Applied Statistics I

(IMT224β/AMT224β)

Department of Mathematics
University of Ruhuna

A.W.L. Pubudu Thilan

Chapter 2

Summarizing data

Summarizing numerical data

It is very important to present collected data in summarized form. In the case of summarizing numerical data, we can use,

1. Frequency tables
2. Frequency tables with class intervals
3. Relative frequency tables
4. Histograms
5. Relative frequency histograms
6. Frequency polygons
7. Cumulative frequency

[1] 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.

[1] Frequency tables

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.

[1] Frequency tables

Example⇒Solution

Mark	Tally	Frequency
4		
5		
6		
7		
8		
9		
10		

Figure: Frequency table

[1] Frequency tables

Example⇒Solution⇒Cont...

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

Figure: Frequency table

[1] Frequency tables

Computer exercise

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 using Minitab statistical software package.

[2] 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.

[2] Frequency tables with class intervals

Computer exercise

The marks of 200 students are given below. Can we draw a frequency table to summarize these data?

98 7 18 23 42 33 2 81 51 80 7 57 68 97 8 65 17 42 27 63 81 23 75
33 97 41 72 95 77 10 73 61 91 66 49 13 5 53 12 9 16 85 88 19 60
30 55 42 17 14 43 61 60 85 22 97 3 11 24 56 32 47 72 53 19 70
77 57 95 68 12 32 70 49 61 98 50 65 46 36 47 32 99 36 73 90 42
69 17 73 99 53 47 33 63 81 40 21 38 80 11 79 81 81 2 16 94 52 0
19 21 11 27 99 51 71 47 54 90 81 75 35 72 39 84 81 68 60 37 96
32 55 89 3 80 60 99 69 64 17 36 22 28 46 15 67 61 48 18 9 28 48
29 69 48 30 13 52 28 56 81 34 59 3 95 26 15 24 36 83 49 51 16
39 93 91 34 78 73 64 7 21 59 68 97 99 27 8 19 71 96 53 54 48 29
23 39 9 73 76

[2] Frequency tables with class intervals

How to determine number of classes?

Some statistician determines number of class k as the smallest integer such that $2^k \geq n$, where n is the sample size.

Eg: Suppose $n = 108$. Determine the value of k.

$$
\begin{aligned}
n &= 108 \\
2^k &\geq 108 \\
k = 6 \Rightarrow 2^6 &\geq 108 \Rightarrow 64 \geq 108 \Rightarrow \text{inequality is wrong} \\
k = 7 \Rightarrow 2^7 &\geq 108 \Rightarrow 128 \geq 108 \Rightarrow \text{inequality is OK} \\
k = 8 \Rightarrow 2^8 &\geq 108 \Rightarrow 256 \geq 108 \Rightarrow \text{inequality is OK} \\
k &= 7 \text{ (smallest value satisfying } 2^k \geq n) \\
\text{Range} &= \text{Largest value-Smallest value} \\
\textbf{Class width} &\simeq \frac{\text{Range}}{k}
\end{aligned}
$$

- **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 width** is the difference between the upper and lower class boundaries of any class.

[2] Frequency tables with class intervals

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 class.

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

(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.

(c) The first class is 200-299.

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

[2] Frequency tables with class intervals

Example 2

The 48 buses of the Ceylon Transport Board 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

[2] Frequency tables with class intervals

Example 2⇒Solution

Smallest value $= 15.25$

Largest value $= 72.17$

$n = 48$

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

[2] Frequency tables with class intervals

Example 2⇒Solution⇒Cont...

So we can select class intervals as:

$$15.00 - 24.99$$
$$25.00 - 34.99$$
$$\vdots$$
$$65.00 - 74.99$$

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

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.

[2] Frequency tables with class intervals

Example 2⇒Solution⇒Cont...

Class intervals	Tally	Frequency
15.00-<25.00		
25.00-<35.00		
35.00-<45.00		
45.00-<55.00		
55.00-<65.00		
65.00-<75.00		

[2] Frequency tables with class intervals

Example 2⇒Solution⇒Cont...

Class intervals	Frequency
15.00-<25.00	9
25.00-<35.00	13
35.00-<45.00	14
45.00-<55.00	6
55.00-<65.00	3
65.00-<75.00	3

[3] Relative frequency tables

- 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.

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

[3] Relative frequency tables

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 table 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

[3] Relative frequency tables

Example⇒Solution

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

[4] Histogram

- A histogram is a graphical display of data using rectangles of different heights.
- The class mid points (also you can use class boundaries) are marked on the horizontal axis.
- The rectangles are drawn such that the area proportional to the class frequencies.
- The rectangles of a histogram are drawn so that they touch each other to indicate that the original variable is continuous.

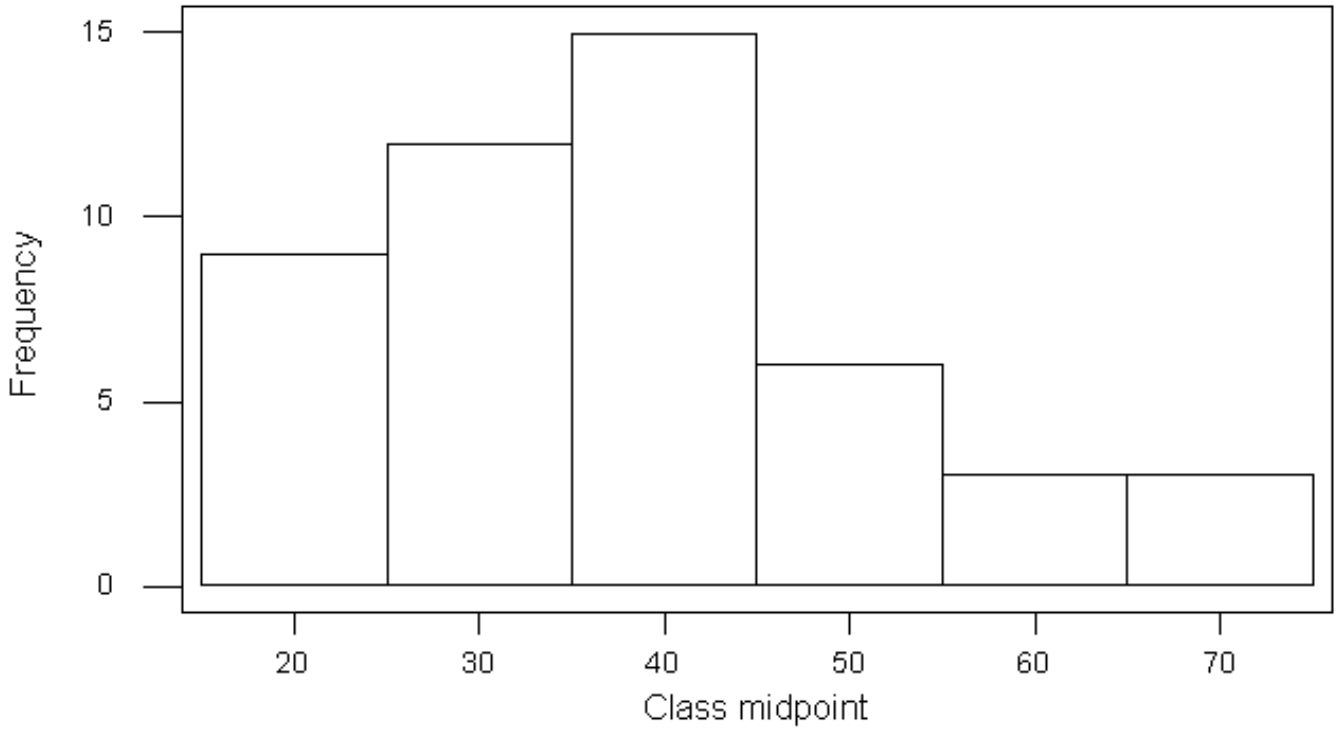
[4] Histogram
Example

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

[4] Histogram

Example⇒Solution



[5] Relative frequency histograms

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

[5] Relative frequency histograms

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 histograms 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

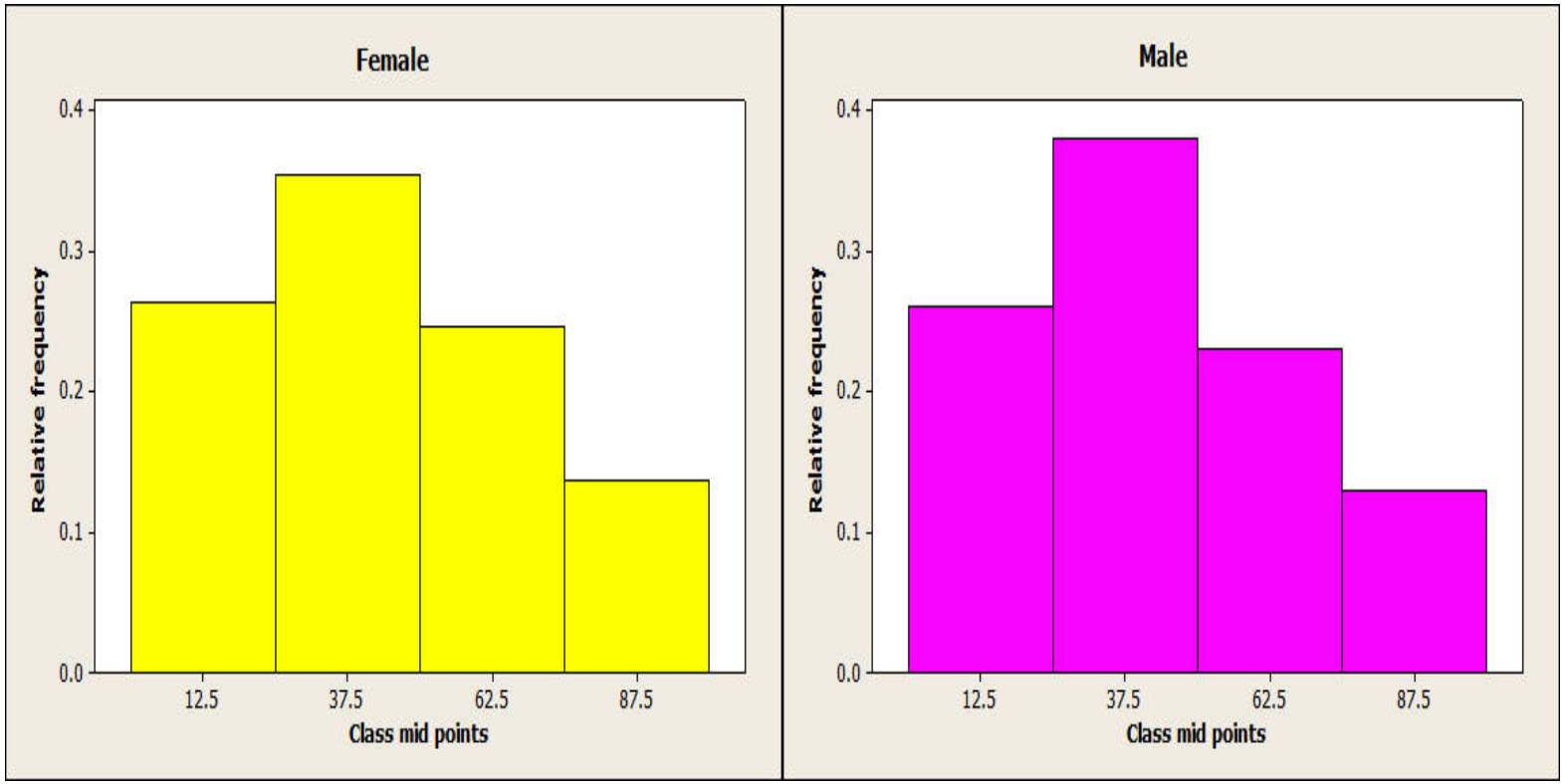
[5] Relative frequency histograms

Example⇒Solution

Marks	Mid point	R.F of males	R.F of females
0-<25	12.5	0.26	0.263
25-<50	37.5	0.38	0.354
50-<75	62.5	0.23	0.246
75-<100	87.5	0.13	0.137

[5] Relative frequency histograms

Example⇒Solution⇒Cont...



[6] Frequency polygon

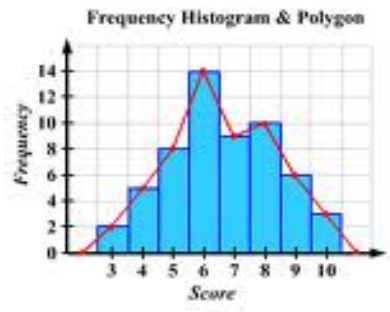
The frequency polygon is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class frequencies.

[6] Frequency polygon

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



[7] Cumulative frequency

- Cumulative 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 cumulative frequency polygon.
- There we plot the cumulative frequencies against the left class boundaries.

[7] Cumulative frequency

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.

[7] Cumulative frequency

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

[7] Cumulative frequency

Example

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

[7] Cumulative frequency

Example⇒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,
  $(150, 0), (155, 4), (160, 26), (165, 82), (170, 114), (175, 119)$.

Summarizing non numerical data

To summarize non numerical data we can use,

1. Frequency tables
2. Bar charts
3. Grouped bar charts
4. Stacked bar charts
5. Pie charts

The following cars were seen in a park (B=British, F=French, G=German, I=Italian).

F B B B G I I B F F B F G F F I B B G B I B B G I B G B G B F B

Construct a frequency table for these data.

[1] Frequency tables

Example⇒Solution

Type of car	Frequency
British	14
French	7
German	6
Italian	5

[2] Bar chart

- A bar shows a category.
- The length of a bar represents the amount, frequency or percentage of values falling into a category.

[2] Bar chart

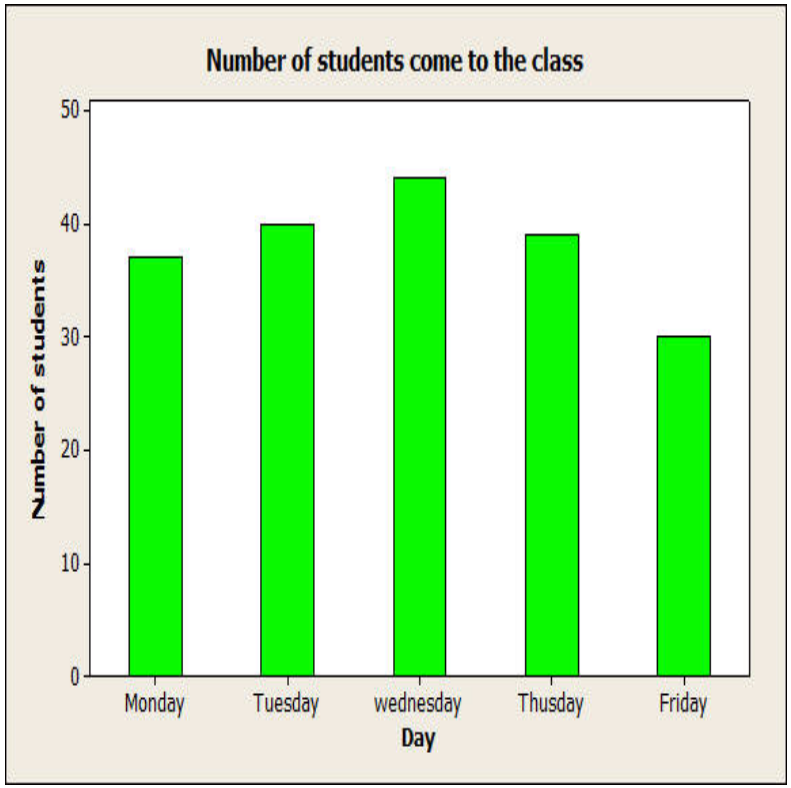
Example

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

Day	Number of Students
Monday	37
Tuesday	40
Wednesday	44
Thursday	39
Friday	30

[2] Bar chart

Example⇒Solution



[3] Grouped bar chart

- It is also known as clustered bar chart.
- Graphing data organized by two grouping variables.
- Grouped bar chart presents bars clustered in groups of more than one.

[3] Grouped bar chart

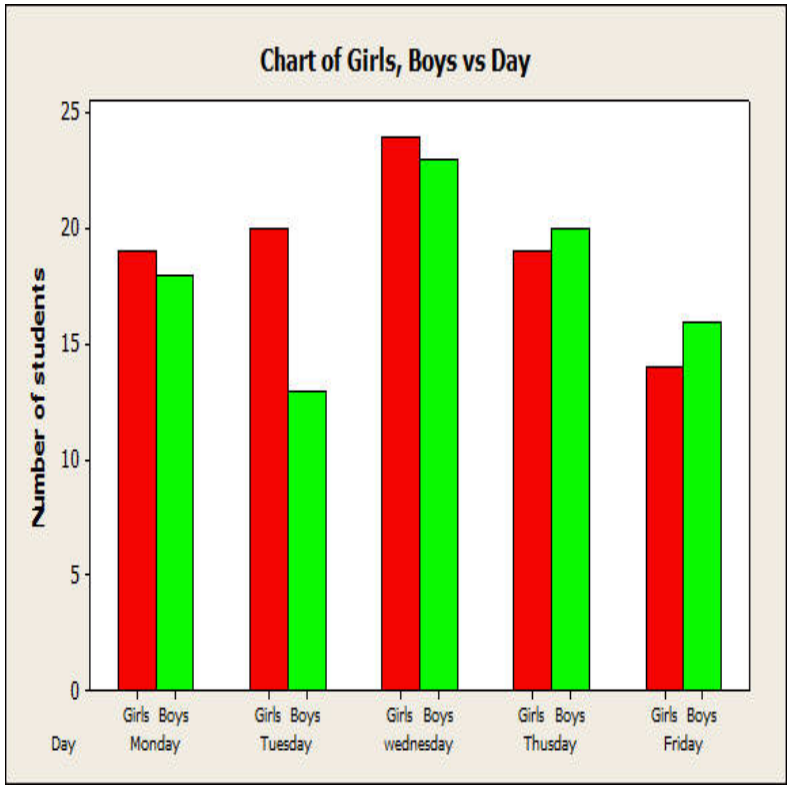
Example

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

Day	Girls	Boys
Monday	19	18
Tuesday	20	20
Wednesday	24	20
Thursday	19	20
Friday	14	16

[3] Grouped bar chart

Example⇒Solution



[4] Stacked bar chart

- Stacked bar graph is a graph that is used to compare the parts to the whole.
- The bars in a stacked bar graph are divided into categories.
- Each bar represents a total.

[4] Stacked bar chart

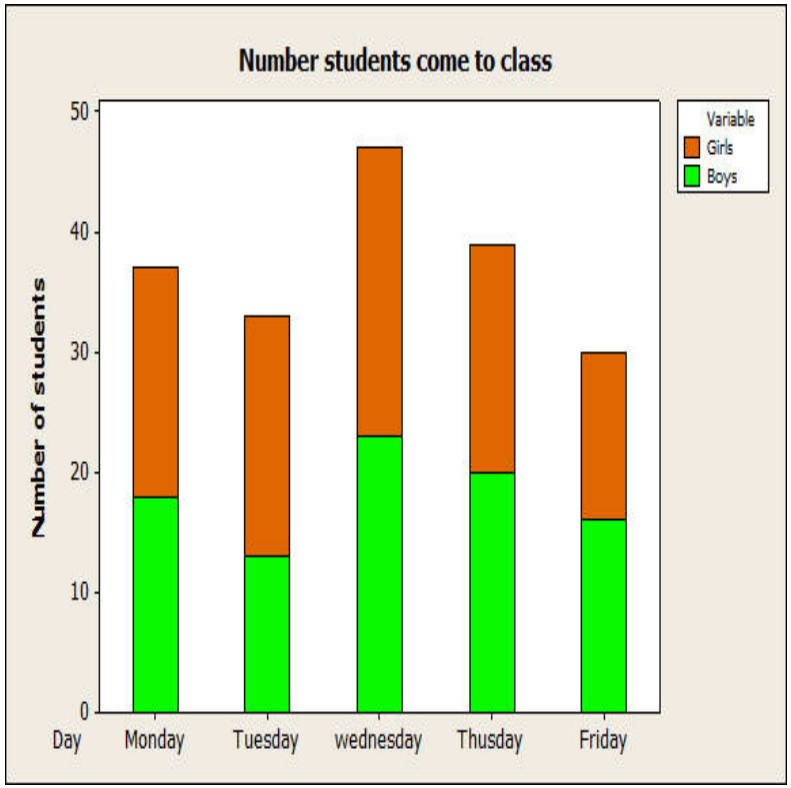
Example

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

Day	Girls	Boys
Monday	19	18
Tuesday	20	20
Wednesday	24	20
Thursday	19	20
Friday	14	16

[4] Stacked bar chart

Example⇒Solution



[5] 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.

[5] Pie chart

Example

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

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

[5] Pie chart

Example⇒Solution

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

$$\frac{73}{100} \times 360 \simeq 263^\circ \Leftarrow \text{Salary}$$

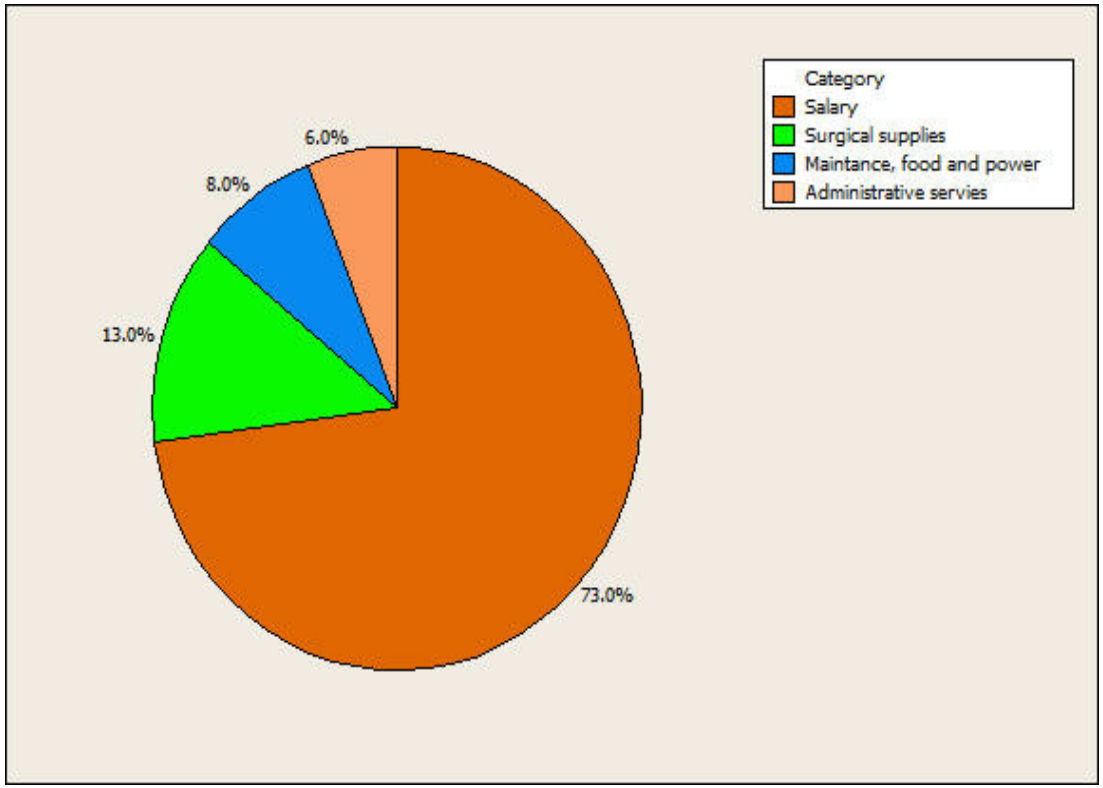
$$\frac{13}{100} \times 360 \simeq 47^\circ \Leftarrow \text{Medical}$$

$$\frac{8}{100} \times 360 \simeq 29^\circ \Leftarrow \text{Maintenance}$$

$$\frac{6}{100} \times 360 \simeq 21^\circ \Leftarrow \text{Admin}$$

[5] Pie chart

Example⇒Solution⇒Cont...



Thank You