CM 1110 Fundamentals of Mathematics and Statistics

Dr. Priyanga D. Talagala

2020-02-19

Contents

Course Syllabus

Pre-requisites

None

Learning Outcomes

On successful completion of this module, students will be able to apply fundamental concepts in Mathematics and Statistics for real world problem solving.

Outline Syllabus

- Number Systems
- Sequences and Series
- Introduction to Logic
- Boolean Algebra
- Differentiation and Integration
- Descriptive Statistics
- Sets and Relations
- Probability
- Correlation and Regression

Method of Assessment

- Mid-semester examination
- End-semester examination

Lecturer

Dr. Priyanga D. Talagala

Schedule

Lectures:

- Monday [1.15 pm -3.15 pm]
- Thursday [1.15 pm -3.15 pm]

Tutorial:

- Monday [3.15 pm -4.15 pm]
- Thursday [3.15 pm -4.15 pm]

Consultation time:

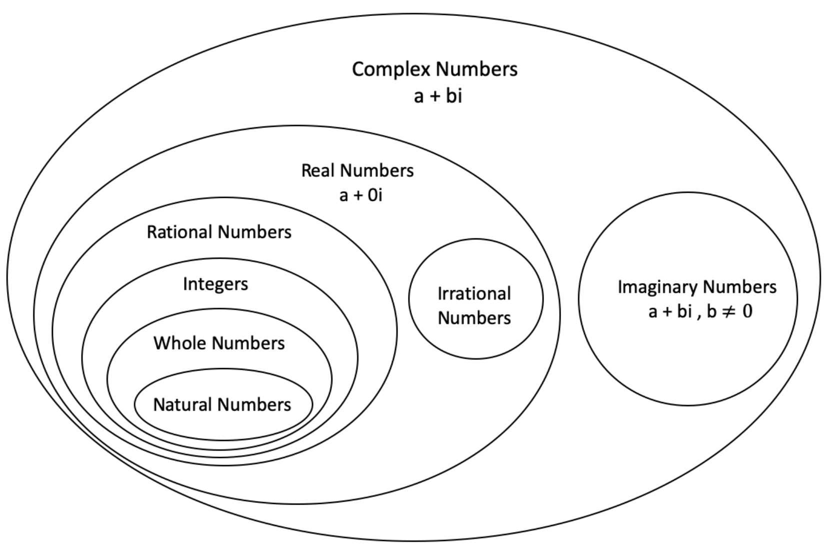
- Tuesday [11.30 am to 12.30 pm]

Chapter 1

Number Systems

Numbers can be classified according to how they are represented or according to the properties that they have.

1.1 Main types



1.1.1 Complex numbers

- Every number in number system taken as a complex number

- A number of the form $a+ib$ is called a complex number when a and b are real numbers and $i=\sqrt{-1}$.
- We call 'a' the real part and 'b' the imaginary part of the complex number $a+ib$.
- IF $a=0$ the number ib is said to be purely imaginary, if $b=0$ the number a is real.
- A pair of complex number $a+ib$ and $a-ib$ are said to be conjugate of each other.
- Show that the sum and product of a complex number and its conjugate complex are both real.

Let $x+iy$ be a complex number and $x-iy$ its conjugate complex.

$$Sum = (x+iy)+(x-iy) = 2x$$

$$Product = (x+iy).(x-iy) = x^2+y^2$$

- Let $a+ib$ and $c+id$ be two complex numbers. Then

Addition. $(a+ib)+(c+id) = (a+c)+i(b+d)$

Subtraction $(a+ib)-(c+id) = (a-c)+i(b-d)$

Multiplication $(a+ib)\times(c+id) = ac-bd+i(ad+bc)$

Addition. $\frac{a+ib}{c+id} = \frac{a+ib}{c+id}.\frac{c-id}{c-id} = \frac{ac+bd}{c^2+d^2}+i\frac{bc-ad}{c^2+d^2}$

- Complex numbers are denoted by $\mathbb{C}$

1.1.2 Imaginary numbers

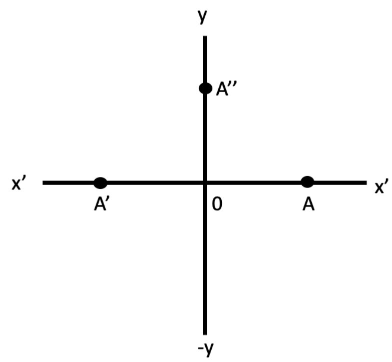
- A number does not exist in the number line is called imaginary number.
- For example square root of negative numbers are imaginary numbers. It is denoted by i. i.e

$$\sqrt{-1} = i$$

$$i^2 = 1$$

- So there is no real number i that satisfies the above equation.
- The quantity i is called the unit imaginary number.

Geometrical Representation of imaginary numbers



- Let OA be positive numbers which is represented by x and OA' by $-x$
- And $-x = (i)^2 x = i(ix)$ is on OX'
- It means that the multiplication of the real number x by i twice amounts to the rotation of OA through two right angles to reach OA'.
- Thus, it means that multiplication of x by i is equivalent to the rotation of x through one right angle to reach OA''.
- Hence, y-axis is known as imaginary axis.
- Multiplication by i rotates its direction through right angle.

1.1.3 Real numbers

- All numbers that can be represented on the number line are called real numbers.
- The real numbers is the set of numbers containing all of the rational numbers and all of the irrational numbers.
- The real numbers are "all the numbers" on the number line.
- Real Numbers are denoted by $\mathbb{R}$.

1.1.4 Rational numbers

- $\mathbb{Q}$
- A rational number is defined as number of the form x/y where x and y are integers and $y \neq 0$.
- i.e Any number which can be expressed as in the form of p/q where p and q are the integers and $q \neq 0$.
- The set of rational numbers encloses the set of integers and fractions.
- The rational numbers that are not integral will have decimal values. These values can be of two types

- Terminating decimal fractions (finite decimal factors): For example $1/5 = 0.5$, $13/5 = 2.6$.
- Non Terminating decimal fractions : The non terminating decimal fractions having two types.
  i) Non terminating periodic fractions
  ii) Non terminating non periodic fractions

i) Non terminating periodic fractions

a. These are non terminating decimal fractions of the type $a.b1b2b3b4b5.....bmb1b2b3b4b5.....bm$

b. Examples

   $19/6 = 3.16666666.....$

   $18/7 = 2.57142857142857.......$

   $21/9 = 2.3333.......$

ii) Non terminating non periodic fractions

a. These are non terminating and there is no periodic decimal places for that number.
b. i.e $a.b1b2b3b4b5.....bmc1c2.........$
c. for example 6.789542587436512………

- **So from above terminating and non terminating periodic fraction numbers belongs to rational numbers.**

1.1.5 Irrational numbers

- Irrational numbers are denoted by $\mathbb{I}$
- An Irrational numbers are non terminating and non periodic fractions.
- i.e irrational number is a number that cannot be written as a ratio x/y form (or fraction).
- In decimal form, is never ends or repeats.
- Examples for irrational numbers are $\sqrt{2} = 1.414213......$, $\pi = 3.14159265.......$, $\sqrt{3}$, $\sqrt{5}$ etc.

1.1.6 Integers

- All numbers that do not having the decimal places in them are called integers.
- All whole numbers including Negative number + Positive number
- $\mathbb{Z} = \{..., -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, ...\}$

- i.e the integer it may positive or negative or zero.
- The set of integers generally written $\mathbb{Z}$for short.
- Any integers are added, subtracted, or multiplied the result is always is an integer.
- When any integers multiplied , each of the multiplied integer is called a factor or divisor of the resulting product

1.1.7 Whole numbers

- The set of whole numbers means narrator numbers and 0
- Whole numbers $= \mathbb{W} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, ...\}$

1.1.8 Natural numbers

- The counting numbers start with 1 and their end is not defined. Generally it is denoted by "N"
- i.e $\mathbb{N} = \{1, 2, 3, 4, ...\}$

Reading :

Dass, H. K. (2008). 'Complex Numbers', *Advanced Engineering Mathematics*. S. Chand Publishing. pp. 474-520.

1.2 Number representations

Decimal: The standard Hindu–Arabic numeral system using base ten.

Binary: The base-two numeral system used by computers.

Hexadecimal: Widely used by computer system designers and programmers, as they provide a more human-friendly representation of binary-coded values.

Octal: Occasionally used by computer system designers and programmers.

Duodecimal: The most convenient numeral system, due to twelve's divisibility by a wide range of the most elemental numbers {1, 2, 3, 4}.

Sexagesimal: Originated with the ancient Sumerians in the 3rd millennium BC, was passed down to the ancient Babylonians

(See positional notation for information on other bases)

Roman numerals: The numeral system of ancient Rome, still occasionally used today.

Tally marks: usually used for counting things that increase by small amounts and don't change very quickly.

Fractions: A representation of a non-integer as a ratio of two integers. These include improper fractions as well as mixed numbers.

Continued fraction: An expression obtained through an iterative process of representing a number as the sum of its integer part and the reciprocal of another number, then writing this other number as the sum of its integer part and another reciprocal, and so on.

Scientific notation: A method for writing very small and very large numbers using powers of 10. When used in science, such a number also conveys the precision of measurement using significant figures.

Knuth's up-arrow notation, Conway chained arrow notation, and Bowers's operators : Notations that allow the concise representation of some extremely large integers such as Graham's number.

https://en.wikipedia.org/wiki/List_of_types_of_numbers

https://byjus.com/maths/number-system/

https://www.researchgate.net/publication/320677641_Number_System

http://www.compsci.hunter.cuny.edu/~sweiss/resources/BinaryNumbers.pdf

https://www.cs.bu.edu/courses/cs101a1/slides/CS101.Lect15.BinaryNumbers.ppt.pdf

https://www.eecs.wsu.edu/~ee314/handouts/numsys.pdf

http://www.uobabylon.edu.iq/eprints/publication_3_8400_6187.pdf

https://www.cs.princeton.edu/courses/archive/spr15/cos217/lectures/03_NumberSystems.pdf

Number system is used in IT applications

- Number Systems Rational, irrational numbers

Chapter 2

Sequences and Series

2.1 Sequences

- A **Sequence** can be thought of as a list of numbers written in a definite order:
$$a_1, a_2, a_3, a_4, ..., a_n, ...$$
- The number a_1 is called the *first term*, a_2 is the *second term*, and in general a_n is called the n^{th} *term*.
- We will deal exclusively with infinite sequences and so each term a_n will have a successor a_{n+1}
- Notice that for every positive integer n there is a corresponding number a_n and so a sequence can be defined as a function whose domain is the set of positive integers.
- But we usually write a_n instead of the function notation $f(n)$ for the value of the function at the number n.
- **Notation:** The sequence $\{a_1, a_2, a_3, ...\}$ is also denoted by
$$\{a_n\}$$
or
$$\{a_n\}_{n=1}^{\infty}$$
- Some sequences can be defined by giving a formula for the nth term.
- In the following examples we give three descriptions of the sequences:

  i. by using the preceding notation
  ii. by using the defining formula
  iii. by writing out the terms of the sequence

- Notice that n doesn't have to start at 1.

a) $\{\frac{n}{n+1}\}_{n=1}^{\infty}$ $\quad a_n = \frac{n}{n+1}$ $\quad \{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, ..., \frac{n}{n+1}, ...\}$

b) $\{\frac{(-1)^n(n+1)}{3^n}\}_{n=1}^{\infty}$	$a_n = \frac{(-1)^n(n+1)}{3^n}$	$\{-\frac{2}{3}, \frac{3}{9}, -\frac{4}{27}, \frac{5}{81}, ..., \frac{(-1)^n(n+1)}{3^n}, ...\}$
c) $\{\sqrt{n-3}\}_{n=3}^{\infty}$	$a_n = \sqrt{n-3}, n \geq 3$	$\{0, 1, \sqrt{2}, \sqrt{3}, ..., \sqrt{n-3}, ...\}$
d) $\{cos\frac{n\pi}{6}\}_{n=0}^{\infty}$	$a_n = cos\frac{n\pi}{6}, n \geq 0$	$\{1, \frac{\sqrt{3}}{2}, \frac{1}{2}, 0, ..., cos\frac{n\pi}{6}, ...\}$

Example

Find a formula for the general term a_n of the sequence

$$\{\frac{3}{5}, -\frac{4}{25}, \frac{5}{125}, -\frac{6}{625}, \frac{7}{3125}, ...\}$$

assuming that the pattern of the first few terms continues.

SOLUTION

We are given that

2.2 Series

Reading:

Stewart, J., Clegg, D. K., & Watson, S. (2020). 'Infinite Sequences and Series', *Calculus: early transcendentals.* Cengage Learning.

Chapter 3

Introduction to Logic

Chapter 4

Boolean Algebra

Chapter 5

Differentiation and Integration

Chapter 6

Descriptive Statistics

6.1 Introduction to Statistics

6.1.1 Some Basic Terminologies Used in Statistics

i Population

- The set of **all** possible elements in the universe of interest to the researcher

ii Sample

- A Sample is a **subset** (a portion or part) of the population of interest
- The sample must be a representative of the population of interest

iii Element

- Element is an **entity or object** which the information is collected.
- *Eg: Student, household, farm, company, tomato plant*

iv Variable

- A variable is **a feature characteristic which has different 'values' or categories for different elements** (items/subjects/individuals)
- *Eg: Gender of client, brand of mobile phones, risk level, number of emails received per day, age of client, income of client*

v Data

- Data are **measurements or facts** that are collected from a statistical unit/entity of interest
- We collect data on variables
- Data are raw numbers or facts that must be processed (analysed) to get useful information.
- We get information from data.
- *Eg:*

Variable: *Age (in years) of client*

Data: *21, 45, 18, 32, 30, 22, 23, 27*

Information:

The mean age is 27.25 years

The minimum age is 18 years

The range of ages is 18-45

The percentage of clients below 25 years of age: 50%

vi Statistic

- **Characteristic** of a **sample**
- The value which calculated based on sample data

vii Parameter

- **Characteristic** of a **population**
- The value which calculated based on population data

viii Census

- When a researcher **gathers data from the whole population for a given measurement,** it is called a census

ix Sampling

- When a researcher **gathers data from a sample of the population for a given measurement,** it is called sampling
- The process of selecting a sample is also called sampling

Why take a sample instead of studying every member of the population ?

- Prohibitive cost of census
- Destruction of item being studied may be required
- Not possible to test or inspect all members of a population being studied.

6.1.2 Branches of Statistics

i Descriptive Statistics

- Descriptive statistics consists of organizing, summarizing and presenting data in an informative way.
- The main purpose of descriptive statistics is to provide an overview of the data collected.
- Descriptive statistics describes the data collected through frequency tables, graphs and summary measures (mean, variance, quartiles, etc.).

ii Inferential Statistics

- In inferential statistics sample data are used to draw inferences (i.e. derive conclusions) or make predictions about the populations from which the sample has been taken.
- This includes methods used to make decisions, estimates, predictions or generalizations about a population based on a sample.
- This includes point estimations, interval estimation, test of hypotheses, regression analysis, time series analysis, multivariate analysis, etc.

6.1.3 Types of Variables

6.1.3.1 Qualitative / Quantitative Variables

i Qualitative variable (Categorical variable)

- The characteristic is a quality.
- The data are categories.
- They cannot be given numerical values.
- However, it may be given a numerical label
- Qualitative variables are sometimes referred as categorical variables.
- *Eg:*

Gender:

Age group:

Education level:

A/L stream:

Degree type:

Hair colour:

FIT student batch:

Undergraduate level:

Grade that you can obtain for CM 1110/ CM1130

ii Quantitative variable

- The characteristic is a quantity
- The data are numbers
- Quantitative data require numeric values that indicate how much or how many.
- They are obtained by counting or measuring with some scale
- *Eg:*

Number of family members:

Number of emails received per day:

Weight of a student:

Age:

Credit balance in the SIM card:

Time remaining in class:

Temperature:

Marks

6.1.3.2 Discrete/ Continuous Variables

- Quantitative variables can be classified as either discrete or continuous.

i Discrete Variables

- Quantitative
- Usually the data are obtained by counting
- There are impossible values between any two possible values
- *Eg:*

Number of family members:

Number of emails received per day:

ii Continuous Variables

- Quantitative
- Usually, the data are obtained by measuring with a scale

- There are no impossible values between any two possible values.(any value between any two possible values is also a possible value)
- i.e a continuous variable can take any value within a specified range.
- *Eg:*

Weight of a student:

Age:

Credit balance in the SIM card:

Time remaining in class:

Temperature:

Marks

6.1.4 Scales of Measurements

- There are four levels of measurements called, **nominal, ordinal, interval and ratio.**
- Each levels has its own rules and restrictions
- Different levels of measurement contains different amount of information with respect to whatever the data are measuring

i Nominal Scale

- Qualitative
- No order or ranking in categories.
- These categories have to be mutually exclusive, i.e. it should not be possible to place an individual or object in more than one category
- A name of a category can be substituted by a number, but it will be mere label and have no numerical meaning

ii Ordinal Scale

- Qualitative
- Categories can be ordered or ranked
- A name of a category can be substituted by a number, but such a sequence does not indicate absolute quantities.
- Difference between any two numbers on the scale does not have a numerical meaningful.
- It cannot be assumed that the differences between adjacent numbers on the scale are equal.

iii Interval Scale

- Quantitative
- Data can be ordered or ranked
- There is no absolute zero point. Zero is only an arbitrary point with which other values can compare
- Difference between two numbers is a meaningful numerical value
- Ration of two numbers is not a meaningful numerical value.

iv Ratio Scale

- Quantitative
- Highest level of measurement
- There exist an absolute zero point (It has a true zero point)
- Ratio between different measurements is meaningful

6.2 Presentation of Data

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew

[1]

Here's a quick summary of our variables:

Variable Name	Description
PassengerID	Passenger ID (just a row number, so obviously not useful for prediction)
Survived	Survived (1) or died (0)
Pclass	Passenger class (first, second or third)
Name	Passenger name
Gender	Passenger Gender
Age	Passenger age
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Fare
Cabin	Cabin
Embarked	Port of embarkation (S = Southampton, C = Cherbourg, Q = Queenstown)

6.2.1 Tabular Presentations of Data

Raw Data

- Raw data are collected data that have not been organized numerically
- Eg: Passenger age

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
```

[1]Data source: https://www.kaggle.com/varimp/a-mostly-tidyverse-tour-of-the-titanic

```
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                     Name    Sex Age SibSp Parch
## 1                              Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                               Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                             Allen, Mr. William Henry   male  35     0     0
## 6                                     Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
##  [1] 22 38 26 35 35 NA 54  2 27 14  4 58 20 39 14 55  2 NA 31 NA 35 34 15 28  8
## [26] 38 NA 19 NA NA 40 NA NA 66 28 42 NA 21 18 14
```

An array

- An array is an arrangement of raw numerical data in ascending or descending order of magnitude.
- Eg: Passenger age

```
##  [1]  2  2  4  8 14 14 14 15 18 19 20 21 22 26 27 28 28 31 34 35 35 35 38 38 39
## [26] 40 42 54 55 58 66
```

Frequency Table (Frequency Distributions)

- A frequency table (frequency distribution) is a listing of the values a variable takes in a data set, along with how often (frequently) each value occurs
- frequency can be recorded as a
  - **frequency or count:** the number of times a value occurs, or
  - **percentage frequency:** the percentage of times a value occurs
- Percentage frequency can be calculated as,

$$Percentage frequency = \frac{a}{b} \times 100\%$$

- The objective of constructing a frequency table are as follows
  - to organize the data in a meaningful manner
  - to determine the nature or shape of the distribution
  - to draw charts and graphs for the presentation of data
  - to facilitate computational procedures for measures of average and spread
  - to make comparisons between different data sets
- There are two basic types of frequency tables
  1. Simple frequency tables (Ungrouped frequency distribution)
  2. Grouped frequency distribution

6.2.1.1 Simple frequency table (Ungrouped frequency distribution)

- Each possible value or category is taken as a class
- More suitable for
  - Qualitative variables
  - Discrete variables
- Sometimes construct for continuous variables when there is a small number of possible values between the minimum and maximum.

Examples:

CASE I:

Example 1

The native countries of 56 students from a certain education institute are as follows:

```
##  [1] "SL" "BD" "SL" "SL" "SL" "SL" "IN" "SL" "SL" "SL" "BD" "SL" "SL" "SL" "IN"
## [16] "SL" "SL" "BD" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "MD" "SL" "SL"
## [31] "SL" "SL" "SL" "SL" "PK" "MD" "PK" "SL" "SL" "SL" "SL" "SL" "PK" "MD" "SL"
## [46] "SL" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "MD" "MD"
```

BD- Bangladesh, IN-India, MD-Maldives, PK-Pakistan, SL- Sri Lanka

Construct a frequency table

```
##  Native Country Count Percentage (%)
##      Bangladesh     3          5.357
##           India     2          3.571
##        Maldives     5          8.929
##        Pakistan     3          5.357
##       Sri Lanka    43         76.786
##           Total    56        100.000
```

CASE II:

Example 2

The grades of 30 students for Statistics are as follows:

```
##  [1] "B" "C" "B" "D" "B" "C" "C" "A" "B" "C" "C" "B" "E" "B" "B" "D" "D" "F" "B"
## [20] "D" "D" "A" "B" "A" "B" "C" "E" "A" "A"
```

Construct a frequency table

```
##  Grade Count Percentage (%)
##      A     5         17.241
##      B    10         34.483
##      C     6         20.690
##      D     5         17.241
##      E     2          6.897
##      F     1          3.448
##  Total    29        100.000
```

CASE III:

Example 3

The number of family members of a sample of undergraduates of Batch 19 are as follows:

```
##  [1] 7 5 3 4 5 4 3 6 4 4 5 2 7 4 5 6 4 4 3 5
```

Construct a frequency table

```
##  Number of family members Count Percentage (%)
##                         2     1              5
##                         3     3             15
##                         4     7             35
##                         5     5             25
##                         6     2             10
##                         7     2             10
##                     Total    20            100
```

CASE IV:

Example 4

The ages (in years) of a sample of undergraduates of Batch 19 are as follows:

```
##  [1] 21 22 22 23 22 24 24 23 21 22 23 22 22 23 21 21 22 23 22 23
```

Construct a frequency table

```
##  Age (years) Count Percentage (%)
##           21     4             20
##           22     8             40
##           23     6             30
##           24     2             10
##        Total    20            100
```

6.2.1.2 Grouped frequency distribution

- A grouped frequency distribution (table) is obtained by constructing classes (or intervals) for the data and then listing the corresponding number of values in each interval.
- Suitable for quantitative variables with large number of possible values in the range of data.
- Note that when items have been grouped in this way, their individual values are lost.
- When studying about frequency distributions it is very important to know the meaning of the following terms

i Class intervals

- In a frequency distribution the total range of the observations are divided into a number of classes. Those are called *class intervals*
- Eg: Class intervals: 10-14, 15-19, 20-24, …, 40-44

ii Class limits

- Class limits are the smallest and largest piece of data value that can fall into a given class.
- In the class interval 10-14, the end numbers, 10 and 14, are called class limits
- The smaller number (10) is the *lower class limit*
- The larger number (14) is the *upper class limit*

iii Class boundaries

- Class boundaries are obtained by adding the upper limit of one class interval to the lower limit of the next-higher class interval and dividing by 2.
- Class boundaries are also called **True class limits**
- Class boundaries **should not** *coincide with actual observations*

Class interval	Class boundaries
10 - 14	9.5 – 14.5
15 - 19	14.5 – 19.5
20 - 24	19.5 – 24.5
25 - 29	24.5 – 29.5
30 - 34	29.5 – 34.5
35 - 39	34.5 – 39.5
40 - 44	39.5 – 44.5

iv The size or width of a class interval

- The size or width of a class interval is the difference between the *lower and upper class boundaries*
- It is also referred to as the *class width, class size, or class length*
- Eg: The class width for the class 10-14 is = 14.5-9.5 = 5

v The class mark (Midpoint of the class)

- Midpoint of the class
- Also called as *class midpoint*
- Midpoint of the class = $\frac{\text{Lower limit+Upper limit}}{2}$

or

- Midpoint of the class = $\frac{\text{Lower boundary+Upper boundary}}{2}$

vi Open class intervals

- A class interval that, at least theoretically, has either no upper class limit or no lower class limit indicated is called an *open class interval*

- For example, referring to age groups of individuals, the class interval "65 year and over" is an open class interval

Rules and Practices for constructing grouped frequency tables

- Every data value should be in an interval
- The intervals should be mutually exclusive
- The classes of the distribution must be arrayed in size order.
- The number of classes not less than 5 or not greater than 15 is recommended.

- The following formula is often used to determine the number of classes: If n is the number of observations, then

$$\text{Number of classes} = \sqrt{n}$$

$$\text{Width of the class interval} = \frac{Range}{\sqrt{n}} = \frac{Min - Max}{\sqrt{n}}$$

- Data should be represented within classes having limits which the data can attain
- Classes should be continuous
- By convention, the beginning of the interval is given the appropriate exact value, rather than the end.
  Eg: intervals of 0-49, 50-99,100-149 would be preferred over the intervals 1-50, 51-100, 101-150 etc.
- The number f observations falling into each category or class interval (class frequency) can be easily found using *tally marks*.

Examples:

In a grouped frequency distribution, class intervals can be constructed in different ways

Example 1

Class interval	Number of students
10 - 14	4
15 - 19	5
20 - 24	11
25 - 29	9
30 - 34	6
35 - 39	3
40 - 44	2

Example 2

Salary	Number of employees
0 – 1999	1
2000 – 3999	31
4000 – 5999	18
6000 – 7999	4
8000 – 9999	2

Salary	Number of employees
10000 - 11999	1
12000 – 13999	0
14000 – 15999	0
16000 – 17999	1
18000 -19999	1
20000-21999	1

Salary	Number of employees
0 – 1999	1
2000 – 3999	31
4000 – 5999	18
6000 – 7999	4
8000 – 9999	2
10000 - 15999	1
16000 – 21999	3
Total	60

Example 3

Salary	Number of employees
Less than 2000	1
2000 – 2999	11
3000 – 3999	20
4000 – 5999	18
6000 – 9999	6
Greater than or equal to 10000	4
Total	60

6.2.1.3 Two-way frequency table

- Cross tabulation, Cross classification table, Contingency table, Two-way table
- Display the relationship between two or more qualitative variables (categorical variables (nominal or ordinal))

```
## # A tibble: 2 x 4
##   Survived First Second Third
##   <chr>    <dbl>  <dbl> <dbl>
## 1 died        80     97   372
```

```
## 2 Survived   136     87   119
```

```
## # A tibble: 2 x 4
##   Survived First Second Third
##   <chr>    <dbl>  <dbl> <dbl>
## 1 died      0.37   0.53  0.76
## 2 Survived  0.63   0.47  0.24
```

6.2.2 Graphic Presentations of Data

- A diagram is a visual form for presentation of statistical data.
- The form of the diagram varies according to the nature of the data

6.2.2.1 Describing Qualitative Data

- Bar chart / Pie chart
- Suitable for
    - Qualitative variables (nominal or ordinal)
    - Discrete variables (when the number of bars or number of different values is small)

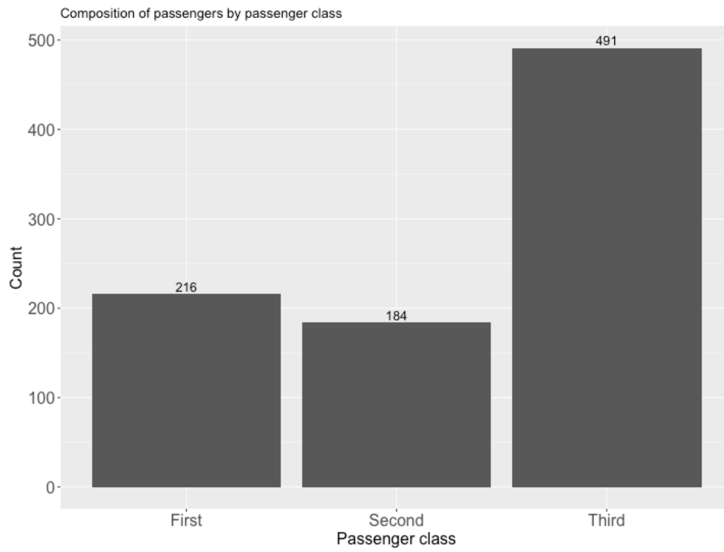
I Bar Chart

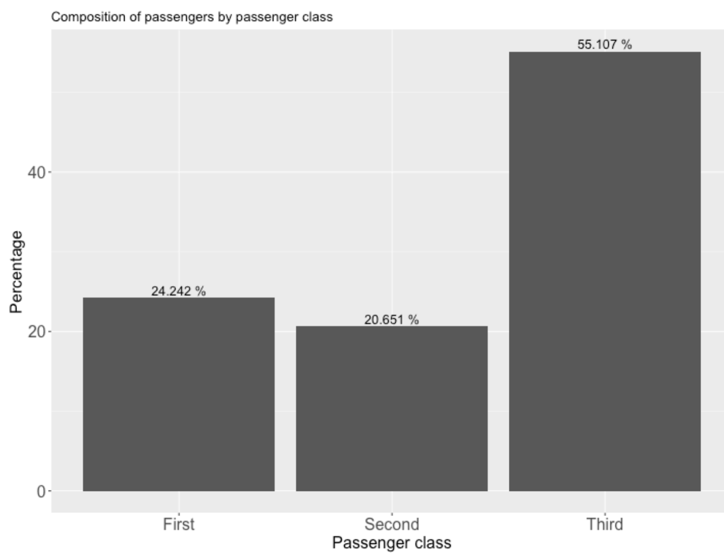
- A bar graph uses bars to represent discrete categories of data
- It can be drawn either on horizontal (more common) or vertical base
- A rectangle of equal width is drawn for each category
- The height (in vertical bar chart) or the length (in horizontal bar chart) of the rectangle is equal to the category's **frequency** or **percentage**.

i Simple Bar Chart

- Only one categorical variable can be presented
- Often used in conjunction with simple frequency tables
- The bars do not touch each other
- The gaps between adjacent bars are same in length

Passenger class	Count	Percentage
First	216	24.242
Second	184	20.651
Third	491	55.107



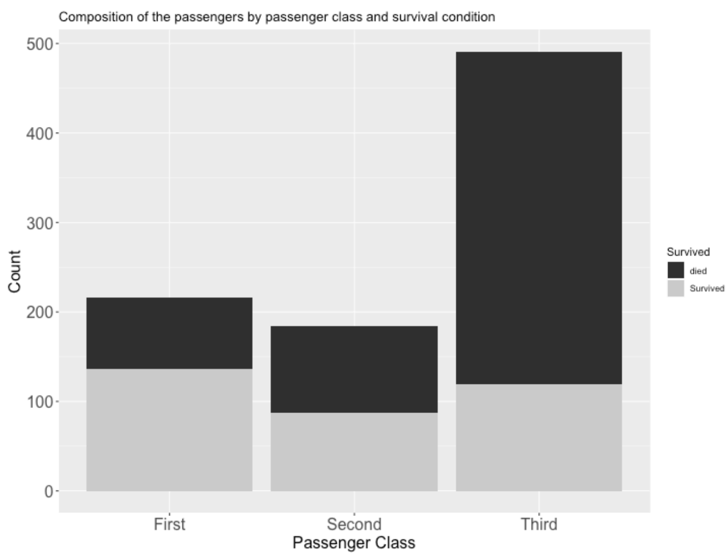


ii Component Bar Chart

- **Sub divided bar chart/ Stacked bar chart**
- Use to compare two or more qualitative variables (nominal or ordinal)
- Often used in conjunction with two way tables
- Start by drawing a simple bar chart with the total figures.
- The bars are then divided into the component parts
- Can be drawn on absolute figures or percentages

- The various components should be kept in the same order in each bar
- To distinguish different components from one another, different colours or shades can be used

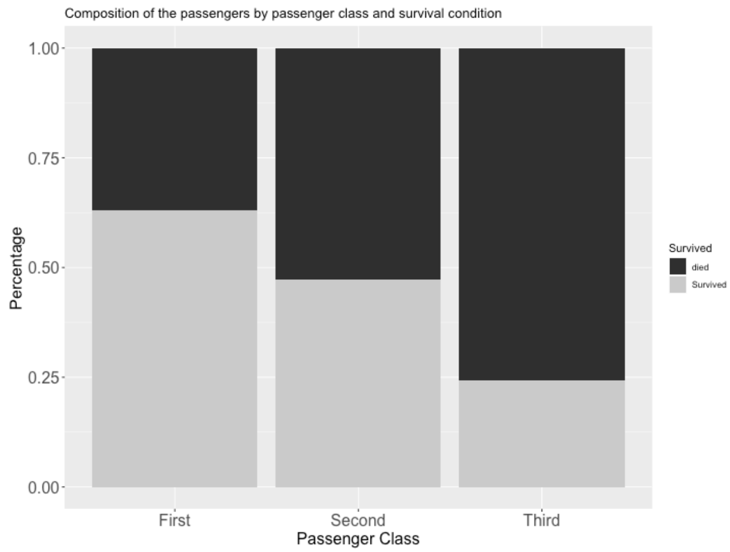
Survived	First	Second	Third
died	80	97	372
Survived	136	87	119



Percentage component bar chart

- When sub-divided bar chart is drawn on percentage basis it is called percentage bar chart
- The various components are expressed as percentage to the total
- All bars are equal in height

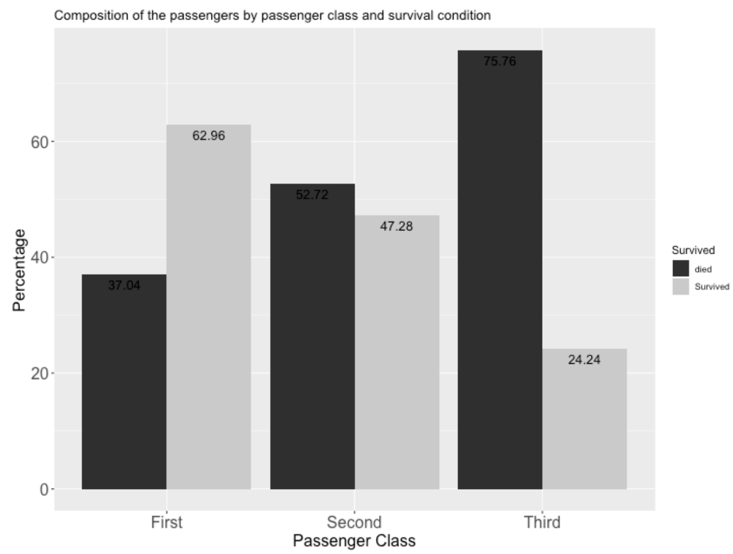
Survived	First	Second	Third
died	0.3703704	0.5271739	0.7576375
Survived	0.6296296	0.4728261	0.2423625



iii Multiple Bar Chart

- Compound bar chart/ Cluster bar chart
- Use to compare two or more qualitative variables (nominal or ordinal)
- Often used in conjunction with two way tables
- These bar charts are drawn side by side

Survived	First	Second	Third
died	37.04	52.72	75.76
Survived	62.96	47.28	24.24



6.2.2.2 Describing Quantitative Data

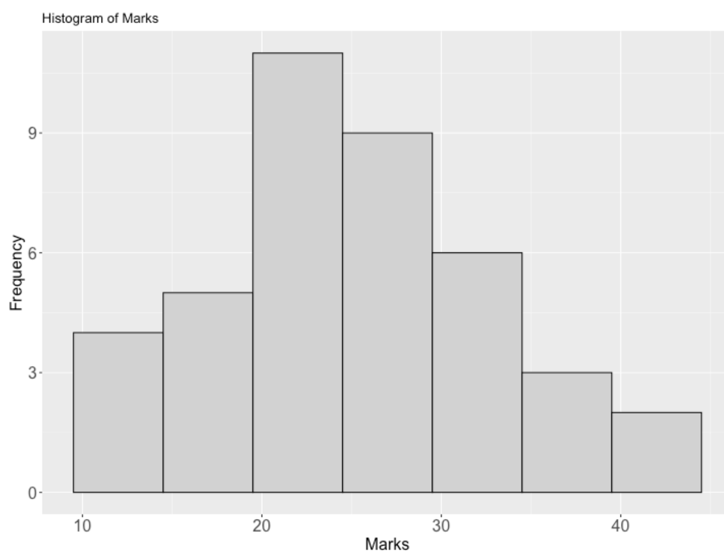
- Histogram/ Dot plot / Box plot/ Scatter plot

II Histogram

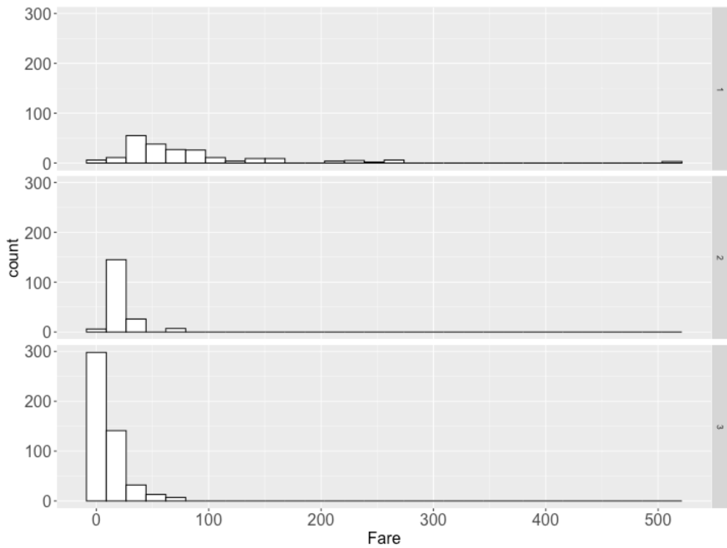
- Histogram looks similar to bar chart since it also has bars.
- However, it is different from a bar chart in a number of aspects.
- One main difference is that in the histogram, the bars are drawn attached to each other; there are no gaps between bars like in a bar chart.
- Histogram is used to show the shape of the distribution of a **continuous variable**.
- However, the histogram is also used for discrete variables when the data are grouped in to class intervals.
- In a histogram, **the area of a bar should be proportional to the frequency of the corresponding class.**
- If all the bars have the same width, then the height of a bar can represent the frequency.
- The bar corresponding to a class interval should be drawn from the lower class boundary to the upper class boundary. In this way there will be no gaps between the bars.

Example: The marks(out of 50) of a group of students are recorded in the accompanying table. Draw a histogram for the data

Marks	Number of students
10 - 14	4
15 - 19	5
20 - 24	11
25 - 29	9
30 - 34	6
35 - 39	3
40 - 44	2
Total	40



Example 2

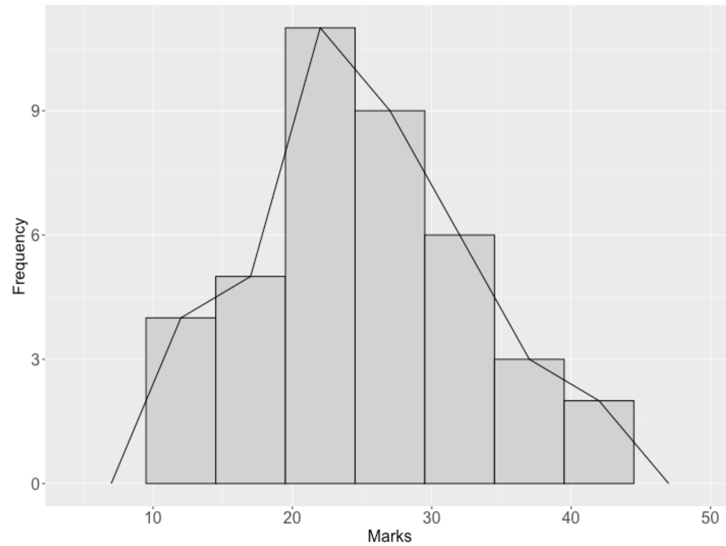


III Frequency polygon

- If the mid-point of the top of each block in a histogram is joined by a straight line, a frequency polygon is produced.
- This is done under the assumption that the frequencies in a class-interval are evenly distributed throughout the class

Example: The marks(out of 50) of a group of students are recorded in the accompanying table. Draw a frequency polygon for the data

Marks	Number of students
10 - 14	4
15 - 19	5
20 - 24	11
25 - 29	9
30 - 34	6
35 - 39	3
40 - 44	2
Total	40

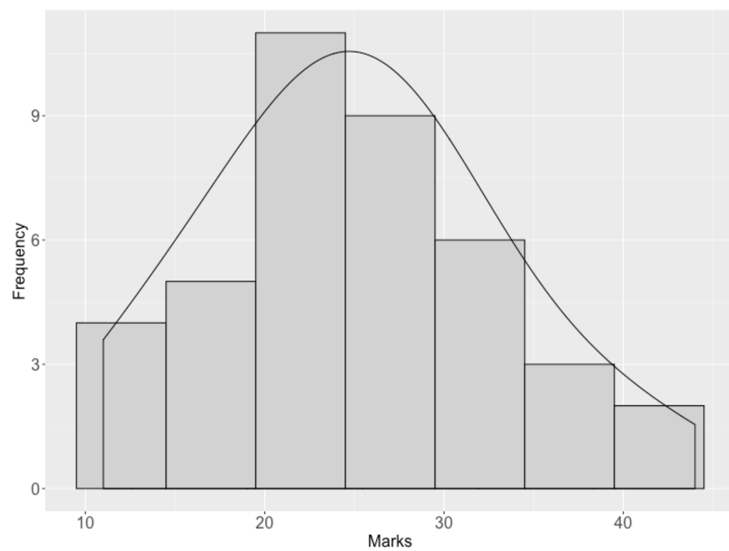


IV Frequency curve

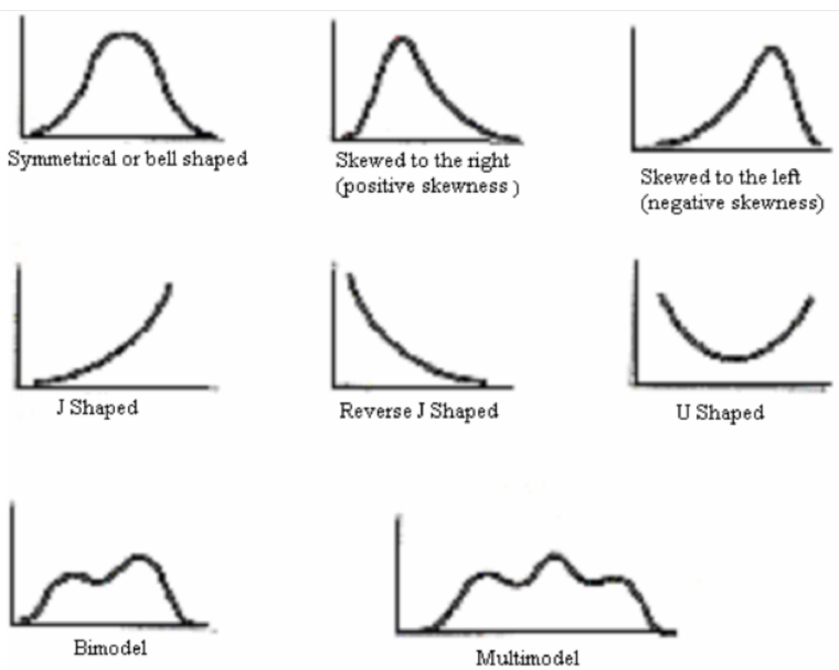
- A frequency curve is drawn by smoothing the frequency polygon.
- It is smooth in such a way that the sharp turns are avoided

Example: The marks(out of 50) of a group of students are recorded in the accompanying table. Draw a frequency curve for the data

Marks	Number of students
10 - 14	4
15 - 19	5
20 - 24	11
25 - 29	9
30 - 34	6
35 - 39	3
40 - 44	2
Total	40



frequency curves arising in practice take on certain characteristics shapes as shown bellow



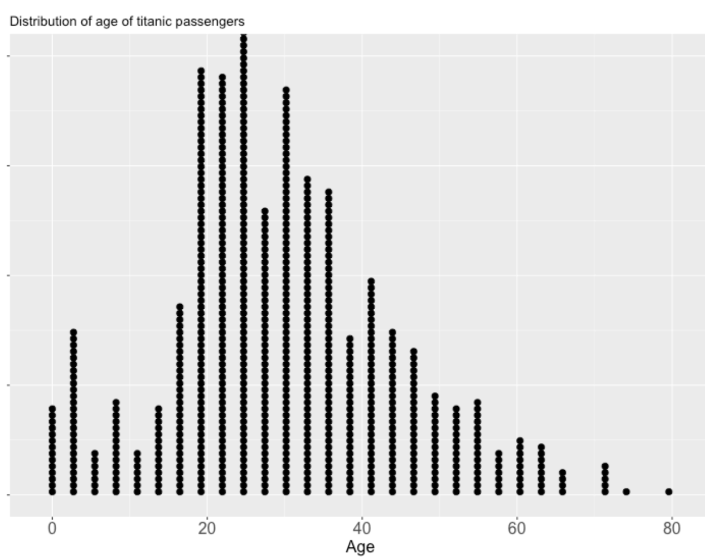
1. The **symmetrical** or **bell shaped** frequency curves are characterized by

the fact that observations equidistant from the central maximum have the same frequency. An important example is the normal curve.

2. In the **moderately asymmetrical** or **skewed** frequency curves the tail of the curve to one side of the central maximum is longer than that to the other. If the longer tail occurs to the right the curve is said to be **skewed to the right** or to have **positive skewness**.While if the reverse is true the curve is said to be **skewed to the left** or to have **negative skewness**.
3. In a **J shaped** or **reverse J shaped** curve a maximum occurs at one end.
4. A **U shaped** frequency curve has maxima at both ends.
5. A **bimodal** frequency curve has two maxima. These appear as two distinct peaks (local maxima) in the frequency curve.When the two modes are unequal the larger mode is known as the major mode and the other as the minor mode.
6. A **multimodal** frequency curve has more than two maxima.

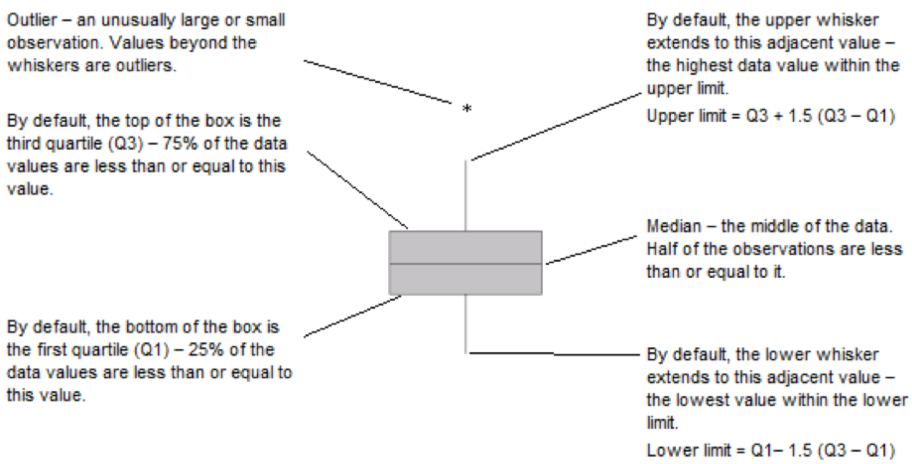
V Dot Plot

- A dot plot is a method of presenting data which gives a rough but rapid visual appreciation of the way in which the data are distributed
- It consists of a horizontal line marked out with divisions of the scale on which the variable is being measured -This graph can be used to represent only the numerical data.



VI Box plot (Box and whisker plot)

- Box plot is also a useful method of representing the behavior of a data set or comparing two or more data sets.
- Box plot is constructed by identifying five statistics from the data set as largest value, smallest values, median, Q1 and Q3.

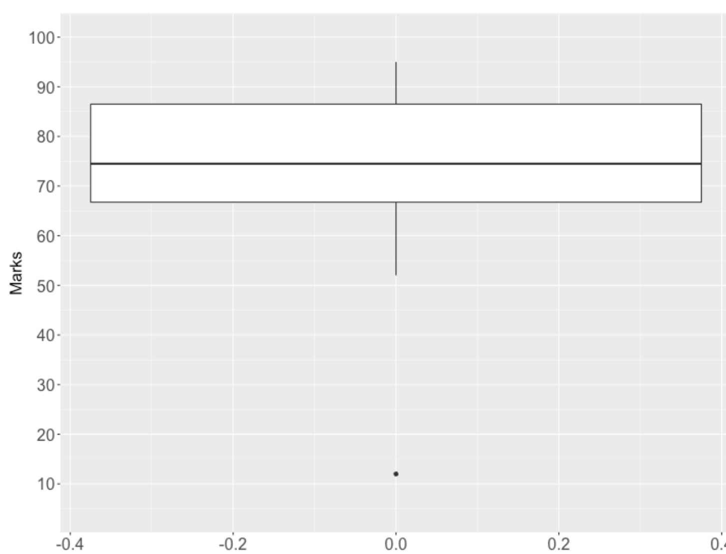


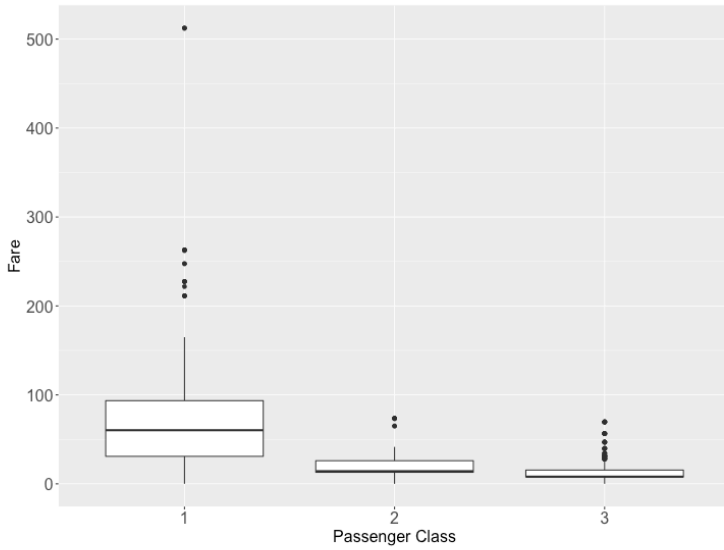
Example:

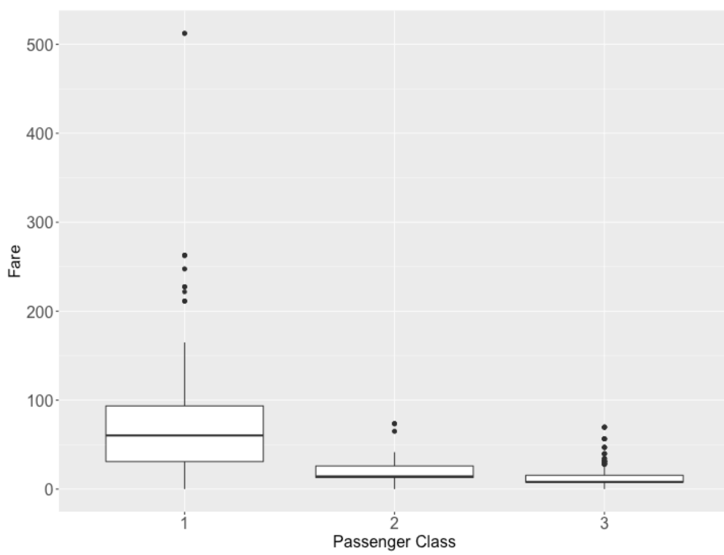
Construct a box plot for the following data set (Marks of students)

52, 88, 56, 79, 72, 91, 85, 88, 68, 63, 76, 73, 86, 95, 12, 69

Xmin = 12 Xmax = 95 Q1 =64.25 Q2 = Median = 74.5 Q3 = 87.5

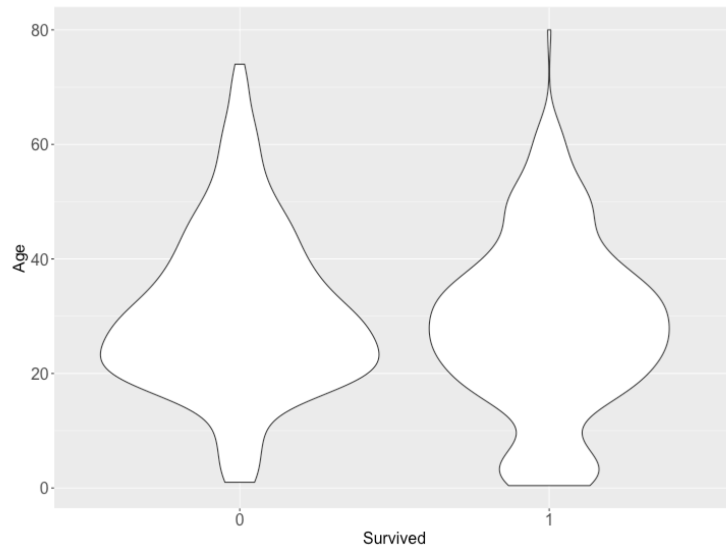

100
90
80
70
60
50
40
30
20
10
Marks
-0.4
-0.2
0.0
0.2
0.4

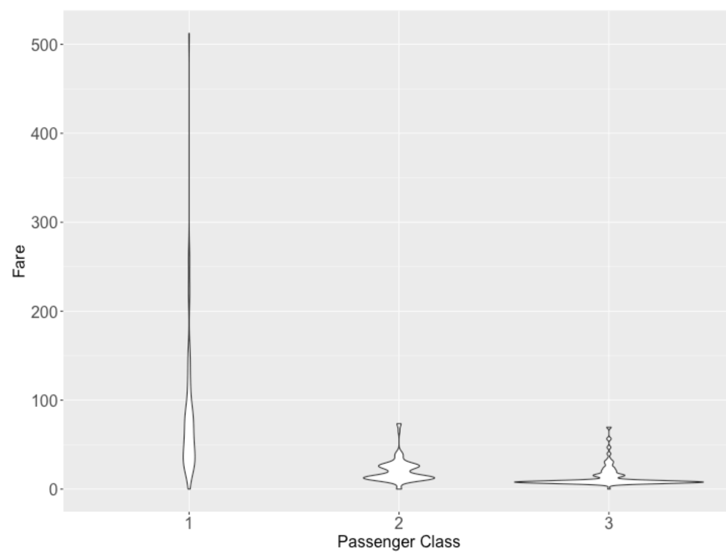




VII Violin plot

- A violin plot is a method of plotting quantitative data.
- It is similar to a box plot, with the addition of a rotated kernel density plot on each side.
- Violin plots are similar to box plots, except that they also show the probability density of the data at different values, usually smoothed by a kernel density estimator.


80
60
40
20
0
Age
0
1
Survived


500
400
300
200
100
0
Fare
1
2
3
Passenger Class

6.3 Summary Measures

- Although frequency distribution serves useful purpose, there are many situations that require other types of data summarization.
- What we need in many instances is the ability to summarize the data by means of a single number called a descriptive measure.
- Descriptive measures may be computed from the data of a sample or the data of a population. To distinguish between them we have the following definitions.

Definitions

- A descriptive measure computed from the data of a sample is called a **statistic**.
- A descriptive measure computed from the data of a population is called a **parameter**.

6.3.1 Measures of Central Tendency

- **Measure of central tendency** yield information about the center, or middle part, of a group of numbers.
- Eg: Mode, Median, Arithmetic Mean, Geometric mean, Harmonic Mean, Quadratic Mean, Quartiles, Deciles, and Percentiles

6.3.1.1 Mode

- The Mode is the most frequently occuring value In a set of data
- Organizing the data into an ordered array (an ordering of the numbers from smallest to largest) helps to locate the mode.
- A series having only one mode is called as **uni-modal**
- In the case of a tie for the most frequent occuring value, two modes are listed. Then the data are set to be **bimodal**
- If a set of data is not exactly bimodal but contains two values that are more dominant than others, some researchers take the liberty of referring to the data set as bimodal even without an *exact tie* for the mode.
- Data sets with more than two modes are referred to as **multimodal**.

- The mode is an appropriate measure of central tendency for nominal-level data.
- The mode can be used to determine which category occurs most frequently.

For ungrouped data

Example 01: Find mode of the following datasets

Dataset 1: 12, 14, 10, 8, 6, 8, 8, 15, 8

Dataset 2: 40, 44, 57, 48, 78

Dataset 3: 42, 45, 55, 50, 45, 40, 55, 45, 52, 55, 54

For grouped frequency data

Example 02: Find mode of the following data

Marks	Number of students
20	8
30	10
40	16
50	8
60	5
70	3

- Advantages and disadvantages of mode

Advantages

- Easy to understand
- Easy to calculate
- Not affected by extreme values in the dataset
- Good for qualitative data

Disadvantages

- Not suitable for further mathematical calculations
- There may be more than one mode for a given dataset
- It is not based upon all the observations
- In some cases, we may not be able to find a mode for a given dataset

6.3.1.2 Median

- The median is the middle value in an ordered array of numbers.
- Median divides the series into equal parts
- The following steps are used to determine the median.
- STEP 1: Arrange the observations in an ordered data array.
- STEP 2: For an **odd number** of terms, find the middle term of the ordered array. It is the median.
- STEP 3: For an **even number** of terms, find the arithmetic mean of the middle two terms. This arithmetic mean is the median.

$$Median = \text{the } (\frac{n+1}{2})\text{th item in the data array}$$

- The level of data measurement must be at least ordinal for a median to be meaningful.

Example 1: Find the median of the dataset 1, 8, 6, 3, 2

Example 2: Find the median of the dataset 8, 9, 1, 2, 14, 12

- Advantages and disadvantages of median

Advantages

- Simple to understand
- Easy to calculate
- Not affected by extreme values in the dataset
- Can be calculated even for qualitative variables (ordinal scale data)

Disadvantages

- It is not based upon all the observations

6.3.1.3 Arithmetic Mean

- The arithmetic mean (usually called mean) is the sum of all observations divided by the total number of observations.

- Population Mean

  - The poupulation mean us represented by the Greek letter *mu* (μ).
  - Let, N is the number of terms in the population.

$$\mu = \frac{\sum x}{N} = \frac{x_1 + x_2 + x_3 + ... + x_N}{N}$$

- Sample Mean

  - The sample mean is represented by $\bar{x}$
  - Let, n is the number of terms in the sample

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

- It is inappropriate to use the mean to analyse data that are not at least interval level in measurement.

Example 1: Calculate the mean from the following data

Student	1	2	3	4	5	6	7	8	9	10
Marks	40	50	53	78	58	60	73	35	43	48

- Advantages and disadvantages of arithmetic mean

Advantages

- Simple to understand
- Easy to calculate
- Based on all the observations
- Well defined
- Unique
- Can be used in further calculation

Disadvantages - Can be affected by extreme values in the dataset - May lead to false conclusions - Only applicable to quantitative data (not applicable to qualitative data)

Empirical relationship between mean, mode, median

- In case of symmetrical distribution, mean, median and mode coincide ($mean = meadian = mode$)

- For a moderately asymmetrical distribution, the following relationship exists $MeanMode = 3(Mean - Median)$

Choice between mean and median

- Mean is very sensitive to outliers.Median is not sensitive to outliers
- When there are outliers in a data set, median is more appropriate than mean

6.3.1.4 Quartiles, Deciles and Percentiles

- Median divides the data set into two equal parts.

- There are other values which divide the data set into a number of equal parts
- Those are Quartiles, Deciles and Percentiles

(a) Quartiles (Q) – Quartiles divide an array into four equal parts

$$Q_i = \text{the } \frac{i}{4}(n+1)\text{th item in the data array}$$

(b) Deciles (D) – Deciles divide an array into ten equal parts

$$D_i = \text{the } \frac{i}{10}(n+1)\text{th item in the data array}$$

(c) Percentiles (P) – Percentiles divide an array into 100 equal parts

$$P_i = \text{the } \frac{i}{100}(n+1)\text{th item in the data array}$$

6.3.2 Measures of Variability

- Statistics that describe the spread or dispersion of a set of data.
- Measure of central tendency yield information about particular points of a data set.
- However, business researchers can use another group of analytic tools to describe a set of data.
- These tools are measures of variability, which describe the spread of the dispersion of a set of data.
- Using measures of variability in conjunction with measures of central tendency makes possible a more complete numerical description of the data.
- This section focuses on seven measures of variability for ungrouped data: range, interquartile range, variance, standard deviation, z score and coefficient of variation.

6.3.2.1 Range

- The range is the difference between the largest value of a data set and the smallest value.

$$Range = Maximum - Minimum$$

- One important use of the range is in quality assurance, where the range is used to construct control charts
- Advantages and disadvantages of range

Advantages

- Easy to understand and calculate

Disadvantages

- Consider only the highest and lowest values of the data and fails to take account of any other observations in the dataset
- Heavily influenced by extreme values

6.3.2.2 Interquartile Range (IQR)

- We use the interquartile range (IQR) to measure the spread of a data around the median (M).
- The interquartile range is the range of values between the first and third quartile.
- Essentially it is the range of the middle 50% of the data and is determined by computing the value of $Q_3 - Q_1$.
- The interquartile range is especially useful in situations where data users are more interested in values towards the middle and less interested in extremes.
- The interquartile range is used in the construction of box and whisker plots.
- By eliminating the lowest 25% and the highest 25% of the items in a series, we are left with the central 50% , which are ordinarily free of extreme values.

Advantages

- Easy to understand and calculate
- Not influenced by extreme values

Disadvantages

- Ignore the first 25% and the last 25% in the dataset

6.3.2.3 Variance and Standard Deviation

- To measure the spread of data around the mean, we use the standard deviation (S).
- The variance and standard deviation are two very popular measures of dispersion.
- These measures are not meaningful unless the data are at least interval-level data.
- Their formulations are categorized into whether to evaluate from a population or from a sample.

NOTE

- Sum of deviations from the arithmetic mean is always zero.

$$\sum (x - \mu) = 0$$

- This property requires considering the alternative ways to obtain measure of variability.

6.3.2.3.1 Variance

- The **variance** is the average of the squared deviations about the mean for a set of numbers.
- The population variance is denoted by σ^2

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

- *The sum of the squared deviations about the mean of a set of values* - called the **sum of squares of** x and sometimes abbreviated as SS_x
- Because the variance is computed from squared deviations, the final result is expressed in terms of squared units of measurements.
- Statistics measured in squared units are problematic to interpret.

6.3.2.3.2 Standard Deviation

- The standard deviation is the square root of the variance.
- The population standard deviation is denoted by σ

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} = \sqrt{\sigma^2}$$

- One feature of standard deviation that distinguishes it from a variance is that the standard deviation is expressed in the same units as the raw data, whereas the variance is expressed in those units squared.

Advantages

- Based on all the observations
- Since this is based on arithmetic mean, it has all the merits of it
- The most important and widely used measure of dispersion

Disadvantages

- Not easy to understand and difficult to calculate
- Gives more weight to extreme values, because the values are squared up

6.3.2.4 Empirical Rule

- The empirical rule is an important rule of thumb that is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data **if the data are normally distributed**
- The empirical rule is used only for three numbers of standard deviations: 1σ, 2σ, 3σ

Distance from the mean	Values within distance
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

- If a set of data is normally distributed, or bell shaped, approximately 68% of the data values are within one standard deviation of the mean, 95% are within two standard deviations, and almost 100% are within three standard deviations.

6.3.2.5 Population versus sample variance and standard deviations

- The sample variance is denoted by s^2 and the sample standard deviation by s.
- The main use for sample variances and standard deviations is as estimators of population variances and standard deviations.
- Thus, computation of the sample variance and standard deviation differs slightly from computation of the population variance and standard deviation.
- Both the sample variance and sample standard deviation use $n - 1$int he denominator instead of n because using n in the denominator of a sample variance results in a statistic that tends to underestimate the population variance.
- While discussion of the properties of *good estimator* is beyond the scope of this course, one of the properties of a good estimator is being *unbiased.*
- Whereas, using n in the denominator of the sample variance makes it a *biased* estimator, using $n - 1$ allows it to be an *unbiased* estimator, which is a desirable property in inferential statistics.

Sample variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Sample standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{s^2}$$

6.3.2.6 Computational formulas for variance and standard deviation

- An alternative method of computing variance and standard deviation, sometimes referred to as the computational method or shortcut method, is available.

- Algebraically,

$$\sum (x - \mu)^2 = \sum x^2 - \frac{(\sum x)^2}{N}$$

and

$$\sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

- Substituting these equivalent expressions into the original formulas for variance and standard deviation yields the following computational formulas.

Computational formula for population variance and standard deviation

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

Computational formula for sample variance and standard deviation

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

$$s = \sqrt{s^2}$$

- For situations in which the mean is already computed or is given, alternative forms of these formulas are:

$$\sigma^2 = \frac{\sum x^2 - N\mu^2}{N}$$

$$s^2 = \frac{\sum x^2 - n(\bar{x})^2}{n-1}$$

6.3.2.7 Coefficient of variation

- In general, for two variables measured with he same units (eg: two groups of people both weighed in kg), the group with the larger variance and standard deviation has n=more variability among their scores.
- The unit of measure affects the size od the variance.
- The same population weights, expressed in 'grams' rather than kg would have a larger variance and standard deviation.
- The *coefficient of variation*, a measure of relative variability gets around this difficulty and makes it possible to compare variability across variables measured in different units.
- The coefficient of variation is the ratio of the standard deviation to the mean, expressed as a percentage and is denoted CV.

$$CV = \frac{\sigma}{\mu}(100)$$

Using median and quartile deviation

$$CV = \frac{\frac{Q_3 - Q_1}{2}}{Median}(100)$$

- The coefficient of variation essentially is a relative comparison of a standard deviation to its mean.
- The coefficient of variation can be useful in comparing standard deviations that have been computed from data with different means.

- The choice of whether to use a coefficient of variation or raw standard deviations to compare multiple standard deviations is a matter of preference

- The coefficient of variation also provides and optional method of interpreting the value of a standard deviation.

6.3.3 Measures of Shape

- Measures of shape are tools that can be used to describe the shape of a distribution of data.

- In this section, we examine two measures of shapes: skewness and Kurtosis.

6.3.3.1 Skewness

- A distribution of data in which the right half is a mirror image of the left half is said to be *symmetrical.*

- One example of a symmetrical distribution is the normal distribution, or bell curve.

- **Skewness** is a measure of symmetry, or more precisely, the lack of symmetry

- The measures of asymmetry are called as measures of skewness.

- The skewed portion is the long, thin part of the curve

Skewness and the relationship of the mean, median and mode

- The concept of skewness helps us to understand the relationship between the mean, median and mode.

- In a unimodal distribution (distribution with a single peak or mode) that is skewed, the mode is the apex (high point) of the curve and the median is the middle value.

- The mean tends to be located toward the tail of the distribution, because the mean is affected by all values, including the extreme ones.

- A bell-shaped or normal distribution with the mean, median and mode all at the centre of the distribution has no skewness.

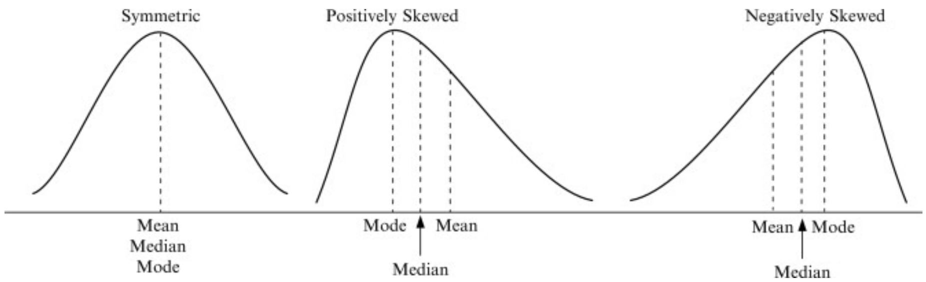


Figure 6.1: Relationship of mean, median and mode for different types of skewness

6.3.3.1.1 Pearsonian coefficient of skewness

- This coefficient compares the mean and median in light of the magnitude of the standard deviation

$$S_k = \frac{3(\mu - M_d)}{\sigma}$$

where S_k = coefficient of skewness, M_d = median

- Note that if the distribution is symmetrical, the mean and median are the same value and hence the coefficient of skewness is equal to zero.
- If the value of S_k is positive, the distribution is positively skewed.
- If the value of S_k is negative, the distribution is negatively skewed.
- The greater the magnitude of S_k, the more skewed is the distribution.

6.3.3.2 Kurtosis

- Kurtosis describes the amount of peakedness of a distribution.
- **Kurtosis** is a measure of whether the data are peaked or flat relative to a normal distribution
- Distributions that are high and thin are referred to as **leptokurtic** distributions.
- Distributions that are flat and spread out are referred to as **platykurtic** distributions.
- Between the above two types are distributions that are more ‘normal’ in shape, referred to as **mesokurtic** distributions

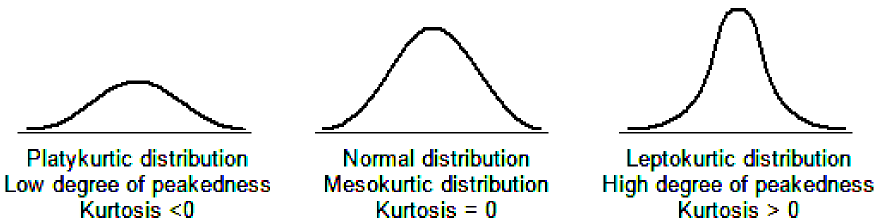


Figure 6.2: Types of kurtosis

References

Black, K., Asafu-Adjaye, J., Khan, N., Perera, N., Edwards, P., & Harris, M. (2007). *Australasian business statistics.* John Wiley & Sons.

Tutorial

Chapter 6: Descriptive Statistics

1. A power company in Sri Lanka designs and manufactures power distribution switchboards and control centres for hospitals, bridges, airports, highways and water treatment plants. Power company director of marketing wants to determine client satisfaction with their products and services. He developed a questionnaire that yields a satisfaction score between 10 and 50 for participant responses. A random sample if 35 of the company's 900 clients is asked to complete a satisfaction survey. The satisfaction scores for the 35 participants are averaged to produce a mean satisfaction score.

a. What is the population for this study?
b. What is the sample for this study?
c. What is the statistic for this study?
d. What would be a parameter for this study?

2. Classify each of the following as nominal, ordinal, interval or ratio data

a. The time required to produce each type on an assembly line
b. the number of litres of milk a family drinks in a month
c. The ranking of four machines in your plan after they have been designated as excellent, good, satisfactory or poor
d. The telephone area code of clients in New Zealand
e. The age of each on your batch mates
f. The sales at the local pizza restaurant each month
g. A student index number
h. The response time of emergency services

Chapter 7

Sets and Relations

Chapter 8

Probability

Chapter 9

Correlation and Regression