on data structure and scale of measurement of the variables.

Structured and Unstructured Data

- Data at a macro-level can be classified as structured and unstructured data. ***Structured data means that the data is described in a matrix form with labelled rows and columns.Any data that is not originally in the matrix form with rows and columns is an unstructured data.***
- For example, e-mails, click streams, textual data, images (photos and images generated by medical devices), log data, and videos.
- ***Machine generated data such as images generated by satellite, magnetic resonance imaging (MRI), electrocardiogram (ECG) and thermography are few examples of unstructured data.*** There is an increasing trend in the generation of unstructured data due to social media platforms such as Facebook and YouTube and analysis of unstructured data is important for effective management. Internet of things (IoT) is another source unstructured data.
- ***The importance of unstructured data in decision making has increased many folds in the recent past due to its applications to different sectors of the industry.***
- For example, analysing social media data is important for companies to understand the

sentiments expressed by the customers about their products/ services and take necessary remedial measures. Significant proportion of social media data is natural language (text) apart from images and videos. Apart from social media, machine-generated data are usually unstructured (e.g. data generated from medical devices such as ECG, MRI, etc.).

- ❖ High percentage of Big Data problems constitute unstructured data. One of the main challenges in analysing unstructured data is in the conversion of unstructured data to structured data, which then enables model development.
- ❖ ***Examples of structured and unstructured data are shown in Tables 2.1 and 2.2.*** The data in Table 2.2 is a clickstream data (search behaviour of an internet user that capturesthe websites visited by the user). Clickstream data is useful for understanding the behaviour of internet users. Based on their surfing (internet browsing) behaviour, individuals are targeted with advertisement for products and services.
- ❖ The unstructured data as shown in Table 2.2 does not have matrix structure as in the case of structured data in Table 2.1. Before any analytics model can be built, unstructured data has to be converted into a structured data.

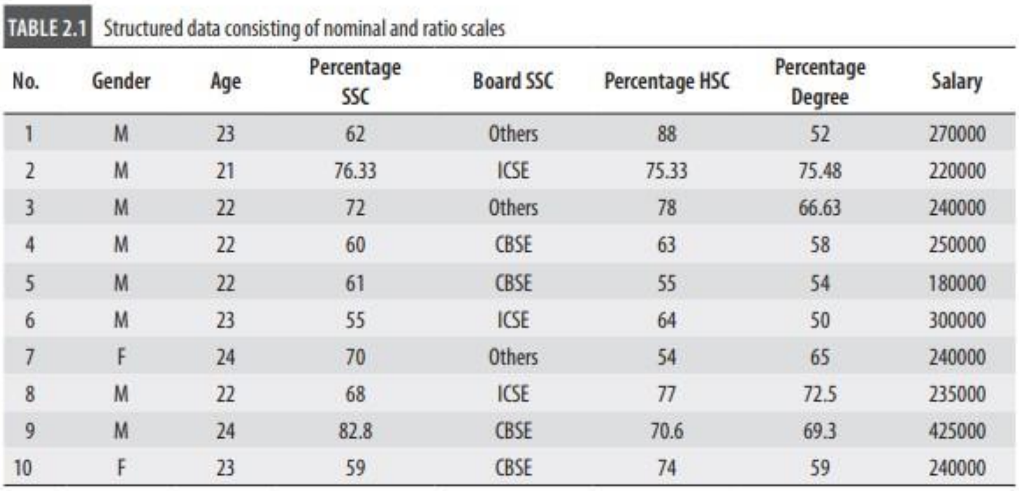


TABLE 2.1 Structured data consisting of nominal and ratio scales

No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000
6	M	23	55	ICSE	64	50	300000
7	F	24	70	Others	54	65	240000
8	M	22	68	ICSE	77	72.5	235000
9	M	24	82.8	CBSE	70.6	69.3	425000
10	F	23	59	CBSE	74	59	240000

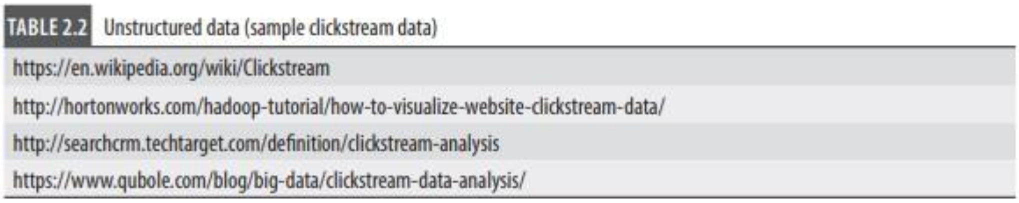


TABLE 2.2 Unstructured data (sample clickstream data)

https://en.wikipedia.org/wiki/Clickstream
http://hortonworks.com/hadoop-tutorial/how-to-visualize-website-clickstream-data/
http://searchcrm.techtarget.com/definition/clickstream-analysis
https://www.qubole.com/blog/big-data/clickstream-data-analysis/

Cross-sectional, Time Series, and Panel Data Another important classification of data is based on the type of data collected. Based on the type of data collected, the data is grouped into the following three classes:

1. Cross-Sectional Data:

- ❖ ***A data collected on many variables of interest at the same time or duration of time is called cross-sectional data. For example, consider data on movies such as budget, box-office collection, actors, directors, genre of the movie during year 2017.***

2. Time Series Data:

- ***A data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.) is called a time series data.***

3. Panel Data:

- ***Data collected on several variables (multiple dimensions) over several time intervals iscalled panel data*** (also known as longitudinal data). Example of a panel data is data collected on variables such as gross domestic product (GDP), Gini index, and unemployment rate for several countries over several years.

Views of Data: Understanding Data as an interdisciplinary framework for learning methodologies: covering statistics, neural networks, and fuzzy logic Measurement & Scaling Concepts: Measurement of variables and commonly used Statistical Tools: Number of procedures for measurement of the variables, Categorization procedures, Scale construction procedures and Techniques of data processing for qualitative as well as quantitative data.

TYPES OF DATA MEASUREMENT SCALES

- ***Structured data can be either numeric or alpha numeric and may follow different scalesof measurement (level of measurement).***
- ***It is important to understand the type of variables within the data with respect to the measurement scale since the model specification while building analytics models suchas regression may depend on the scale of measurement.***
- ***Nominal Scale (Qualitative Data) Nominal scale refers to variables that are basicallynames (qualitative data) and also known as categorical variables.***
- ***For example, variables such as marital status (single, married, divorced) and industrytype (manufacturing, healthcare, banking and finance) fall under nominal scale. Duringdata collection, it is usual to assign a numerical code to represent a***
- ***nominal variable. For example, the data collector may have used number 1 to represent singles, 2 for married, and 3 for divorced category for categorical variable marital status. The codes 1, 2, and 3 used here do not have any value attached to them.***
- ***That is, basic mathematical operations are meaningless in a nominal scale (e.g., subtraction: married – unmarried or ratio: married/unmarried are meaningless). While developing statistical models, nominal scale data are usually transformed before building the model.***
- ***For example, when developing a regression model, categorical variables are convertedusing dummy variables before building the regression model***
- ***Ordinal scale is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.***
- ***For example, in many survey data, Likert scale is used. Likert scale is finite (usually a 5 point scale) and the data collector would have defined the order of preference.***
- ***For example, assume that a feedback is collected on a training program using 5-point Likert scale in which 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, and 5 = Excellent. In this case, we know that 5 is better than 4 and 4 is better than 3; however, the difference 5 – 4 (Excellent – Very Good) is meaningless.***
- ***Interval Scale Interval scale corresponds to a variable in which the value is chosen from an interval set. Variable such as temperature measured in centigrade (°C) or intelligencequotient (IQ) score are examples of interval scale. In interval scale, the ratios do not make sense.***
- ***For example, 40°C is not twice hot as 20°C. Similarly, a person with an IQ score of 160is not twice smarter than a person with an IQ score of 80. However, 40°C is 20°C morethan 20°C, IQ score of 160 is 80 more than an IQ score of 80.***

- ***In an interval scale, the reference is fixed arbitrarily, for example 0°C is fixed based onthe freezing point of water.***
- ***Ratio Scale Any variable for which the ratios can be computed and are meaningful is called ratio scale. Most variables come under this type; for example: demand for a product, market share of a brand, sales, salary, and so on. If Ms Hawai Sundari's salaryis 40,000 per month and Ms Dawai Sundari's salary is 90,000 per month then we can interpret that Dawai Sundari earns 2.25 times the salary of Hawai Sundari.***

POPULATION AND SAMPLE

- ***Population is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem. The size of the population can be very large in many cases.***
- For example, in 2014, close to 834.08 million people were eligible to vote in the Indian general elections (Source: Election Commission of India). Thus, the population size of the voters in 2014 was 834.08 million which included all eligible voters. During every election, media and other organizations collect data to predict likely winner of election through opinion polls (and they rarely get it right due to complexities associated with collecting right sample).
- It is very difficult (also practically impossible) to collect data from all 834.08 million eligible voters about their choice of candidate, so the opinion polls are based on opinion expressed by a subset of voters called sample. Population (also known as universal set) is the set of all possible data for a given context whereas sample is the subset taken from a population.
- In many analytical problems, we make inference about the population based on the sample data. There are many challenges in sampling (process of selecting an observation from the population). An incorrect sample may result in bias and incorrect inference about the population.

MEASURES OF CENTRAL TENDENCY

- Measures of central tendency are the measures that are used for describing the data using a single value. Mean, median and mode are the three measures of central tendency and are frequently used to compare different data sets. Measures of central tendency help users to summarize and comprehend the data.

MEAN (OR AVERAGE) Value Mean is the arithmetical average value of the data and is one of the most frequently used measures of central tendency. Assume that the data has n observations in a sample, and let Xi be the value of the i th observation. Then the mean is given by

$$\text{Mean} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n}$$

Symbol X is frequently used to represent the estimated value of the mean from a sample.

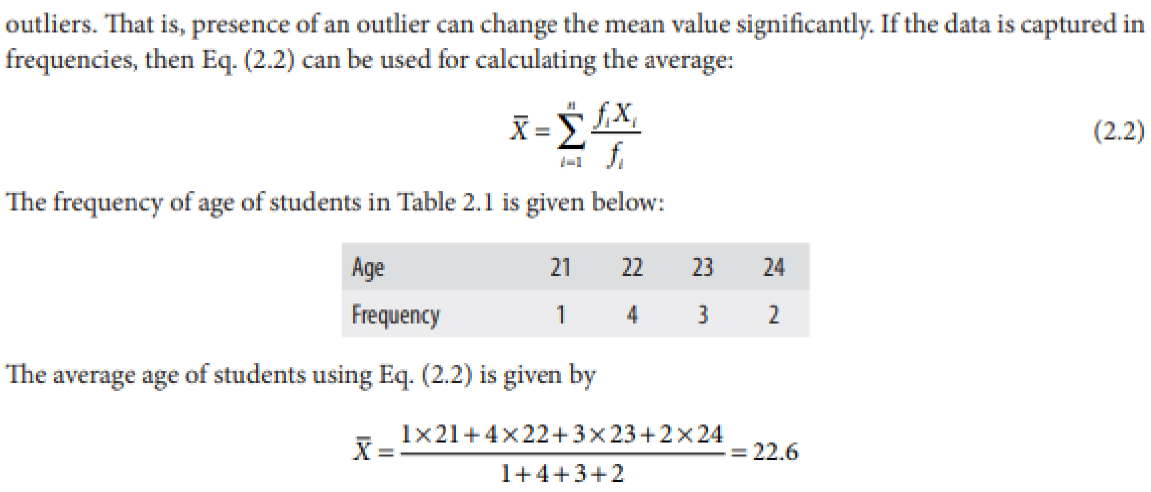
- If the entire population is available and if we calculate mean based on the entire population, then we get the population mean which is denoted by m. Among the measures of central tendency, mean is the most frequently used measure since it uses all the observations (all Xi values) in the data set (either sample or population) to calculate the mean value. Table 2.1 has the salary of graduating students from a business school; the average salary is given by

$$\bar{X} = \frac{(270+220+240+250+180+300+240+235+425+240)\times 1000}{10} = 260000$$

- The average (or mean) salary is 260000. Note that the average value need not be a part of the data set, that is, none of the graduating student’s salary is 260000. In Microsoft Excel, function ‘Average(array)’ can be used for calculating the mean value of the data. Mean can be interpreted as the centre of gravity of the distribution of the data. An important property of mean is that the summation of deviation of observations from the mean is zero, that is

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

- Associated with the mean is a phenomenon often called “wisdom of crowd”, according to which the collective wisdom of people is better than any individual person’s knowledge.
- For example, in 1906, Francis Galton attended a contest in Plymouth, UK in which the villagers were asked to guess the weight of an Ox, the one who guessed the closest won the prize. Around 800 villagers participated in the contest.
- Francis Galton found that the average of all the weights entered was very close to the actual weight. In fact, the difference was less than a pound. Also, the average turned to be better than the guess by the winner of the contest (Surowiecki, 2004).
- One should be careful about taking decisions based on the mean value of the data. There is a famous joke in statistics which says that, “if someone’s head is in freezer and leg is in the oven, the average body temperature would be fine, but the person may not be alive”.
- Making decisions solely based on mean value is not advisable. In capital asset procurement such as procurement of fighter aircraft and weapons, defence services across the world use mean time between failures (MTBF) as one of the measures of system reliability (performance).
- However, MTBF (which is the mean value of the time between failure data) in itself is not a useful measure to assess the reliability of the asset and not very useful in taking operational decisions. It has to be used along with other measures and measures of variability for better understanding of the data. Another issue with mean is, it is affected significantly by presence of



outliers. That is, presence of an outlier can change the mean value significantly. If the data is captured in frequencies, then Eq. (2.2) can be used for calculating the average:

$$\bar{X} = \sum_{i=1}^{n} \frac{f_i X_i}{f_i} \tag{2.2}$$

The frequency of age of students in Table 2.1 is given below:

Age	21	22	23	24
Frequency	1	4	3	2

The average age of students using Eq. (2.2) is given by

$$\bar{X} = \frac{1\times 21 + 4\times 22 + 3\times 23 + 2\times 24}{1+4+3+2} = 22.6$$

MEDIAN (OR MID)

- ***Value Median is the value that divides the data into two equal parts, that is, the proportion of observations below median and above median will be 50%. Easiest way to find the median value is by arranging the data in the increasing order and the medianis the value at position (n + 1)/2 when n is odd. When n is even, the median is the average value of (n/2)th and (n + 2)/2th observation after arranging the data in the increasing order.***
- Consider the example of a bank. The number of deposits in a branch of a bank in a week is shown in Table 2.3

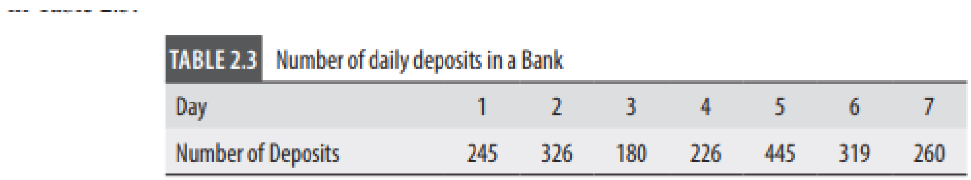
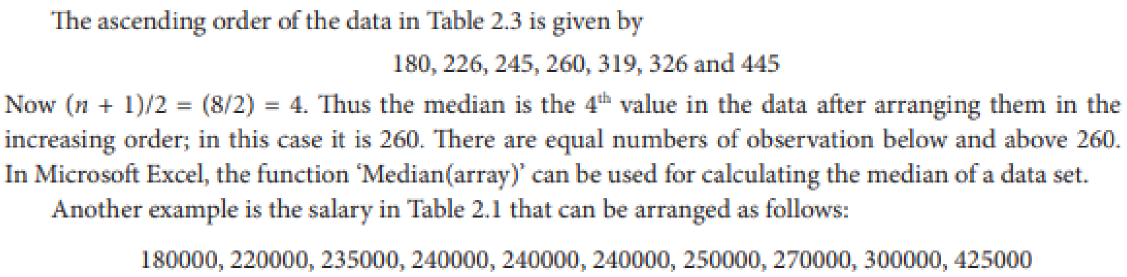


TABLE 2.3 Number of daily deposits in a Bank

Day	1	2	3	4	5	6	7
Number of Deposits	245	326	180	226	445	319	260



The ascending order of the data in Table 2.3 is given by

180, 226, 245, 260, 319, 326 and 445

Now $(n + 1)/2 = (8/2) = 4$. Thus the median is the 4^{th} value in the data after arranging them in the increasing order; in this case it is 260. There are equal numbers of observation below and above 260. In Microsoft Excel, the function 'Median(array)' can be used for calculating the median of a data set.

Another example is the salary in Table 2.1 that can be arranged as follows:

180000, 220000, 235000, 240000, 240000, 240000, 250000, 270000, 300000, 425000

- The 5th and 6th observations are 240000 and 240000 and the average is 240000. Thus, the median salary for the data in Table 2.1 is 240000. Median is much more stable than the mean value, that is adding a new observation may not change the median significantly.
- However, the drawback of median is that it is not calculated using the entire data like in the case of mean. We are simply looking for the midpoint instead of using the actual values of the data.

MODE

- Mode is the most frequently occurring value in the data set. For example, in the data ‘salary’ in Table 2.1, the value 240000 is appearing three times and is the mode since all other values are observed only once.
- In Microsoft Excel, the function ‘Mode(array)’ can be used for calculating mode. Mode is the only measure of central tendency which is valid for qualitative (nominal) data since the mean and median for nominal data are meaningless.
- For example, assume that a customer data with a retailer has the marital status of customer, namely, (a) Married, (b) Unmarried, (c) Divorced Male, and (d) Divorced Female.
- Mean and median are meaningless when we try to use them on a qualitative data such as marital status. On the other hand, mode will capture the customer type in terms of marital status that occurs most frequently in the database. In the bar chart (and histogram), mode is the tallest column.
- It is possible that a data set may not have any mode at all. For example, if each value in the data set appears only once, then there is no mode in the data set.

PERCENTILE, DECILE, AND QUARTILE

- Percentile, decile and quartile are frequently used to identify the position of the observation in the data set. Percentile score is frequently used in education to identify the position of a student in the group.
- Another frequent application of percentile is the percentile life used in asset management. Percentile, denoted as Px , is the value of the data at which x percentage of the data lie below that value.
- For example, P10 denotes the value below which 10 percentage of the data lies. To find Px , we have to arrange the data in the increasing order and the value of Px is the position in the data calculated using Eq. (2.3):

$$\text{Position corresponding to } P_x = \frac{x(n+1)}{100} \qquad (2.3)$$

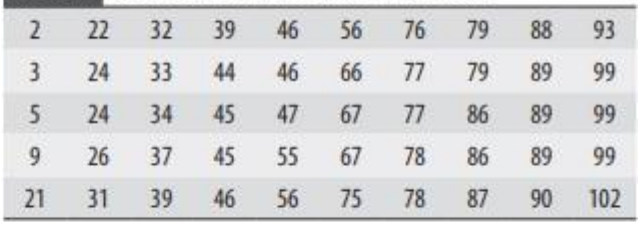
where n is the number of observations in the data.

- Note that the value obtained from Eq. (2.3) can be non-integer, in which case we can either round it to the nearest integer or use an approximation.
- Decile corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on.
- Similarly, Quartile divides the data into 4 equal parts. The first quartile (Q1) contains first 25% of the data, Q2 contains 50% of the data and is also the median. Quartile 3 (Q3) accounts for 75% of the data.
- In Microsoft Excel, the function ‘Percentile(array, k)’ provides Px value. That is, Percentile(array, 0.1) will give 10th percentile.

EXAMPLE 2.1

Time between failures (in hours) of a wire cutter used in a cookie manufacturing oven is given in Table 2.4. The function of the wire-cut is to cut the dough into cookies of desired size.

TABLE 2.4 Time between failures of wire-cut (in hours)



2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

(a) Calculate the mean, median, and mode of time between failures of wire-cuts.

(b) The company would like to know by what time 10% (ten percentile or P10) and

90% (ninety percentile or P90) of the wire-cuts will fail?

(c) Calculate the values of P25 and P75.

(a) Mean = 57.64, median = 56, and mode = 46, 89 and 99.

(b) Note that the data in Table 2.4 is arranged in increasing order in columns. The position of P10 = 10 × (51)/100 = 5.1.

We can round off 5.1 to its nearest integer which is 5. The corresponding value from table is 21 (10 percentage of observations in Table 2.4 have a value of less than or equal to 21). That is, by 21 hours, 10% of the wire-cuts will fail. In asset management (and reliability theory), this value is called P10 life.

Instead of rounding the value obtained from Eq. (2.3), we can use the following approximation: Position corresponding to P10 = 10 × (51)/100 = 5.1 Value at 5th position is 21. Value at position 5.1 is approximated as 21 + 0.1 × (value at 6th position – value at 5th position) = 21 + 0.1(1) = 21.1

Position corresponding to P90 = 90 × 51/100 = 45.9 The value at position 45 is 90 and the value at position 45.9 is 90 + 0.9 (value at 46th position - value at 45th position) = 90 + 0.9 × (3) = 92.7 That is, 90% of the wire-cuts will fail by 92.7 hours.

(c) Position corresponding to P25

(1st Quartile or Q1) = 25 × 51/100 = 12.75

Value at 12th position is 33, so P25 = 33 + 0.75 (value at 13th position – value at 12th position) = 33 + 0.75 (1) = 33.75

Position corresponding to P75 (3rd Quartile or Q3) = 75 × 51/100 = 38.25

Value at 38th position is 86, so P75 = 86 + 0.25 (value at 39th position – value at 38th position) = 86 + 0.25 (0) = 86

MEASURES OF VARIATION

- One of the primary objectives of analytics is to understand the variability in the data. Predictive analytics techniques such as regression attempt to explain variation in the outcome variable (Y) using predictor variables (X).
- Variability in the data is measured using the following measures: 1. Range 2. Inter-Quartile Distance (IQD) 3. Variance 4. Standard Deviation

RANGE Range is the difference between maximum and minimum value of the data. It captures the data spread. In the data in Table 2.4, the range = 102 – 2 = 100.

INTER-QUARTILE DISTANCE (IQD) Inter-quartile distance (IQD), also called inter-quartile range (IQR), is a measure of the distance between Quartile 1 (Q1) and Quartile 3 (Q3).

- For the data in Table 2.4, we calculated Q1 as 33.75 and Q3 as 86.
  Thus, the IQD = 86 – 33.75 = 52.25. IQD is a useful measure for identifying outliers in the data.

OUTLIER is an observation which is far away (on either side) from the mean value of the data. Values of data below Q1 – 1.5 IQD and above Q3 + 1.5 IQD are classified as outliers.

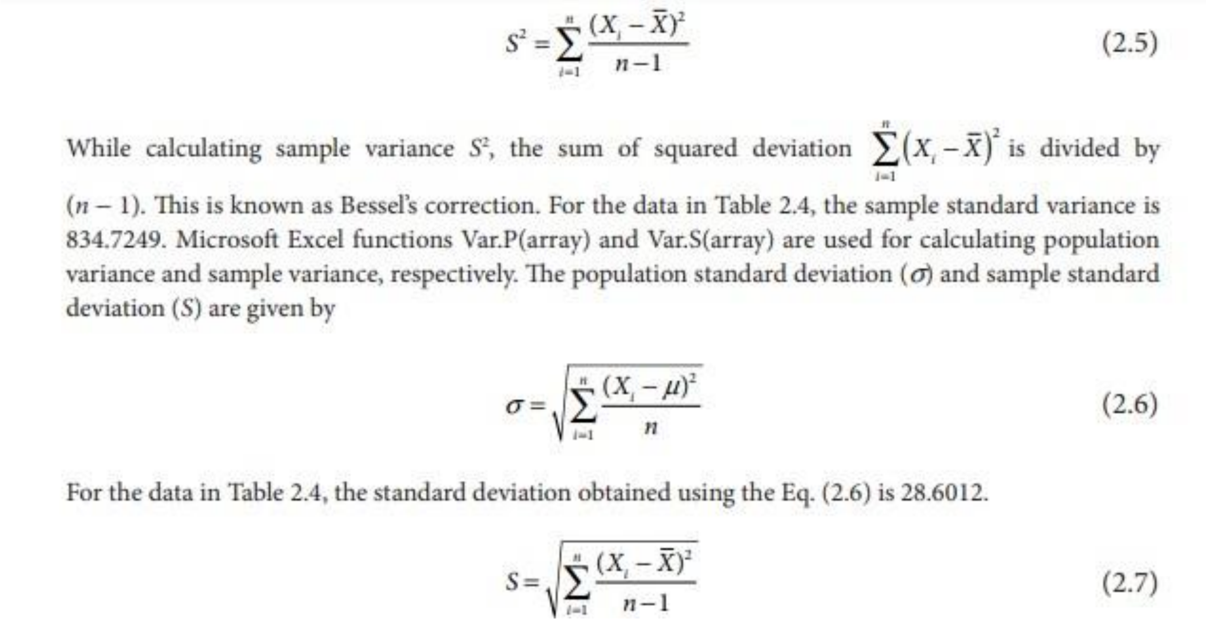
For the data in Table 2.4

Q1 – 1.5 IQD = 33.75 – 1.5 × 52.25 = −44.625 Q3 + 1.5 IQD = 86 + 1.5 × 52.25 = 164.375

- In Table 2.4, there are no values either below – 44.625 or above 164.375, thus there are no outliers. Note that IQD is one of the approaches used for identifying outliers;
- Also, using IQD for identifying outliers is appropriate only in the case of univariate data (data with one dimension). In the case of multivariate data, we use distance measures such as Mahalanobis distance to identify outliers (discussed in Chapters 9 and 10). 2.7.3 | Variance and Standard Deviation Variance is a measure of variability in the data from the mean value. Variance for population, s2 , is calculated using

$$\text{Variance} = \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n} \qquad (2.4)$$

- Note that, in Eq. (2.4), deviation from mean is squared since sum of deviations from mean will always add up to zero. The variance for the data in Table 2.4 is 818.0304 [using Eq. (2.4)]. In case of a sample, the Sample Variance (S2) is calculated using



$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \quad (2.5)$$

While calculating sample variance S^2, the sum of squared deviation $\sum_{i=1}^{n}(X_i - \bar{X})^2$ is divided by $(n-1)$. This is known as Bessel's correction. For the data in Table 2.4, the sample standard variance is 834.7249. Microsoft Excel functions Var.P(array) and Var.S(array) are used for calculating population variance and sample variance, respectively. The population standard deviation (σ) and sample standard deviation (S) are given by

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}} \quad (2.6)$$

For the data in Table 2.4, the standard deviation obtained using the Eq. (2.6) is 28.6012.

$$S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}} \quad (2.7)$$

- For the data in Table 2.4, the standard deviation obtained using the Eq. (2.7) is 28.8916. In Microsoft Excel, functions Stdev.P(array) and Stdev.S(array) are used for calculating population standard deviation and sample standard deviation respectively.
- There are two arguments for dividing the sum of squared deviations from mean by (n − 1) instead of n in Eqs. (2.5) and (2.7). One argument is that, when we take a sample and estimate the mean from the sample X —, we tend to underestimate the sum of squared deviations from the mean.
- For example, take a sample consisting of first 5 (first column) and last 5 (last column) observations from Table 2.4. The sample is given in Table 2.5.

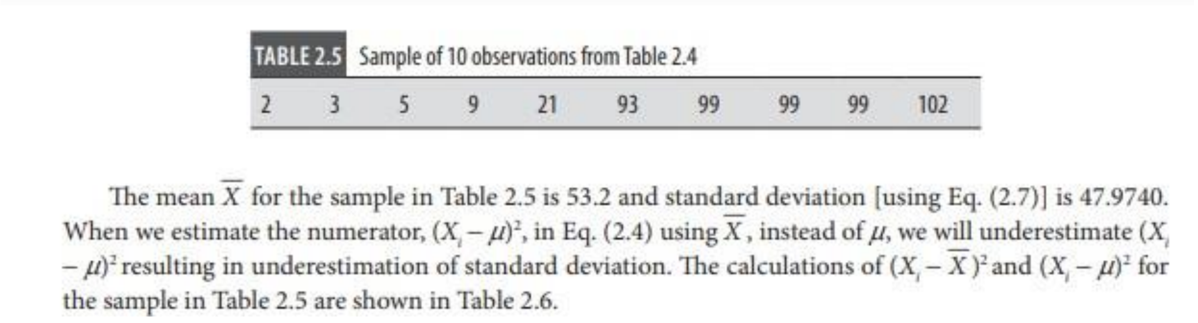


TABLE 2.5 Sample of 10 observations from Table 2.4

2	3	5	9	21	93	99	99	99	102

The mean $\overline{X}$ for the sample in Table 2.5 is 53.2 and standard deviation [using Eq. (2.7)] is 47.9740. When we estimate the numerator, $(X_i - \mu)^2$, in Eq. (2.4) using $\overline{X}$, instead of μ, we will underestimate $(X_i - \mu)^2$ resulting in underestimation of standard deviation. The calculations of $(X_i - \overline{X})^2$ and $(X_i - \mu)^2$ for the sample in Table 2.5 are shown in Table 2.6.

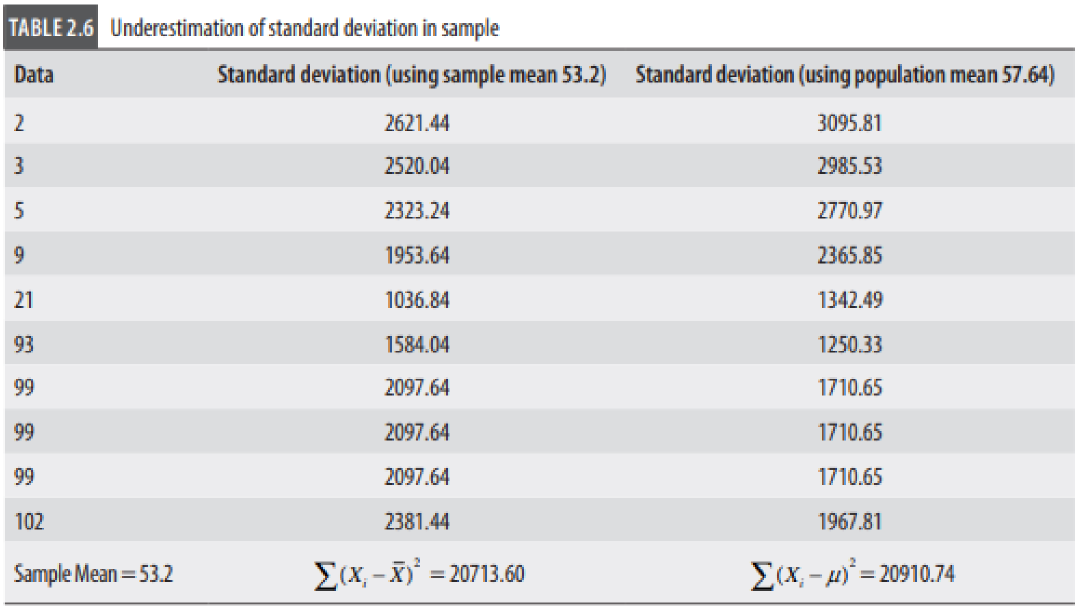


TABLE 2.6 Underestimation of standard deviation in sample

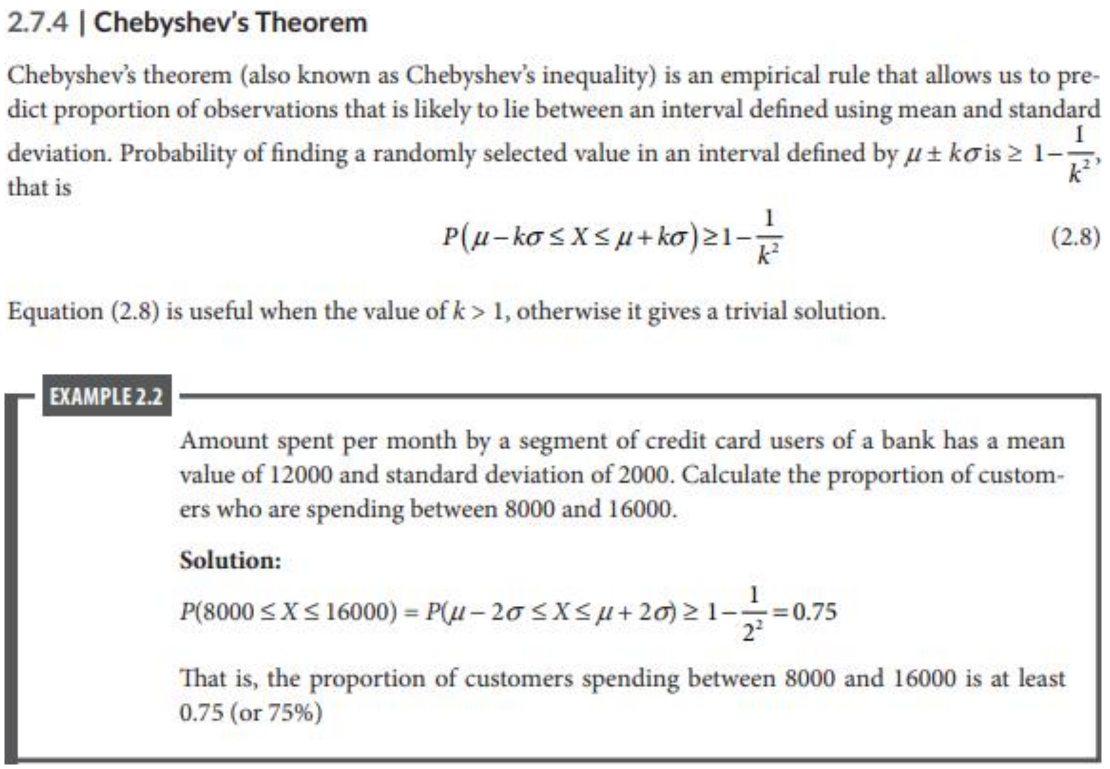
Data	Standard deviation (using sample mean 53.2)	Standard deviation (using population mean 57.64)
2	2621.44	3095.81
3	2520.04	2985.53
5	2323.24	2770.97
9	1953.64	2365.85
21	1036.84	1342.49
93	1584.04	1250.33
99	2097.64	1710.65
99	2097.64	1710.65
99	2097.64	1710.65
102	2381.44	1967.81
Sample Mean = 53.2	$\sum(X_i - \bar{X})^2 = 20713.60$	$\sum(X_i - \mu)^2 = 20910.74$

- In Table 2.6, we can see that the numerator in Eq. (2.4) is underestimated (20713.60) when we use the sample average against population average (20910.74). This will result in underestimation of the standard deviation, a phenomenon called downward bias.
- To overcome this bias, we divide ∑( ) X X i − 2 with (n − 1) instead of n. Another argument of using Eq. (2.5) is through the concept of degrees of freedom.
- The following two definitions are used for degrees of freedom (Pandey and Bright, 2008):

1. **Degrees of freedom is equal to the number of independent variables in the model** (Trochim, 2005). For example, we can create any sample of size n with mean value of X — by randomly selecting (n − 1) values. We need to fix just one out of n values. Thus the number of independent variables in this case is (n − 1).

2. **Degrees of freedom is defined as the difference between the number of observations in the sample and number of parameters estimated** (Walker 1940, Toothaker and Miller, 1996). If there are n observations in the sample and k parameters are estimated from the sample, then the degrees of freedom is (n − k). While using Eq. (2.5) or Eq. (2.7), the value of X —is estimated from the sample.

Thus, the degrees of freedom is (n − 1). Whenever we estimate a parameter from a sample, we lose a degree of freedom. While estimating standard deviation from a sample, we tend to underestimate since mean is also estimated from the sample itself. The downward bias is addressed by dividing the sum of squared deviation from mean with (n − 1) instead of n



2.7.4 | Chebyshev's Theorem

Chebyshev's theorem (also known as Chebyshev's inequality) is an empirical rule that allows us to predict proportion of observations that is likely to lie between an interval defined using mean and standard deviation. Probability of finding a randomly selected value in an interval defined by $\mu \pm k\sigma$ is $\geq 1 - \frac{1}{k^2}$, that is

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2} \tag{2.8}$$

Equation (2.8) is useful when the value of $k > 1$, otherwise it gives a trivial solution.

EXAMPLE 2.2

Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000.

Solution:

$$P(8000 \leq X \leq 16000) = P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75$$

That is, the proportion of customers spending between 8000 and 16000 is at least 0.75 (or 75%)

MEASURES OF SHAPE – SKEWNESS AND KURTOSIS

- ***Skewness is a measure of symmetry or lack of symmetry.***
- ***A data set is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. That is, the proportion ofdata between*** μ and $\mu - k\sigma$ is same as μ and $\mu + k\sigma$, where k ***is some positive constant. Thisimplies that the distribution (or proportion) of the data on either side of mean (and median) is same.***
- Measure of skewness can be used to identify whether the distribution is left skewed (longer tail on left side of the distribution) or right skewed (longer tail on the right side of the distribution). There are many different approaches to measuring skewness.
- Pearson's moment coefficient of skewness for a data set with n observations is given

$$g_1 = \frac{\sum_{i=1}^{n} (X_i - \mu)^3 / n}{\sigma^3} \tag{2.9}$$

by

The value of g1 will be close to 0 when the data is symmetrical. A positive value of g1 indicates a positive skewness and a negative value indicates negative skewness. The formula in Eq. (2.9) is adjusted for sample size when skewness is calculated from a sample. The following formula is used usually for a sample with n observations (Joanes and Gill, 1998):

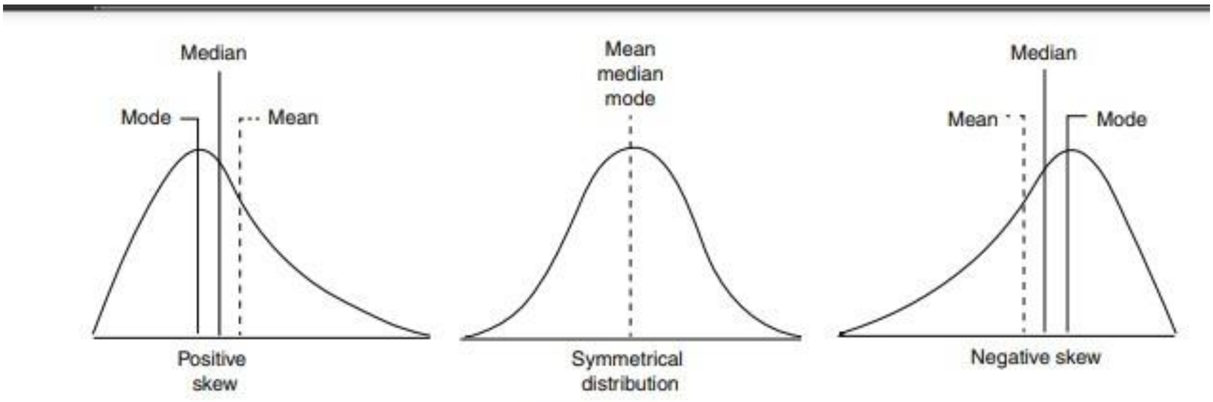


FIGURE 2.1 Skewness.

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \qquad (2.10)$$

The value of $\frac{\sqrt{n(n-1)}}{n-2}$ will converge to 1 as the value of n increases. For the data in Table 2.4, the

value of G1 is −0.232. Since the value of G1 is negative, we can conclude that the data is left skewed.

- In Microsoft Excel, function ‘SKEW(array)’ can be used for calculating the value of skewness (G1) calculated from a sample. In Figure 2.1, the positive skewed (right tailed), normal, and negative skewed (left tailed) distributions are shown.
- Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis is measured using the following equation:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4 / n}{\sigma^4} \qquad (2.11)$$

- Kurtosis value of less than 3 is called platykurtic distribution and greater than 3 is called leptokurtic distribution.
- The kurtosis value of 3 indicates standard normal distribution (also called mesokurtic).
- The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is given by

$$\text{Excess Kurtosis} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4 / n}{\sigma^4} - 3 \qquad (2.12)$$

- For the data in Table 2.4, excess Kurtosis = –1.0968 (that is kurtosis is 1.9032). Figure 2.2 shows shapes of platykurtic, mesokurtic, and leptokurtic distributions.
- In Micrmisoft Excel, ‘KURT(array)’ can be used for calculating the excess kurtosis.

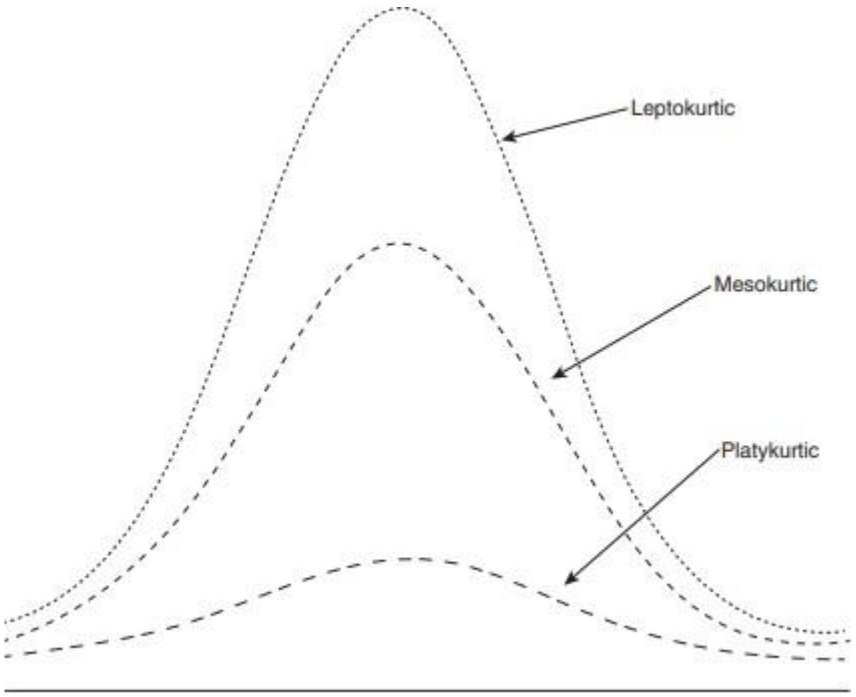


FIGURE 2.2 Leptokurtic, mesokurtic, and platykurtic distributions.

FUZZY SET APPROACH-DATA

- ***Fuzzy Set Approaches Rule-based systems for classification have the disadvantage thatthey involve sharp cutoffs for continuous attributes. For example, consider the following rule for customer credit application approval.***
- ***The rule essentially says that applications for customers who have had a job for two ormore years and who have a high income (i.e., of at least $50,000) are approved:***
- ***IF (years employed ≥ 2) AND (income ≥ 50,000) THEN credit = approved. (9.24) By Rule (9.24), a customer who has had a job for at least two years will receive credit if her income is, say, $50,000, but not if it is $49,000.***
- Such harsh thresholding may seem unfair. Instead, we can discretize income into categories (e.g., {low income, medium income, high income}) and then apply fuzzy logic to allow “fuzzy” thresholds or boundaries to be defined for each category (Figure 9.15).
- Rather than having a precise cutoff between categories, fuzzy logic uses truth values between 0.0 and 1.0 to represent the degree of membership that a certain value has in a given category.
- Each category then represents a fuzzy set. Hence, with fuzzy logic, we can capture the notion that an income of $49,000 is, more or less, high, although not as high as an income of $50,000.
- Fuzzy logic systems typically provide graphical tools to assist users in converting attribute values to fuzzy truth values. Fuzzy set theory is also known as possibility theory. It was proposed by Lotfi Zadeh in 1965 as an alternative to traditional two-value logic and probability theory.

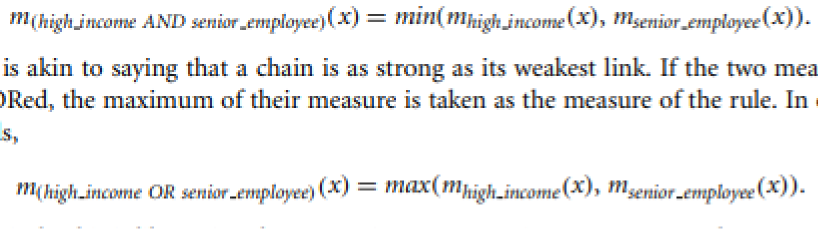
- It lets us work at a high abstraction level and offers a means for dealing with imprecisedata measurement. Most important, fuzzy set theory allows us to deal with vague or inexact facts.
- For example, being a member of a set of high incomes is inexact (e.g., if $50,000 is high, then what about $49,000? or $48,000?) Unlike the notion of traditional “crisp” sets where an element belongs to either a set S or its complement, in fuzzy set theory, elements can belong to more than one fuzzy set.
- For example, the income value $49,000 belongs to both the medium and high fuzzy sets, but to differing degrees. Using fuzzy set notation and following Figure 9.15, this can be shown as

*

$$m_{medium\_income}(\$49{,}000) = 0.15 \; and \; m_{high\_income}(\$49{,}000) = 0.96,$$

where m denotes the membership function, that is operating on the fuzzy sets of medium income and high income, respectively.

- ***In fuzzy set theory, membership values for a given element, x (e.g., for $49,000), do not have to sum to 1. This is unlike traditional probability theory, which is constrainedby a summation axiom.***
- ***Fuzzy set theory is useful for data mining systems performing rule-based classification.It provides operations for combining fuzzy measurements.***
- Suppose that in addition to the fuzzy sets for income, we defined the fuzzy sets junior employee and senior employee for the attribute years employed. Suppose also that we have a rule that, say, tests high income and senior employee in the rule antecedent (IF part) for a given employee, x. If these two fuzzy measures are ANDed together, the minimum of their measure is taken as the measure of the rule. In other words,



$$m_{(high\_income \; AND \; senior\_employee)}(x) = min(m_{high\_income}(x), m_{senior\_employee}(x)).$$

This is akin to saying that a chain is as strong as its weakest link. If the two measures are ORed, the maximum of their measure is taken as the measure of the rule. In other words,

$$m_{(high\_income \; OR \; senior\_employee)}(x) = max(m_{high\_income}(x), m_{senior\_employee}(x)).$$

- Intuitively, this is like saying that a rope is as strong as its strongest strand. Given a tuple to classify, more than one fuzzy rule may apply. Each applicable rule contributes a vote for membership in the categories.
- Typically, the truth values for each predicted category are summed, and these sums are combined. Several procedures exist for translating the resulting fuzzy output into a defuzzified or crisp value that is returned by the system.
- Fuzzy logic systems have been used in numerous areas for classification, including market research, finance, health care, and environmental engineering.

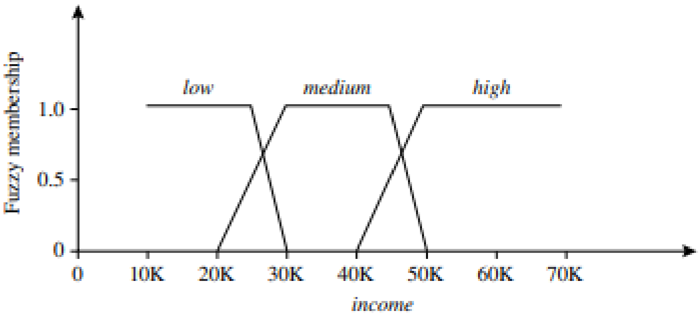


Figure 9.15 Fuzzy truth values for *income*, representing the degree of membership of *income* values with respect to the categories {*low, medium, high*}. Each category represents a fuzzy set. Note that a given income value, *x*, can have membership in more than one fuzzy set. The membership values of *x* in each fuzzy set do not have to total to 1.

ROUGH SET APPROACH

- ***Rough Set Approach Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. It applies to discrete-valued attributes.***
- ***Continuous-valued attributes must therefore be discretized before its use. Rough set theory is based on the establishment of equivalence classes within the given training data.***
- ***All the data tuples forming an equivalence class are indiscernible, that is, the samples are identical with respect to the attributes describing the data. Given real-world data, it is common that some classes cannot be distinguished in terms of the available attributes. Rough sets can be used to approximately or "roughly" define such classes. A rough set definition for a given class, C, is approximated by two sets—a lower approximation ofC and an upper approximation of C.***
- ***The lower approximation of C consists of all the data tuples that, based on the knowledge of the attributes, are certain to belong to C without ambiguity. The upper approximation of C consists of all the tuples that, based on the knowledge of the attributes, cannot be described as not belonging to C.***
- ***The lowerand upper approximations for a class C are shown in Figure 9.14, where each rectangular region represents an equivalence class. Decision rules can be generated foreach class.***
- ***Typically, a decision table is used to represent the rules. Rough sets can alsobe used for attribute subset selection (or feature reduction, where attributes that do not contribute to the classification of the given training data can be identified and removed) and relevance analysis (where the contribution or significance of each attribute is assessed with respect to the classification task).***
- ***The problem of finding the minimal subsets (reducts) of attributes that can describe all the concepts in the given data set is NP-hard. However, algorithms to reduce the computation intensity have been proposed. In one method, for example, a discernibility matrix is used that stores the differences***

between attribute values for each pair of data tuples. Rather than searching on the entiretraining set, the matrix is instead searched to detect redundant attributes.

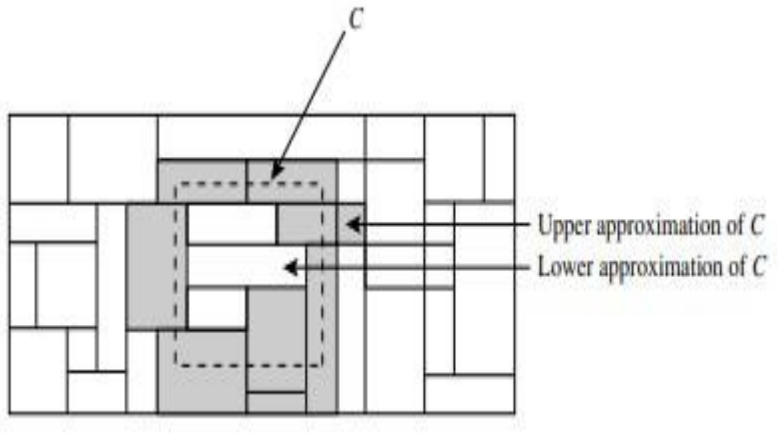


Figure 9.14 A rough set approximation of class C's set of tuples using lower and upper approximation sets of C. The rectangular regions represent equivalence classes.

Data Preprocessing in Data Mining

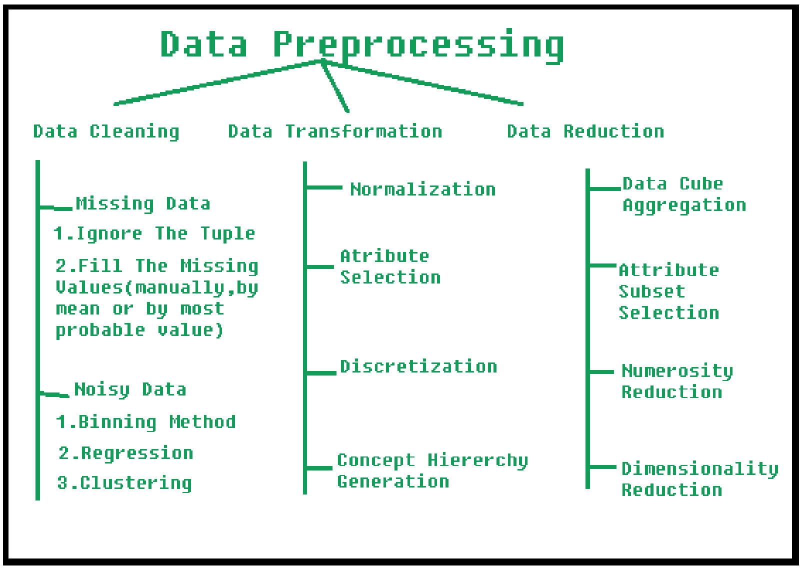
- Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Some common steps in data preprocessing include:

- Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:
- **Data Cleaning:** This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.
- **Data Integration:** This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.
- **Data Transformation:** This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.
- **Data Reduction:** This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as

feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

- **Data Discretization:** This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.
- **Data Normalization:** This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.
- Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.
- By performing these steps, the data mining process becomes more efficient and the results become more accurate.
- **Preprocessing in Data Mining:**
  Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **(a). Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  Some of them are:
  1. **Ignore the tuples:**
     This approach is suitable only when the dataset we have is quite large and multiple

values are missing within a tuple.

2. **Fill the Missing values:**
   There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- **(b). Noisy Data:**
  Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**
     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

3. Data Reduction:
Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

Feature Selection: This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

Feature Extraction: This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

Sampling: This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

Clustering: This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.

Compression: This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

PROBLEMS FOR PRACTICE

1) The cumulative grade point average (CGPA) of 40 students are shown in Table 2.8.

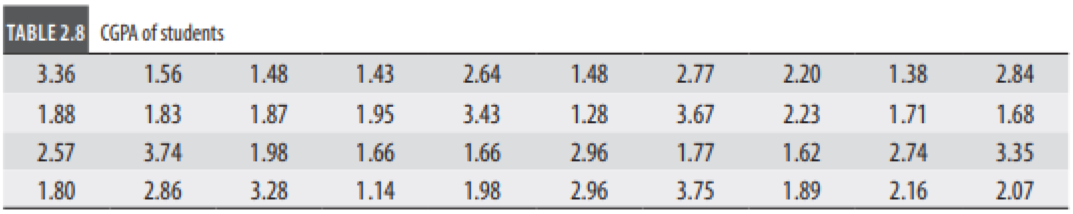


TABLE 2.8 CGPA of students

3.36	1.56	1.48	1.43	2.64	1.48	2.77	2.20	1.38	2.84
1.88	1.83	1.87	1.95	3.43	1.28	3.67	2.23	1.71	1.68
2.57	3.74	1.98	1.66	1.66	2.96	1.77	1.62	2.74	3.35
1.80	2.86	3.28	1.14	1.98	2.96	3.75	1.89	2.16	2.07

(a) Calculate the mean, median and mode. Calculate the standard deviation.

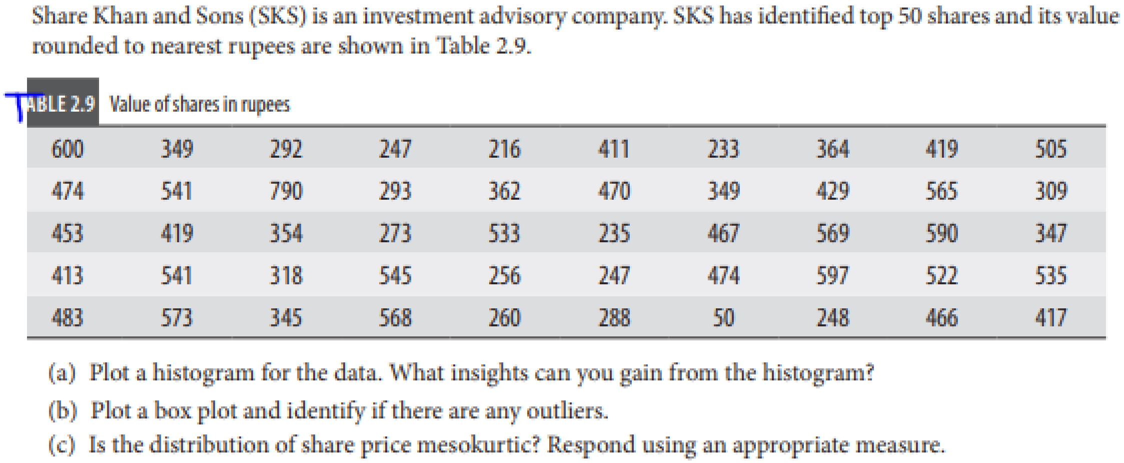
(b) Calculate the 90th and 95th percentile of CGPA.

(c) Calculate the inter quartile range (IQR).

(d) The Dean of the school believes that the CGPA is a right tailed distribution. Is there anevidence to support dean’s belief?

(e) Create a histogram for the data, what should be the ideal number of bins in thehistogram

2) Consider



Share Khan and Sons (SKS) is an investment advisory company. SKS has identified top 50 shares and its value rounded to nearest rupees are shown in Table 2.9.

TABLE 2.9 Value of shares in rupees

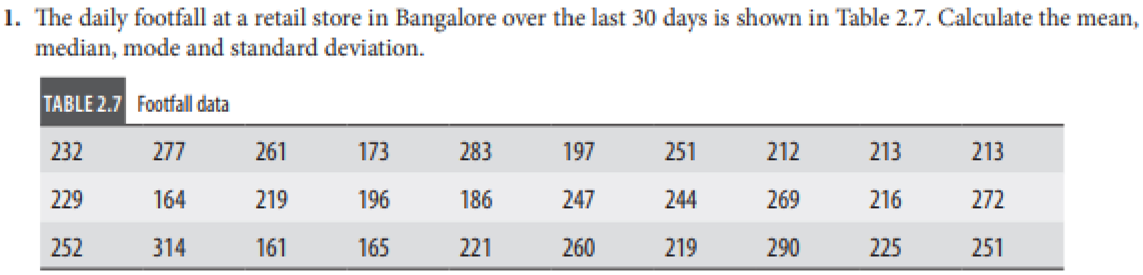
600	349	292	247	216	411	233	364	419	505
474	541	790	293	362	470	349	429	565	309
453	419	354	273	533	235	467	569	590	347
413	541	318	545	256	247	474	597	522	535
483	573	345	568	260	288	50	248	466	417

(a) Plot a histogram for the data. What insights can you gain from the histogram?

(b) Plot a box plot and identify if there are any outliers.

(c) Is the distribution of share price mesokurtic? Respond using an appropriate measure.

3)



1. The daily footfall at a retail store in Bangalore over the last 30 days is shown in Table 2.7. Calculate the mean, median, mode and standard deviation.

TABLE 2.7 Footfall data

232	277	261	173	283	197	251	212	213	213
229	164	219	196	186	247	244	269	216	272
252	314	161	165	221	260	219	290	225	251

4) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the mean of the data? What is the median?

(b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

(c) What is the midrange of the data?

(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

(e) Give the five-number summary of the data.

(f) Show a boxplot of the data.

(g) How is a quantile–quantile plot different from a quantile plot?

5) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the Euclidean distance between the two objects.

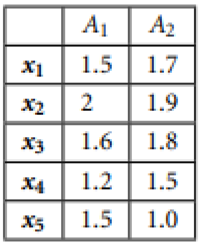
(b) Compute the Manhattan distance between the two objects.

(c) Compute the Minkowski distance between the two objects, using q = 3.

(d) Compute the supremum distance between the two objects.

6 It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on

the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following 2-D data set:



	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

(a) Consider the data as 2-D data points. Given a new data point, x = (1.4,1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.