Applied Statistics I

(IMT224β/AMT224β)

Department of Mathematics
University of Ruhuna

A.W.L. Pubudu Thilan

Syllabus

- Collecting data
- Samples and Populations
- Summarizing data
- Joint distributions of data
- Linear regression
- Statistical Applications with probability models

Time allocation

Number of lecture hours = 30

Number of tutorial hours = 15

References

- Basic Business Statistics Concept and Applications, Mark L. Berenson (519.5BER).
- Statistics concept and applications, Harray Jrankan, Steven C. Althoen (519.5FRA).
- Sampling Theory, William G. Cochram.
- Applied Statistics for Public Administration, Jeffrey L Bradney.
- www.maths.ruh.ac.lk/∼pubudu

Introduction

- **Statistics** is the study of the collection, organization, and interpretation of data.
- **Mathematical statistics**, which is concerned with the theoretical basis of the subject.
- **Applied statistics** is concerned about application of the subject.

Real world application of statistics

- **Weather Forecasts**
  Computer models are built using statistics that compare prior weather conditions with current weather to predict future weather.
- **Medical Studies**
  Scientists must show a statistically valid rate of effectiveness before any drug can be prescribed.

Cont...

- **Quality Testing**
  Company uses statistics to test just a few, called a sample, of what they make.
- **Stock Market**
  Stock analysts also use statistical computer models to forecast what is happening in the economy.

Collecting data

Collecting data

- Data are the raw numbers or facts which must be processed to give useful information.

  Eg: The marks 78, 64, 36, 70 and 52 are data which could be processed to give the information that the average mark of five students in an exam was 60%.

- Data collection should be designed after deciding the use of the data.

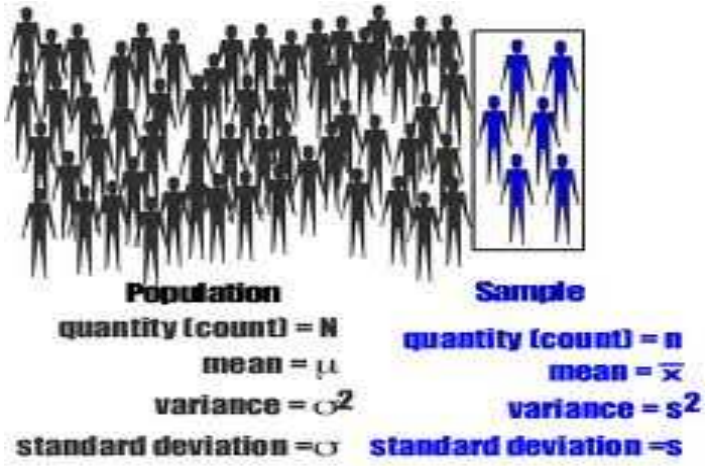
Cont...

- One problem with data collection is knowing how much to collect.
- Data collection and processing inevitably costs money and collecting unnecessary data is wasteful.
- In principle there is an optimal amount of data which should be collected for any purpose.
- The optimal amount of data is not usually calculated, but is suggested in the light of previous experience.

Population vs sample

Population vs sample

- The term **population** is used in statistics to represent all possible measurements or outcomes that are of interest to us in a particular study.
- The term **sample** refers to a portion of the population that is representative of the population from which it was selected.



Finite and infinite population

- A population is called finite if it is possible to count its individuals.

  Eg: The number of vehicles crossing a bridge every day, the number of births per years.

- Sometimes it is not possible to count the units contained in the population. Such a population is called infinite or uncountable.

  Eg: The number of germs in the body of a patient of malaria is perhaps something which is uncountable.

Hypothetical population

A population that does not exist really but exists only in minds of statisticians is called the hypothetical population.

Homogeneous/Non homogeneous populations

A population that is well mixed with respect to the characteristics we measure is called homogeneous population.

Eg : Taste of a curry.

Non homogeneous population is one that is not mixed well.

Eg : Population of Sri Lanka.

Remark
A population that is homogeneous is uniform in composition or character; one that is heterogeneous lacks uniformity in one of these qualities.

Quantitative and qualitative data

- Quantitative data are anything that can be expressed as a number.

  Eg: Scores on achievement tests, number of hours of study.

- Data that cannot be expressed as numerical values are called qualitative.

  Eg: Gender, economic status, religious preference.

Representative and non representative samples

- The sample that represent the corresponding population well is termed as a representative sample. If not it is a non representative sample.
- In statistical sampling, people gather data from a small group and try to extrapolate the results to make generalizations about a larger group.

Sampling methods

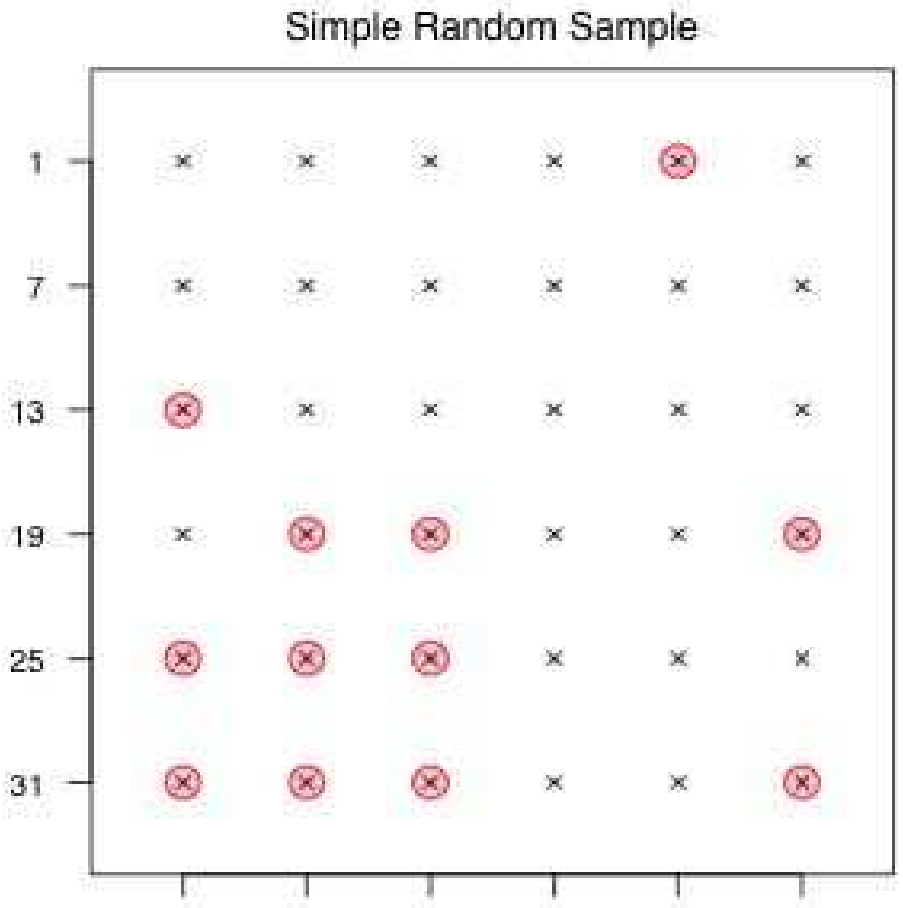
When the set of all possible items in a population is very large it may be too costly or time consuming to do a comprehensive analysis of all of the items. To legitimately be able to use a sample to extrapolate the results to the whole population requires the use of one of statistical sampling methods. These methods are,

- Simple random sample.
- Stratified random sample.
- Cluster sample.
- Systematic sample.

Simple random sampling

- If each unit of the population has an equal chance of being selected for the sample it is called simple random sample.
- The population should be homogeneous and the list of all the items in the population should be available.
- A simple random sample is usually selected by without replacement.

Cont...



Cont...

The following methods are used for the selection of a simple random sample:

- Lottery method
- Using a random number table
- Using the computer

Lottery method

- All the units of the population are numbered from 1 to N.
- These numbers are written on the small slips of paper.
- The slips are thoroughly mixed and a slip is picked up.
- Again the population of slips is mixed and the next unit is selected.
- In this manner, the number of slips equal to the sample size n is selected.

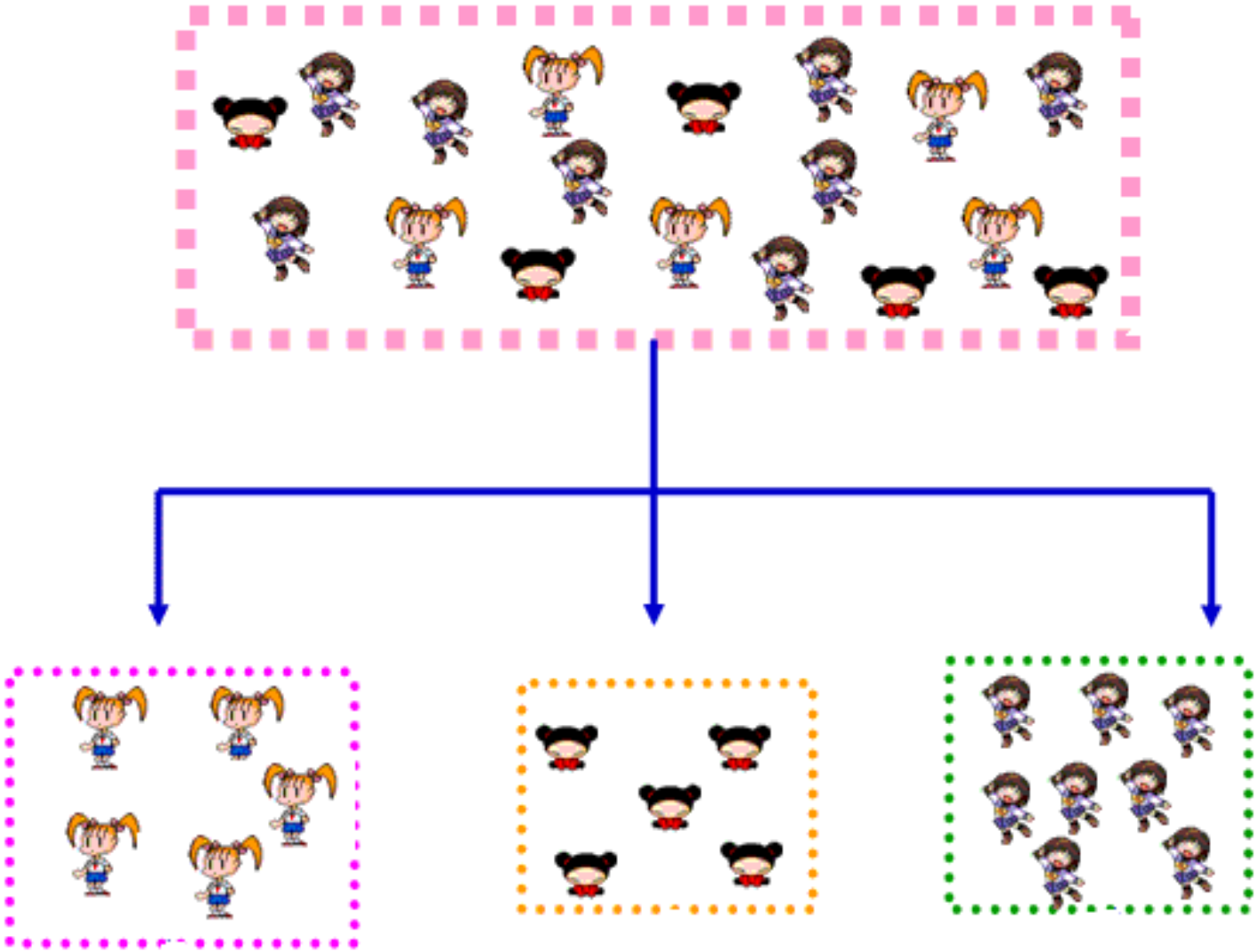
Using a random number table

- Suppose the size of the population is 80 and we have to select a random sample of 8 units.
- The units of the population are numbered from 1 to 80.
- We read two-digit numbers from the table of random numbers.
- We can take a start from any columns or rows of the table.
- Any number above 80 will be ignored and if any number is repeated, we shall not record it if sampling is done without replacement.

Using the computer

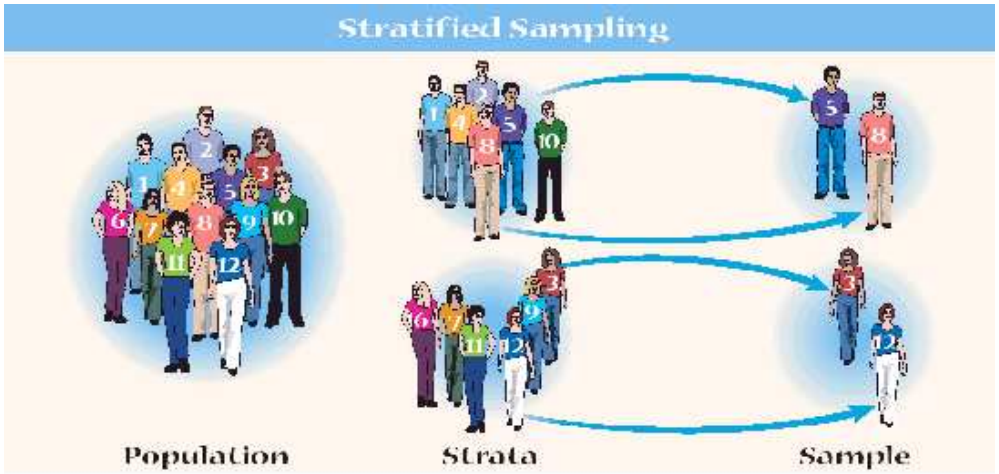
- The facility of selecting a simple random sample is available on the most of the statistical software packages.
- We can use statistical software like SPSS, Minitab, R for that purpose.

What is happend if population is not homogenious?



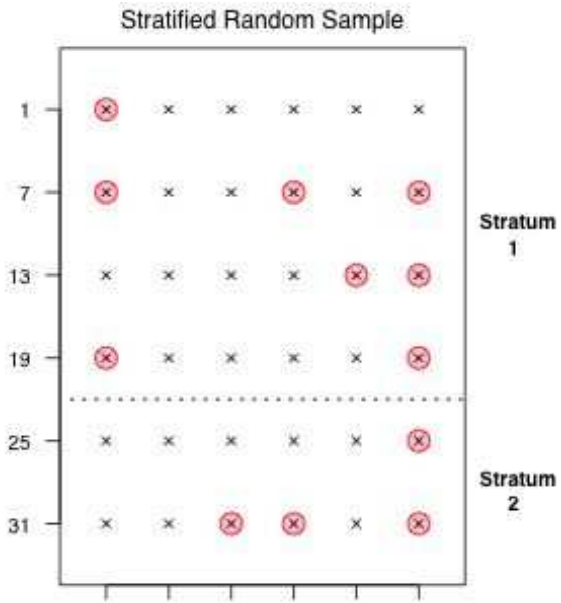
Stratified random sample

- When the population is not homogeneous, the simple random sampling technique is not appropriate.
- We may use stratified random sampling technique, if the population can be separated into subpopulations, which are homogeneous.
- We call these subpopulations as **strata**.
- The sample taken is termed as **stratified random sample**.



Cont...

- To select a simple random sample from a stratum we usually use the proportional allocation method.
- That is we select the size of samples proportional to the sizes of strata.



Example

Suppose that in a company there are the following staff.

Category	Number of persons
male, full time	90
male, part time	18
female, full time	9
female, part time	63

We are asked to take a sample of 40 staff, stratified according to the above categories.

Cont...

The total number of staff =180.

Calculate the percentage in each group.

$$\text{male, full time} = \frac{90}{180} \times 100 = 0.5 \times 100 = 50\%.$$

$$\text{male, part time} = \frac{18}{180} \times 100 = 0.1 \times 100 = 10\%.$$

$$\text{female, full time} = \frac{9}{180} \times 100 = 0.05 \times 100 = 5\%.$$

$$\text{female, part time} = \frac{63}{180} \times 100 = 0.35 \times 100 = 35\%.$$

Cont...

This tells us that of our sample of 40,

50% should be male, full time. 50% of 40 is 20.

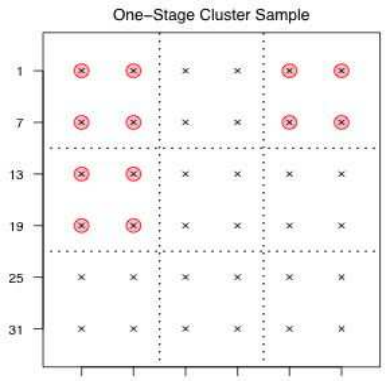
10% should be male, part time. 10% of 40 is 4.

5% should be female, full time. 5% of 40 is 2.

35% should be female, part time. 35% of 40 is 14.

Cluster sample

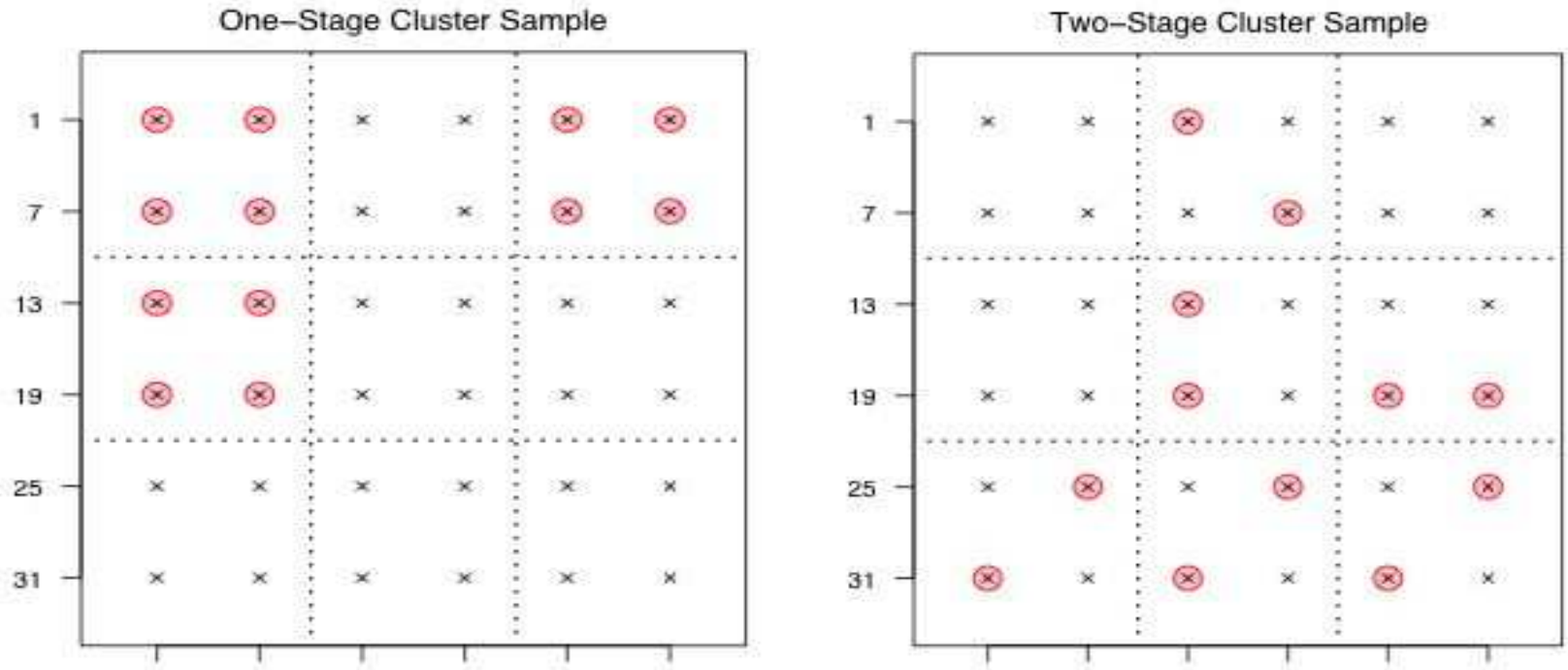
- The entire population of interest is divided into groups, or clusters, and a random sample of these clusters is selected.
- Each cluster must be mutually exclusive and together the clusters must include the entire population.
- After clusters are selected, then all units within the clusters are selected.
- This differs from stratified sampling, in which some units are selected from each group.



Cont...

- When all the units within a cluster are selected, the technique is referred to as **one-stage cluster** sampling.
- If a subset of units is selected randomly from each selected cluster, it is called **two-stage cluster** sampling.
- Cluster sampling can also be made in three or more stages: it is then referred to as **multistage cluster** sampling.

Cont...



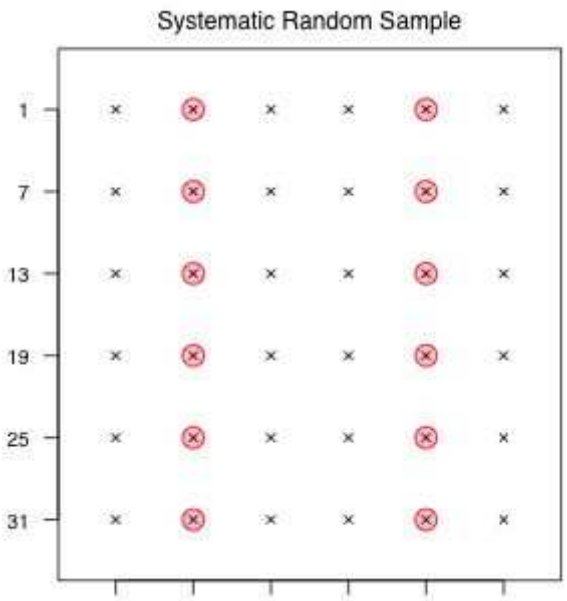
(a) (b)

Figure: One and two stage cluster sampling.

Example

Suppose we want to measure the Mathematics knowledge of grade 5 students. If we compare two grade 5 classes in Sri Lanka, the knowledge of students of two such classes may not be much different. However the knowledge of students within a class may var. The one such class (similar subgroup) is termed as a cluster. By investigating a few cluster, we can estimate the knowledge of students in the whole population.

Systematic sample



- Methodology for sampling in which units are selected from the population at a regular interval.
- First, the statistician has to select an integer k, which is approximately equal to the ratio of population size and sample size.

Cont...

- If the population size is unknown we can guess k. Guessing the value of k does not have much effect on sampling.
- Now select any integer i from 1 to k.
- Then select i^{th} unit of the population to the sample.
- Thereafter select every k^{th} unit till we select the desired sample size.
- Thus the sample contains $i, i+k, i+2k, ..., i+(n-1)k$ units of the above serial numbers.

Example

If there are 120 names on the list. How we obtain a systematic sample of 20 names?

Solution
Let $k = 120/20 = 6$.

Therefore, we would randomly select a number between 1 and 6.

Primary unit selected $= 2$.

After selecting this primary unit, we would include every 6^{th} unit in the sample.

Secondary units in the sample are
$8, 14, 20, 26, 32, 38, 44, 50, 56, 62, 68, 74, 80, 86, 92, 98, 104, 110, 116.$

Advantage and disadvantage

Advantage

- The principal advantage of this technique is its simplicity.

Disadvantage

- This sampling technique is not necessary if the population items are naturally arranged in a periodic order.

Multistage sampling techniques

- Given a population it is possible to select a sample at one stage or at number of stages.
- In one stage sampling we directly get the measurements/observations from the selected sample.
- In the second stage sampling technique we select a sample from the sample we selected at the first stage.

Example

- Suppose we have selected cluster sample at the first stage.
- If it is not possible to study the whole cluster (due to lack of resource/time limits), we can select another sample as second stage sampling by using an appropriate sampling technique.
- This sampling scheme is referred to as two stage sampling. The process can be extended to multistage sampling.

Summarizing data

Summarizing data

- It is very important to present collected data in summarized form.
- In the case of summarizing numerical data, frequency tables, relative frequency tables, histograms, frequency polygons are used.
- For summarizing non numerical data we use frequency tables, bar charts, pie charts etc.

Frequency tables

- The **frequency** of a particular data value is the number of times the data value occurs.
- A **frequency table** is constructed by arranging collected data values in ascending order of magnitude with their corresponding frequencies.

Example

The marks awarded for an assignment set for a Year 8 class of 20 students were as follows:

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

Present this information in a frequency table.

Cont...

Solution:

Mark	Tally	Frequency
4		
5		
6		
7		
8		
9		
10		

Figure: Frequency table

Cont...

Solution:

Mark	Tally	Frequency
4	||	2
5	||	2
6	||||	4
7	$\cancel{\text{IIII}}$	5
8	||||	4
9	||	2
10	|	1

Figure: Frequency table

Frequency tables with class intervals

- When the set of data values are spread out, there will be too many rows in the table.
- So we group the data into class intervals.
- The frequency of a group is the number of data values that fall in the range specified by that group.
- Ideally, we should have between five and ten rows in a frequency table.
- Some statistician determines number of class k as the smallest integer such that $2^k \geq n$, where n is the sample size.

Cont...

$$
\begin{aligned}
n &= 108 \\
2^k &\geq 108 \\
2^6 = 64,\ 2^7 &= 128 \\
\text{Therefore } k &= 7 \\
\text{Range} &= \text{Largest value-Smallest value} \\
\textbf{Class width} &\simeq \frac{\text{Range}}{k}
\end{aligned}
$$

Class limits, Boundaries and Intervals

- **Class limits** are the smallest and largest observations in each class. Therefore, each class has two limits: a lower and upper.
- **Class Boundaries** are the midpoints between the upper class limit of a class and the lower class limit of the next class in the sequence. Therefore, each class has an upper and lower class boundary.
- **Class interval** is the difference between the upper and lower class boundaries of any class.

Example 1

Class	Frequency
200 - 299	12
300 - 399	19
400 - 499	6
500 - 599	2
600 - 699	11
700 - 799	7
800 - 899	3
Total Frequency	60

(a) What are the lower and upper class limits for the second class?

(b) Determine the class boundaries of the second classes.

(c) Determine the class intervals for the first class.

Cont...

(a) The lower class limit is 300.

The upper class limit is 399.

(b) The lower class boundary is the midpoint between 299 and 300, that is 299.5.

The upper class boundary is the midpoint between 399 and 400, that is 399.5.

Cont...

(c) The first class is 200-299.

$$\begin{aligned}
\text{The class interval} &= \text{Upper class boundary-lower class boundary} \\
\text{Upper class boundary} &= 299.5 \\
\text{Lower class boundary} &= 199.5 \\
\text{Therefore, the class interval} &= 299.5 - 199.5 = 100.
\end{aligned}$$

Example 2

The 48 buses of the Cyclone Transport Bord of Matara obtain their weekly fuel consumption in liters as below. Represent the data in group frequency table.

72.17 28.95 37.87 69.49 37.51 20.11 44.63 43.57 24.75 52.83
34.69 31.22 33.21 18.75 38.67 53.41 41.88 41.35 49.30 42.70
42.45 60.75 30.24 26.27 37.80 33.80 31.55 21.45 15.25 36.00
36.45 71.88 50.55 47.82 27.63 38.76 22.16 33.68 64.50 40.58
24.65 25.68 20.45 56.13 39.01 30.56 45.14 23.65

Cont...

Smallest value $= 15.25$

Largest value $= 72.17$

$n = 48$

$$
\begin{aligned}
2^k \geq n &= 48 \\
k &= 6 \\
\text{Range} &= 72.17 - 15.25 \\
&= 56.92 \\
\text{Class width} &= \frac{\text{Range}}{k} = \frac{56.92}{6} \\
&\simeq 9.32 = 10
\end{aligned}
$$

Cont...

So we can select class intervals as

$$15.00 - 24.99$$
$$25.00 - 34.99$$
$$.$$
$$.$$
$$.$$
$$65.00 - 74.99$$

Then $15.00, 25.00, ..., 65.00$ are termed as lower class limits and $24.99, 34.99, ..., 74.99$ are termed as upper class limits.

Cont...

However we prefer to have continuous intervals with no gaps. Thus select the class intervals as

$$15.00- < 25.00$$
$$25.00- < 35.00$$
$$35.00- < 45.00$$
$$45.00- < 55.00$$
$$55.00- < 65.00$$
$$65.00- < 75.00$$

Then $15.00, 25.00, ..., 65.00$ are termed as lower class boundaries and $35.00, 45.00, ..., 75.00$ are termed as upper class boundaries.

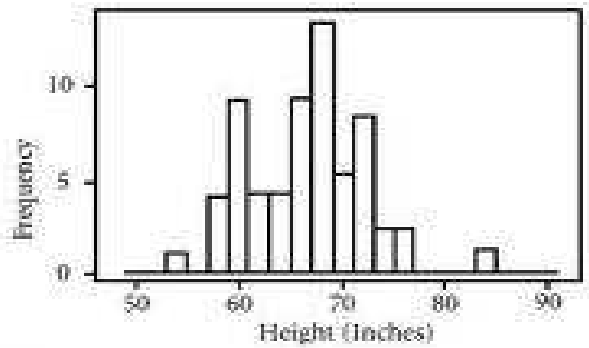
Cont...

Class intervals	Mid point	Frequency
15.00-<25.00	20	9
25.00-<35.00	30	13
35.00-<45.00	40	14
45.00-<55.00	50	6
55.00-<65.00	60	3
65.00-<75.00	70	3

Histogram

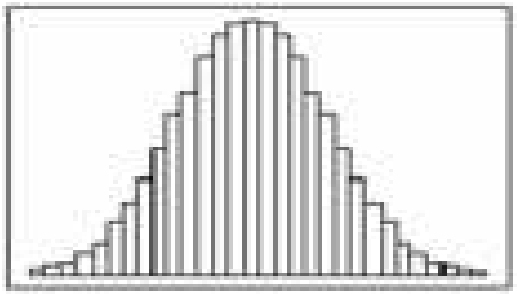
The histogram graphically shows the following:

- Center (i.e., the location) of the data;
- Spread (i.e., the scale) of the data;
- Skewness of the data;
- Presence of outliers; and
- Presence of multiple modes in the data.

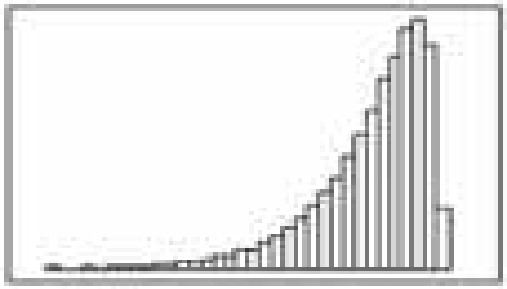


Skewness

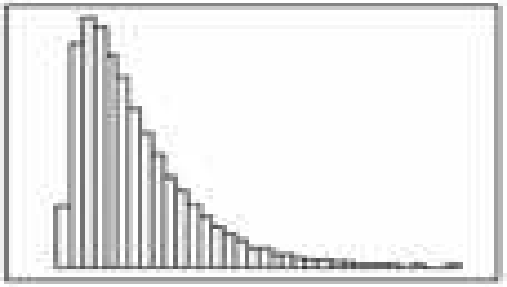
- The shape of the histrogram shows the shape of the distribution.



Symmetric
Bell shaped

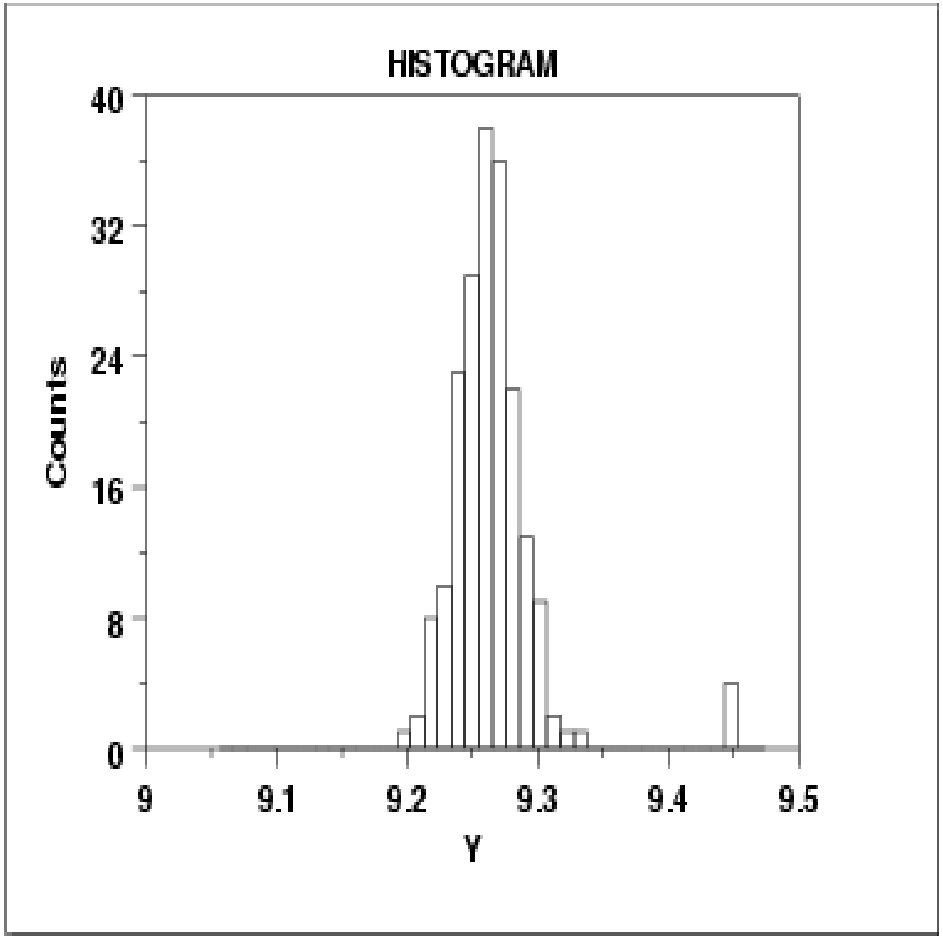


Skewed to
the Left



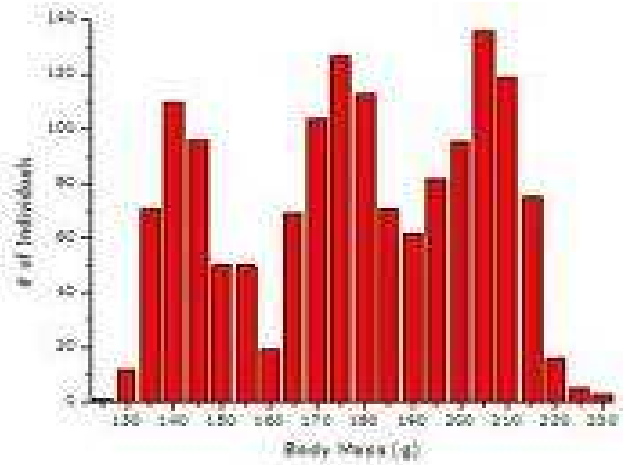
Skewed to
the Right

Outliers



Multiple modes

- When the histrogram shows more than one clear peaks, we assume that data have come from a multimodel population.
- Concentration of data can be seen with higher rectangles.



Draw histrogram

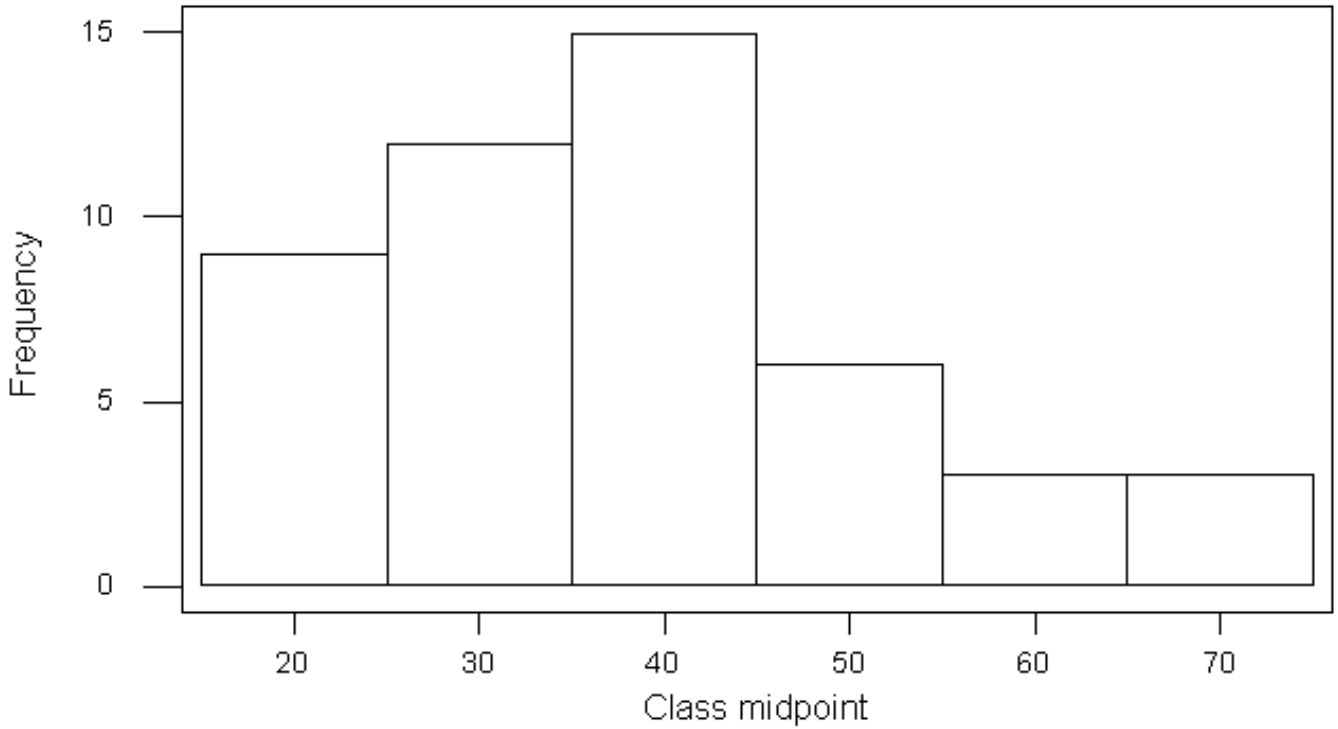
- The class mid points are marked on the horizontal axis.
- Rectangles are drawn such that the area propotional to the class frequencies.

Example

Draw histrogram for following data.

Class intervals	Mid point	Frequency
15.00-<25.00	20	9
25.00-<35.00	30	13
35.00-<45.00	40	14
45.00-<55.00	50	6
55.00-<65.00	60	3
65.00-<75.00	70	3

Solution



Relative frequency

- If the sample sizes are different, it is not sensible to compare the frequencies of falling into different categories. Thus we have to consider relative frequencies.

- The ratio of the observed frequency of some outcome and the total frequency of the random experiment is termed as relative frequency.

Cont...

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Total no. of observations}}$$

- The only difference between a frequency histogram and a relative frequency histogram is that the vertical axis uses relative frequency instead of frequency.

Example

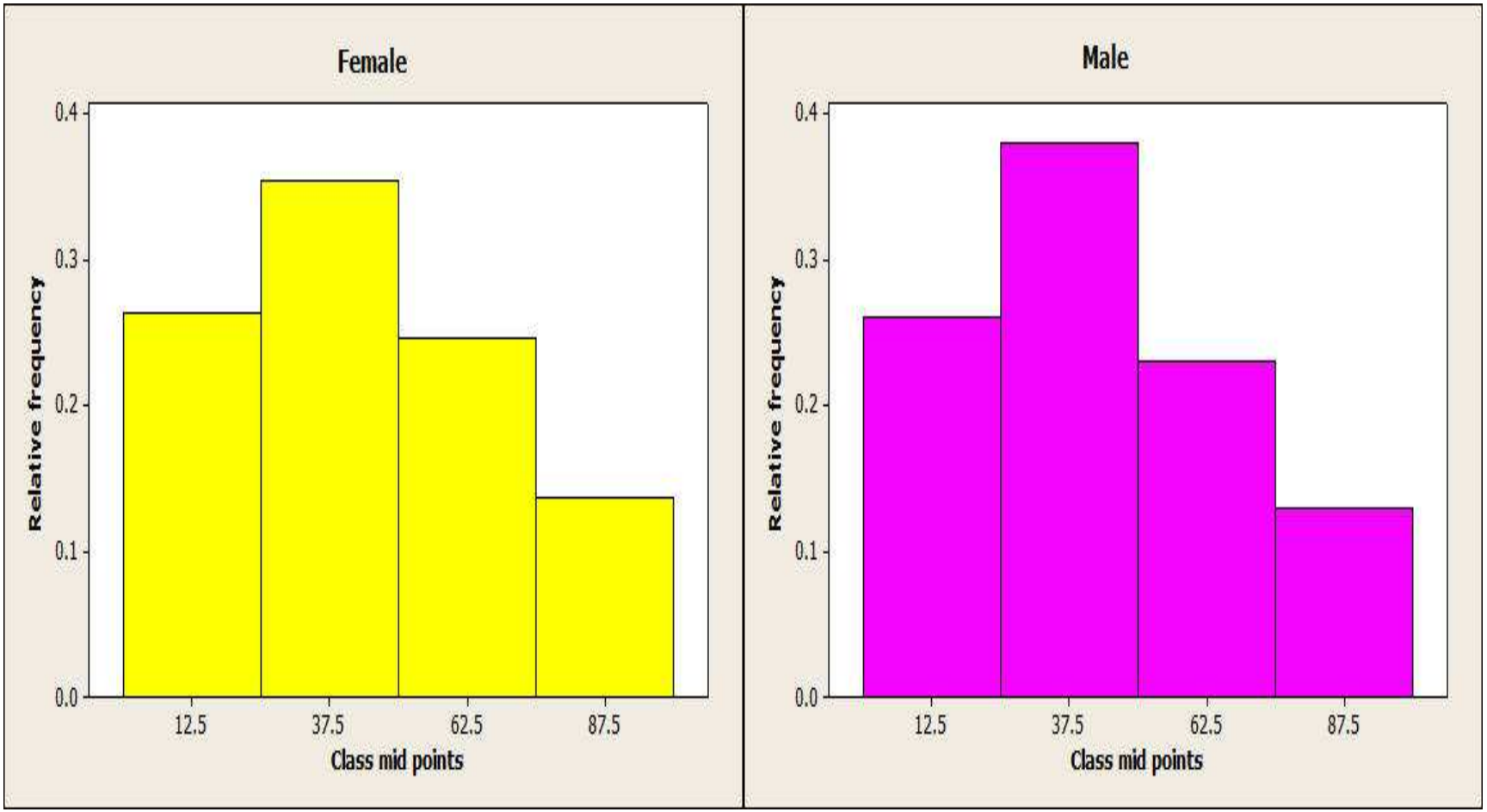
For introductory statistic course unit 100 male students and 175 female students were participated. Their marks distribution is given in the following table. Draw relative frequency histogram separately for male and female students.

Marks	Frequency of male	Frequency of female
0-<25	26	46
25-<50	38	62
50-<75	23	43
75-<100	13	24

Cont...

Marks	R.F of males	R.F of females
0-<25	0.26	0.263
25-<50	0.38	0.354
50-<75	0.23	0.246
75-<100	0.13	0.137

Cont...


Female
Relative frequency
0.4
0.3
0.2
0.1
0.0
12.5
37.5
62.5
87.5
Class mid points
Male
Relative frequency
0.4
0.3
0.2
0.1
0.0
12.5
37.5
62.5
87.5
Class mid points

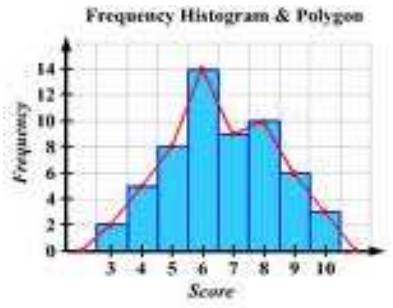
Frequency polygon

The frequency polygon is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class frequencies.

Example

For the set of statistical data draw a frequency histogram and polygon on the same set of axis.

Score	Frequency
3	2
4	5
5	8
6	14
7	9
8	10
9	6
10	3



Remark

- Frequecy polygon and histrogram usually display similar information regarding symmetry, skewness, spread, concentration of data and peak point.
- The concentration of data can be seen with peaks of frequency polygons.
- When there are more than one peaks, the data have come from multimodal population.
- The width of the polygon indicates the spread of the data.

Cumulative frequency

- Cummulative frequency distributions are useful in conveying information about frequency of observations, that are below (or above) a specified level of the response variable.
- **Ogive** is the cummulative frequency polygon.
- There we plot the cummulative frequencies against the left class boundaries.

Draw the cumulative frequency polygon

- The cummulative frequency of the first left class boundary is considered as zero.
- The cummulative frequency for any other left class boundary is taken as the cummulative frequency of immediate precceding class.

Example

Summarize the following height data in cumulative frequency table. Draw the cumulative frequency polygon.

Height (cm)	Frequency
$150 \leq h < 155$	4
$155 \leq h < 160$	22
$160 \leq h < 165$	56
$165 \leq h < 170$	32
$170 \leq h < 175$	5

Cont...

Height (cm)	Cumulative frequency
under 150	0
under 155	0+4
under 160	4+22
under 165	26+56
under 170	82+32
under 175	114+5

Cont...

- From the data table we can see that there are no heights under 150 cm.
- Under 155 cm there are the first 4 height.
- Under 160 cm there are the first 4 height plus a father 22 height that are between 155 cm and 160 cm, giving 26 altogether.
- The cumulative frequency graph can now be plotted using the point in the table,
$(150, 0), (155, 4), (160, 26), (165, 82), (170, 114), (175, 119)$.

Graphical representation of non numerical data

For summarizing non numerical data we use bar charts, pie charts etc.

Bar chart

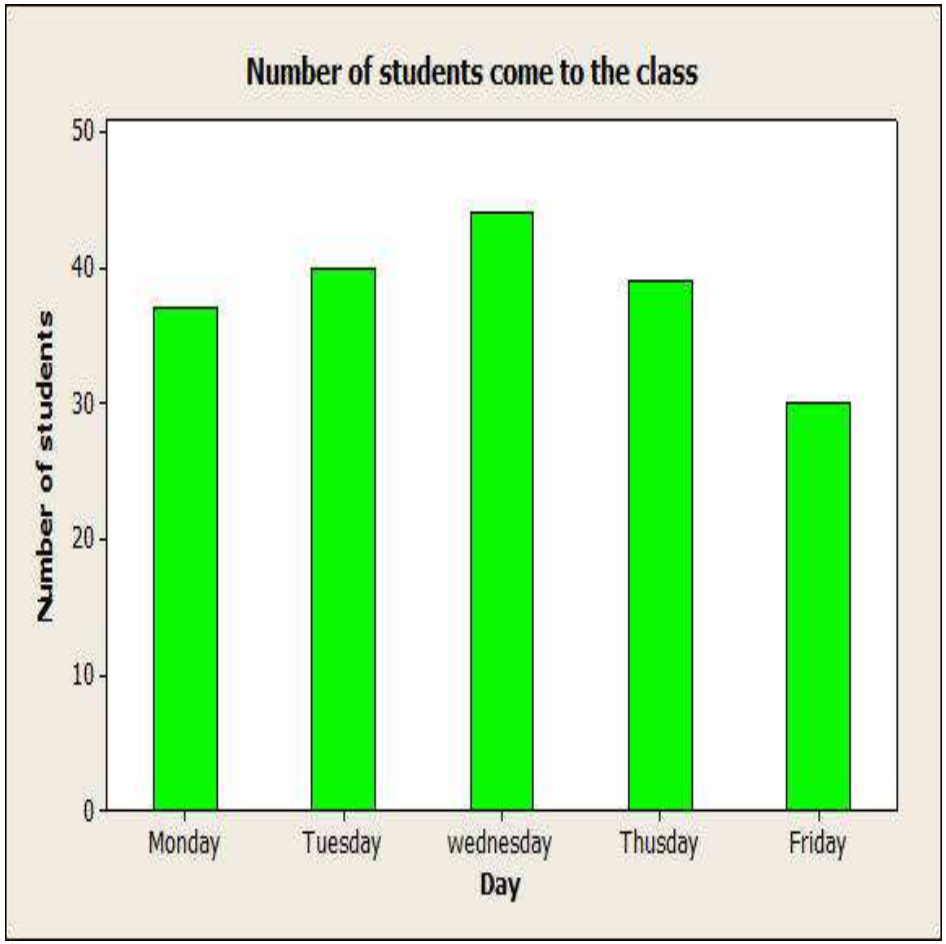
- A bar shows a category.
- The length of a bar represents the amount, frequency or percentage of values falling into a category.
- **Grouped bar graphs** present bars clustered in groups of more than one.
- **Stacked bar graphs** show the bars divided into subparts to show cumulate effect.

Bar chart

The following table presents the number of students come to a class from Monday to Friday on a perticular week. Represent the data graphically using a bar chart.

Day	Number of Students
Monday	37
Tuesday	40
wednesday	44
Thusday	39
Friday	30

Solution


Number of students come to the class
Number of students
50
40
30
20
10
0
Monday
Tuesday
wednesday
Thusday
Friday
Day

Group bar chart

Suppose we have recorded the numbers of boys and girls who come to the class from Monady to Friday on a perticular week. Represent the data graphically using a grouped bar chart.

Day	Girls	Boys
Monday	19	18
Tuesday	20	20
wednesday	24	20
Thusday	19	20
Friday	14	16

Solution

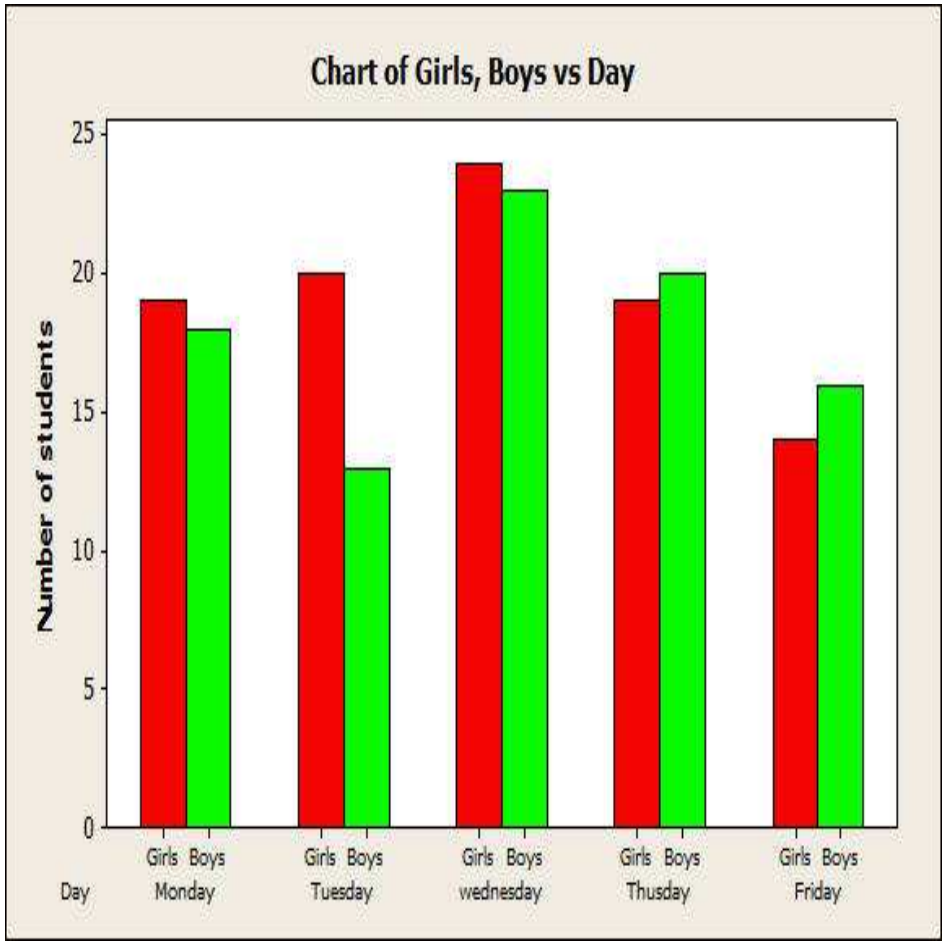


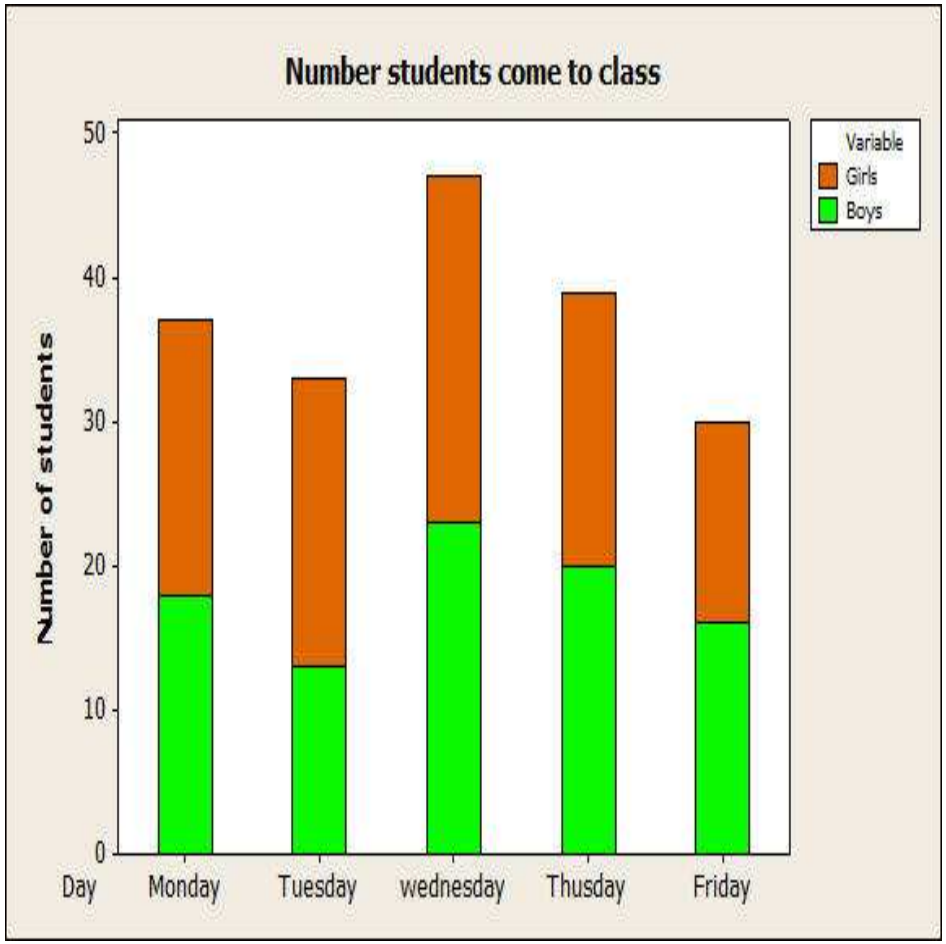
Chart of Girls, Boys vs Day
Number of students
25
20
15
10
5
0
Girls Boys
Day
Monday
Girls Boys
Tuesday
Girls Boys
wednesday
Girls Boys
Thusday
Girls Boys
Friday

Stack bar chart

Suppose we have recorded the numbers of boys and girls who come to the class from Monady to Friday on a perticular week. Represent the data graphically using a stack bar chart.

Day	Girls	Boys
Monday	19	18
Tuesday	20	20
wednesday	24	20
Thusday	19	20
Friday	14	16

Solution


Number students come to class
Number of students
50
40
30
20
10
0
Day
Monday
Tuesday
wednesday
Thusday
Friday
Variable
Girls
Boys

Pie chart

- A pie chart is a circular chart divided into sectors, illustrating proportion.
- In a pie chart, the arc length of each sector is proportional to the quantity it represents.

Example

Draw a pie chart to display the information regarding the expenses of a hospital.

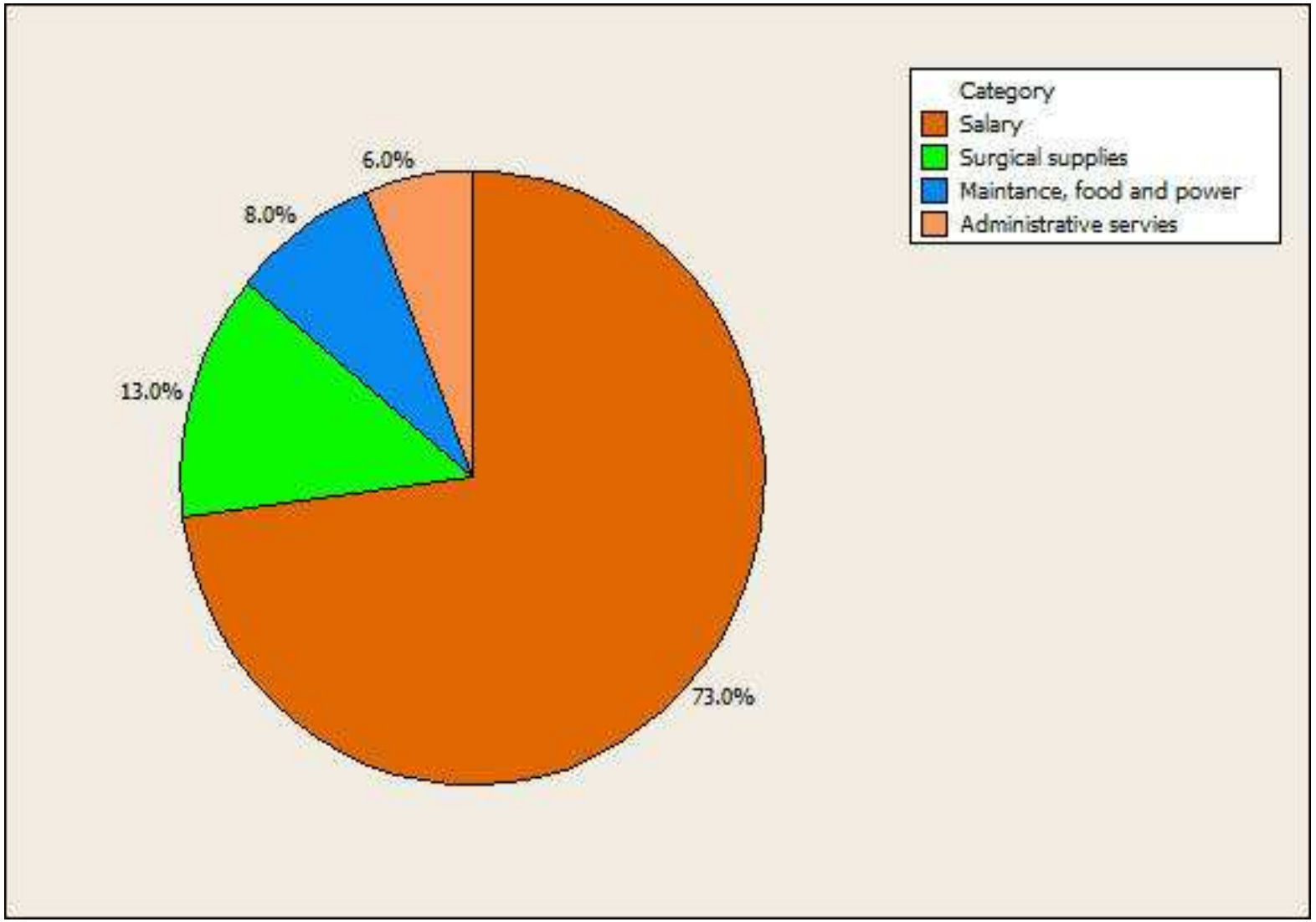
For salary	–	73%
Medical and surgical supplies	–	13%
Maintance, food and power	–	8%
Administrative services	–	6%

Solution

To construct the pie chart, we need to find the corresponding angles.

$$\begin{aligned}
\frac{73}{100} \times 360 &\simeq 263^\circ \Leftarrow \text{Salary} \\
\frac{13}{100} \times 360 &\simeq 47^\circ \Leftarrow \text{Medical} \\
\frac{8}{100} \times 360 &\simeq 29^\circ \Leftarrow \text{Maintanance} \\
\frac{6}{100} \times 360 &\simeq 21^\circ \Leftarrow \text{Admin}
\end{aligned}$$

Cont...


Category
Salary
Surgical supplies
Maintance, food and power
Administrative servies
6.0%
8.0%
13.0%
73.0%

Thanks