# Graphic presentation (histogram, polygon and pie chart)

## Graphs for qualitative and discrete data

Generally discrete quantitative or qualitative data are represented by bar chart, line graph and pie diagram.

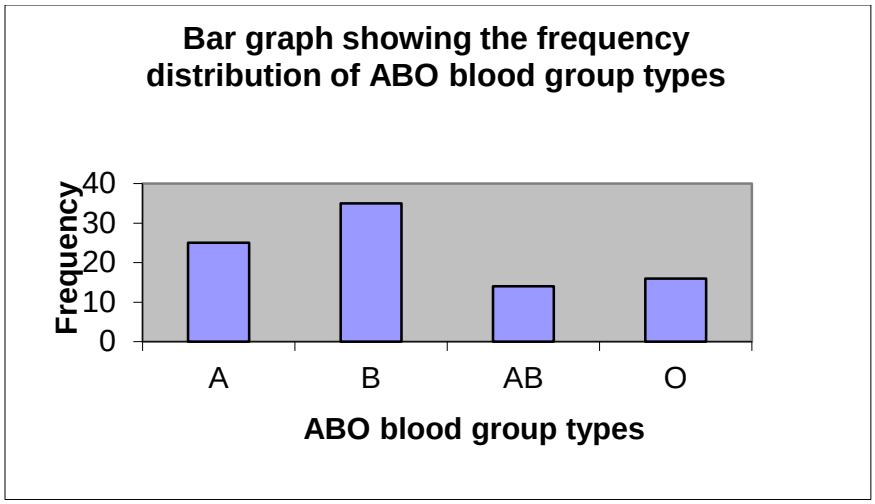
The most straightforward type of graph to produce is a bar chart.

Bar chart: In bar chart, rectangular blocks of equal width are plotted on the X axis, each representing independent variable placed at equal distance from each other. The height of each block represents the frequency of the categories and is proportional to the number of percentage in each category. Because each bar represents a completely separate category, the bars must not touch each other, i.e. always leave a gap between the categories. There is no hard and fast rule for ordering the bars; but some category types (measured in ordinal scale) have a natural ordering, like educational level and monthly income where the bars can be plotted in a logical order from lowest to highest or vis-à-vis, irrespective of the height of the bars.

**Table 1**

| Blood group types | Frequency |
|---|---|
| A | 25 |
| B | 35 |
| AB | 14 |
| O | 16 |

The figure below is the graphical presentation of the table

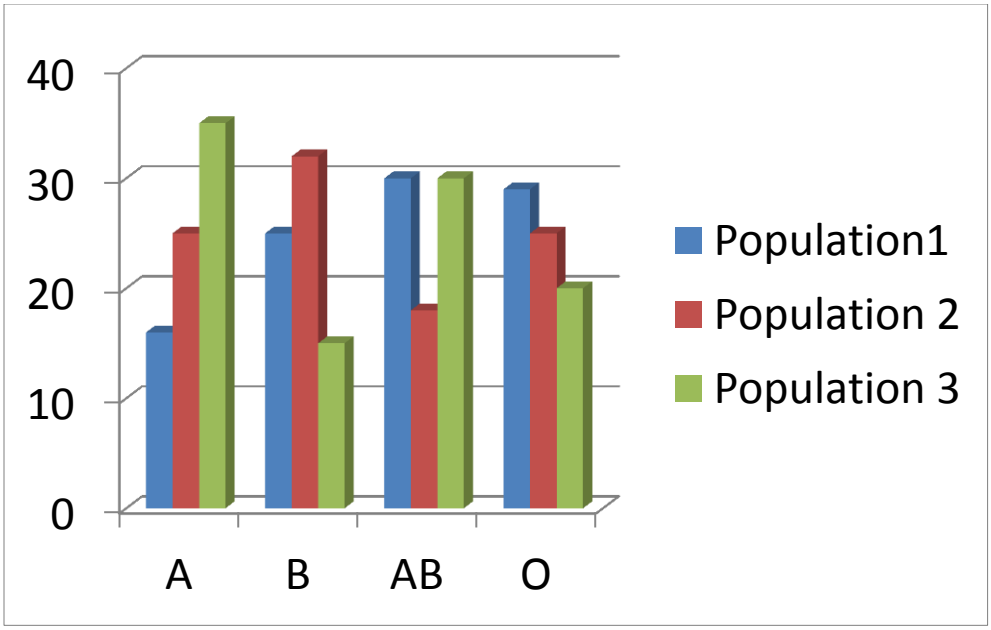



**Multiple bar charts**

Multiple bar charts help to present a comparative view of groups or populations. Suppose you are comparing the frequency of ABO blood groups of two or more population. In such a case, multiple bar charts are useful. You can easily plot two or more bars adjacently, representing the frequency of two or more population against each of the blood group types. But one should convert the actual frequencies to proportion or percentage before comparing two or more groups. This will help to compare data set of unequal sample size.

**Table 2**

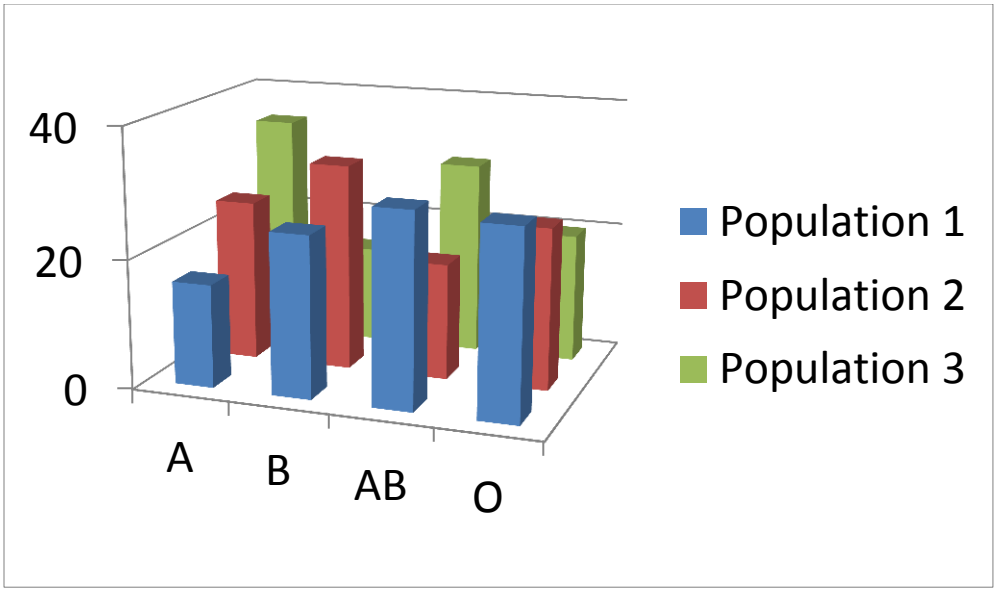
| Blood group types | Population 1 % | Population 2 % | Population 3 % |
|---|---|---|---|
| A | 16 | 25 | 35 |
| B | 25 | 32 | 15 |
| AB | 30 | 18 | 30 |
| O | 29 | 25 | 20 |




How will you interpret this? You have to say that the frequency of A blood group is the highest in Population 3, followed by that of Population 2 and 3. For blood group B, population 2 has the highest frequency, followed by that of Population 1 and 3. For blood group AB, the frequency is the same and highest for Population 1 and 3. Lastly, for blood group O, the frequency is the highest for Population 1, followed by Population 2 and 3.

Or

Using same data set you can make a three dimensional graphical presentation. Here, the bars representing each of the population against a blood group type is not touching with other.

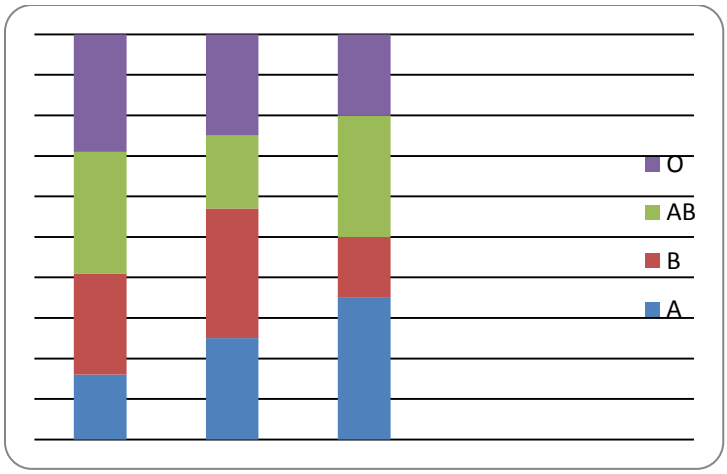



**Stacked bar charts**

You can also make a graphical presentation of the same data using stacked bars where categories make up part of a whole. It makes easier to see what proportion of one category is of the whole. For example, here it will be what proportion of individuals in Population 1, have A, B, AB and O blood groups of a whole. The same can be demonstrated for Population 2 and 3 independently. Again, one can make a visual comparison in the frequency of each of the blood group categories across three populations. For example, the frequency of A and O blood group categories increases and decreases from Population 1 through 3 respectively. One can put the labels in the form of legends, or each section can be labeled at the side of one of the bars.

**Table 3**

| Blood group types | Population 1 | Population 2 | Population 3 |
|---|---|---|---|
| | % | % | % |
| A | 16 | 25 | 35 |
| B | 25 | 32 | 15 |
| AB | 30 | 18 | 30 |
| O | 29 | 25 | 20 |



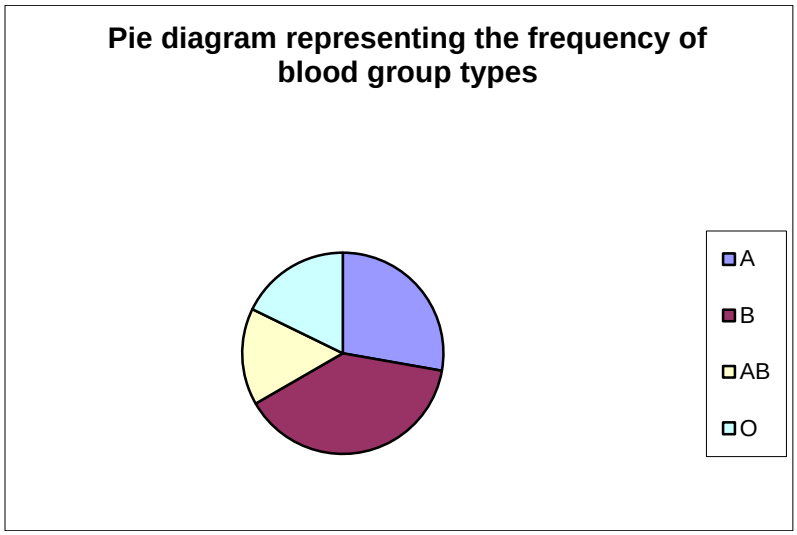

**Pie diagram** is based on a circle where the frequencies of the variables are expressed in degree. This is only applicable for categorical data or attributes. A pie chart presents the categories of data as parts of a circle or 'slices of a pie'. The underlying concept is to slice the Pie into areas that are proportional or percentages observed in the categories.

Table 4

| Blood group types | Frequency | Proportion |
|---|---|---|
| A | 25 | 0.27 |
| B | 35 | 0.38 |
| AB | 14 | 0.15 |
| O | 16 | 0.17 |
| Total | 90 | 1.00 |

In the example provided above, the sum of the frequencies of all the blood group types is 90. A circle represents 360°. The frequency of A blood type when converted to degree becomes 25/90 x 360° = 100° or 0.27x360°. Likewise B = 140°, AB = 56° and O =64°. Now plot the frequencies of ABO blood group types (converted into degrees) as slices of a pie. The labels may be placed at the side or on each of the slices of pie.

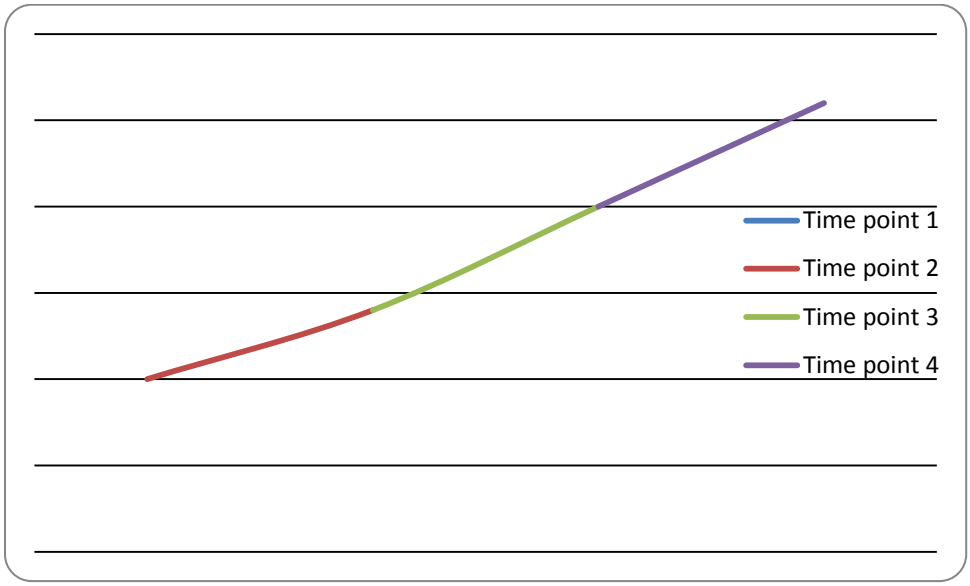
There are few problems associated with pie chart. (1) You cannot make a comparative presentation of frequency distribution of more than one group in a single pie, (2) very small proportions or percentages are difficult to represent since they occupy a small area of the pie, (3) if the categories are more, then the representation from each of these categories become difficult.




The line graph is a graph that uses line segments to connect data points and shows changes in data over time. Multiple Line Graph is a line graph that shows changes in data over time for more than one category.

For example, if you are to plot the mean height of a group of individuals taken over a period of time, at repeated intervals then line graph is the best way to present the data. Thus, in growth studies, if you are to plot distance curve or velocity curve, line graph is used. Distance curve is the increase in body dimension (or distance travelled) in successive time points. Use the X axis for plotting the time points and the Y axis for units of height. As per the data, plot the height of the individual/the mean height of the individuals against various time points. Now, joining the lines, you will get a curve or line graph. For eg., the height of an individual or the mean height of a group of individuals increases with age. Presented below is a distance curve used in growth studies. Here, you will see that the height increases at successive time points.

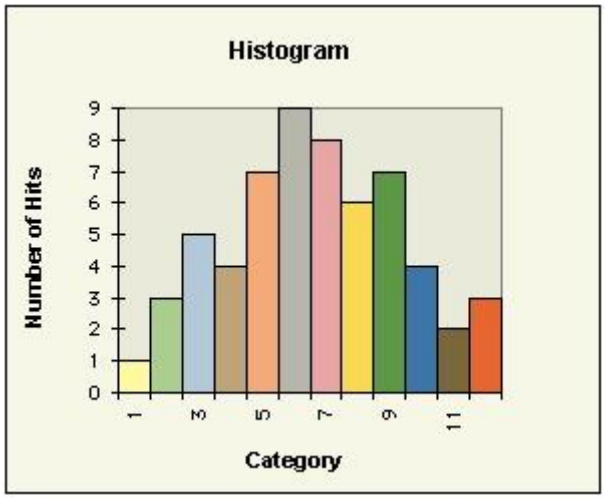
| | Height in cm |
|---|---|
| Time point 1 | 148.5 |
| Time point 2 | 148.9 |
| Time point 3 | 149.5 |
| Time point 4 | 150.1 |




Let us now try with a set of data where you are planning to compare heights of individuals of three different populations of same age.

| | Population 1 | Population 2 | Population 3 |
|---|---|---|---|
| Time point 1 | 144.2 cm | 143.2 cm | 146.2 cm |
| Time point 2 | 145.6 cm | 144.5 cm | 146.9cm |
| Time point 3 | 148.3 cm | 144.9 cm | 151.2 cm |
| Time point 4 | 149.0 cm | 145.2 cm | 152.3 cm |

**Histogram**




In case of continuous data, a histogram is constructed for graphical representation of data. A histogram is a vertical bar chart drawn over a set of class intervals that cover the range of observed data. Histogram speaks of many aspects of data---

1. the shape of the distribution,
2. the typical values of the distribution, the spread of the distribution,
3. and the percentage of distribution falling within a specified range of values.

Before drawing a histogram, organize the data into a *frequency distribution table*, as we do in case of grouped data. The class intervals should be formed with class boundaries, and with equal class width, since histograms are prepared for continuous data. Thus, the upper boundary of a class interval will be the same with the lower boundary of the subsequent class interval. This will be depicted in the graph; the margin of a bar representing the upper boundary of a class and the

margin representing the lower boundary of the subsequent class will overlap. In the graph, the class marks (or the class interval as such) are plotted in the X axis and its corresponding frequency on the Y axis. Unlike the bar graph (presented earlier in course of this discussion), the height of each bar represents the frequency of each class. The width of a bar represents a quantitative variable x, such as age rather than a category. Distribution of characteristics like height, blood pressure and haemoglobin concentration in a population are presented in a histogram.

Let us draw a histogram from the following data set.

Systolic blood pressures of 50 individuals are presented below. Present the data graphically

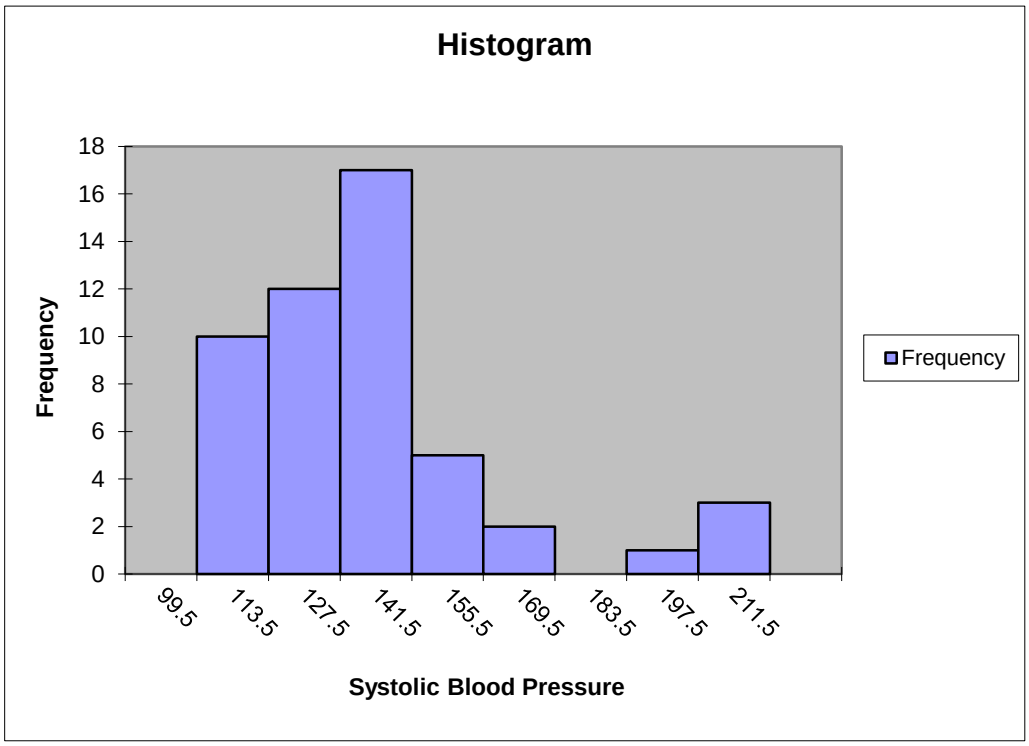
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 102 | 104 | 108 | 108 | 110 | 110 | 112 | 112 | 112 |
| 115 | 116 | 116 | 118 | 118 | 118 | 118 | 120 | 120 | 126 |
| 126 | 126 | 128 | 128 | 128 | 130 | 130 | 130 | 130 | 130 |
| 132 | 132 | 134 | 134 | 136 | 136 | 138 | 140 | 140 | 146 |
| 148 | 152 | 152 | 152 | 156 | 160 | 190 | 200 | 208 | 208 |

Data arranged in a group.

| Lower Boundary | Upper Boundary | Class Mark | Freq. | R. Freq. | C. Freq |
|---|---|---|---|---|---|
| 99.5 | 113.5 | 106.5 | 10 | 0.20 | 10 |
| 113.5 | 127.5 | 120.5 | 12 | 0.24 | 22 |
| 127.5 | 141.5 | 134.5 | 17 | 0.34 | 39 |
| 141.5 | 155.5 | 148.5 | 5 | 0.10 | 44 |
| 155.5 | 169.5 | 162.5 | 2 | 0.04 | 46 |
| 169.5 | 183.5 | 176.5 | 0 | 0.00 | 46 |

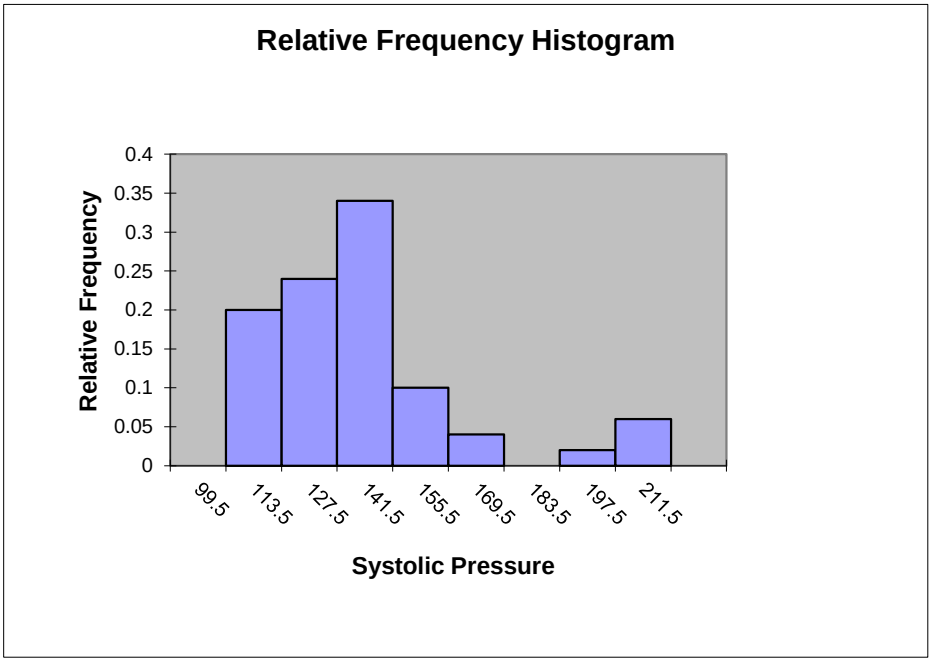
| | | | | | |
|---|---|---|---|---|---|
| 183.5 | 197.5 | 190.5 | 1 | 0.02 | 47 |
| 197.5 | 211.5 | 204.5 | 3 | 0.06 | 50 |

## Frequency histogram of blood pressure data

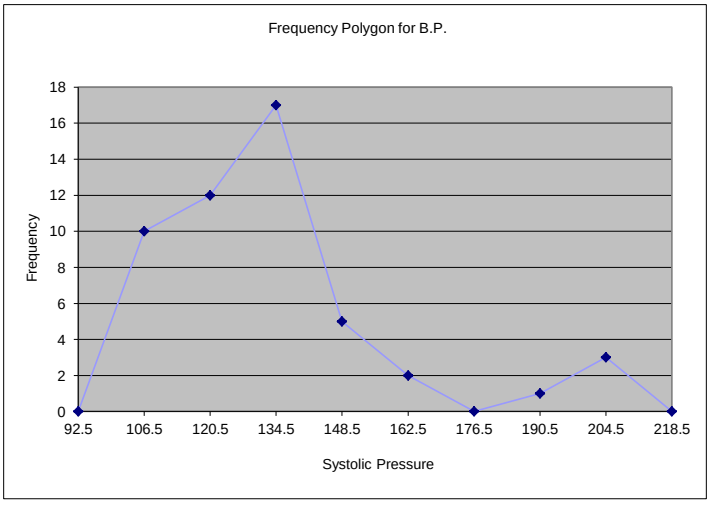



The histogram shows eights vertical bars. Each bar is marked by two figures; these are the class boundaries of each class. For example, the class boundary of the first class interval is (99.5-113.5) and that of second is (113.5-127.5). The height of the first bar reaches the value 10, i.e. the frequency of the class. In a similar manner, the other classes are depicted. It may be noted that since the frequency for the class interval (163.5-189.5) is 0, hence against the class mark 183.5, there is no raised bar. It is also important to note that the frequency of the class preceding to 99.5 and succeeding to 211.5 is zero.

Relative frequency histograms are like frequency histograms except the height of the bars represent relative frequencies. The relative frequency of a class is $f/n$ where f is the frequency of the class, and n is the total of all frequencies. Relative frequency tables are like frequency tables except the relative frequency are given. This type of graphical presentation will provide you the information about the proportion of observation that falls within a particular class interval.




**Frequency Polygon**

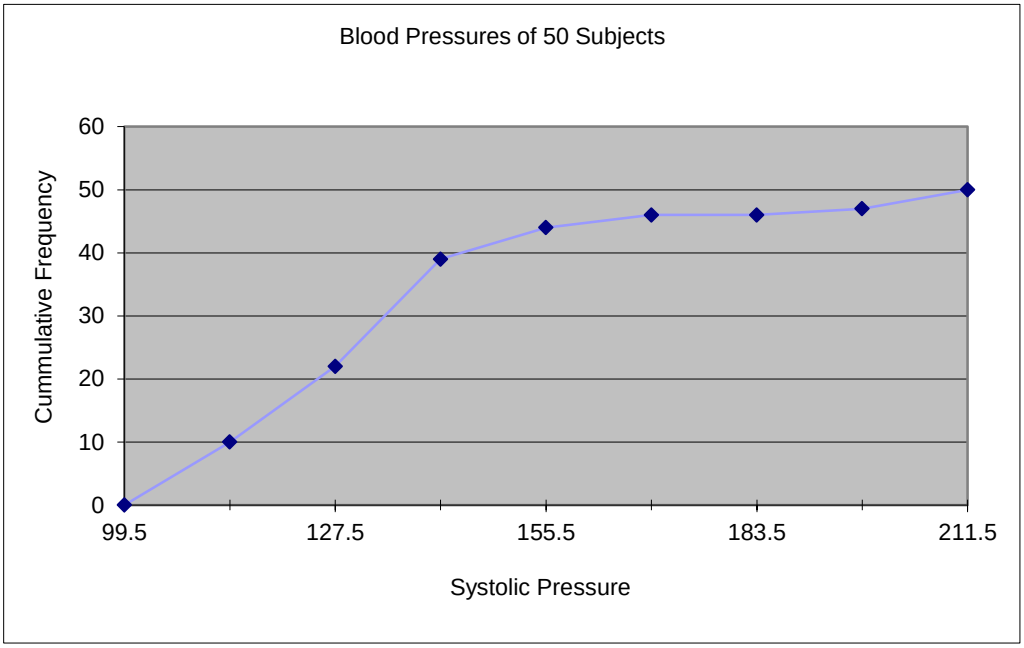


When the centre of the erected bars of a histogram are joined by a straight line we get frequency polygon curve. Larger the data and narrower the class interval, the curve will tend to become smooth. In drawing the frequency polygon curve one must remember that the lines should be extended to both the sides of the histogram touching the X axis showing 0 frequency. This indicates that the frequencies of the class interval preceding the first class interval and the one following the last class interval are zero (0). A frequency polygon drawn from the data presented above will show the lines meeting the X axis at the class mark 92.5 and at the class mark 218.5. These two class marks are calculated on the assumption that there are two class intervals (85.5-99.5) and (211.5-225.5).

The cumulative frequency curve or O'give is the graphical representation of cumulative frequency of a distribution. Make a frequency table showing class boundaries and cumulative frequencies. For each class, put a dot over the upper class boundary at the height of the cumulative class frequency.

| Lower Boundary | Upper Boundary | Class Mark | Freq. | R. Freq. | C. Freq |
|---|---|---|---|---|---|
| 99.5 | 113.5 | 106.5 | 10 | 0.20 | 10 |
| 113.5 | 127.5 | 120.5 | 12 | 0.24 | 22 |
| 127.5 | 141.5 | 134.5 | 17 | 0.34 | 39 |
| 141.5 | 155.5 | 148.5 | 5 | 0.10 | 44 |
| 155.5 | 169.5 | 162.5 | 2 | 0.04 | 46 |
| 169.5 | 183.5 | 176.5 | 0 | 0.00 | 46 |
| 183.5 | 197.5 | 190.5 | 1 | 0.02 | 47 |
| 197.5 | 211.5 | 204.5 | 3 | 0.06 | 50 |

Place dot on horizontal axis at the lower class boundary of the first class. Connect the dots. Against each class interval, the corresponding cumulative frequency is plotted in the graph. The graph appears to be of 'S' shape.




**Box plot**

This is the most commonly used graphical statistics used to estimate the distribution of quantitative data of a population. Boxplot is defined as a graphical statistic that can be used to summarize the observed sample data and can provide useful information on the shape and the tail of the distribution. This is also called as *Box and whisker plot.*

From Boxplot, one can infer about the following points---

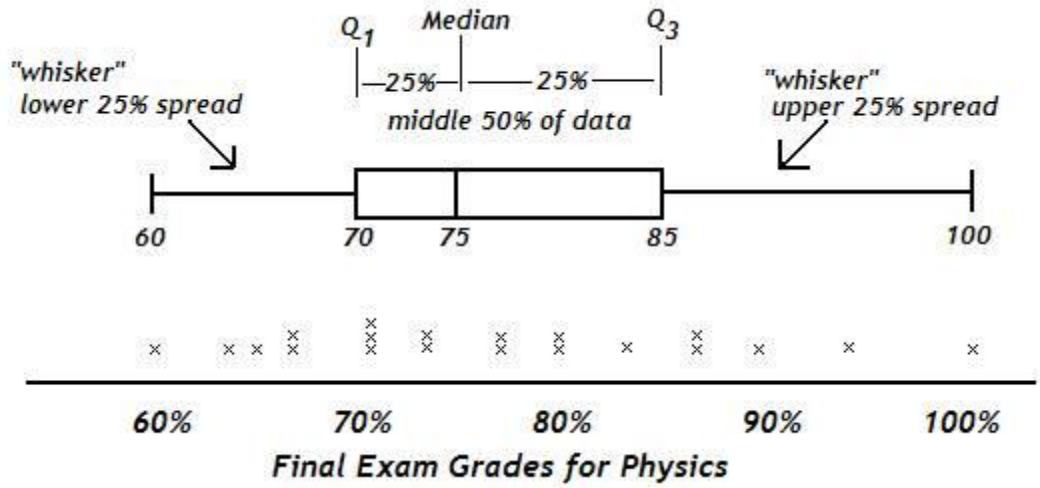
1. tails of the distribution or of the extreme values
2. the typical values in the distribution (minimum, maximum, median)
3. the spread of the population
4. may be used to compare two or more subpopulation.

A boxplot is based on five number summary associated with the data: (a) the minimum value, (b) the maximum value, (c) the first quartile (Q1) value, (d) the second quartile (Q2) value or the median, (e) the third quartile (Q3) value. These five values can create two types of boxplot, namely simple box plot and outlier boxplot. It can be horizontal or vertical

Suppose you have to draw a boxplot or box and whisker plot from the final grade scoring in Physics examination of a class.

**Table 5**

| Minimum score | Maximum score | Q1 | Median or Q2 | Q3 |
|---|---|---|---|---|
| 60 | 100 | 70 | 75 | 85 |

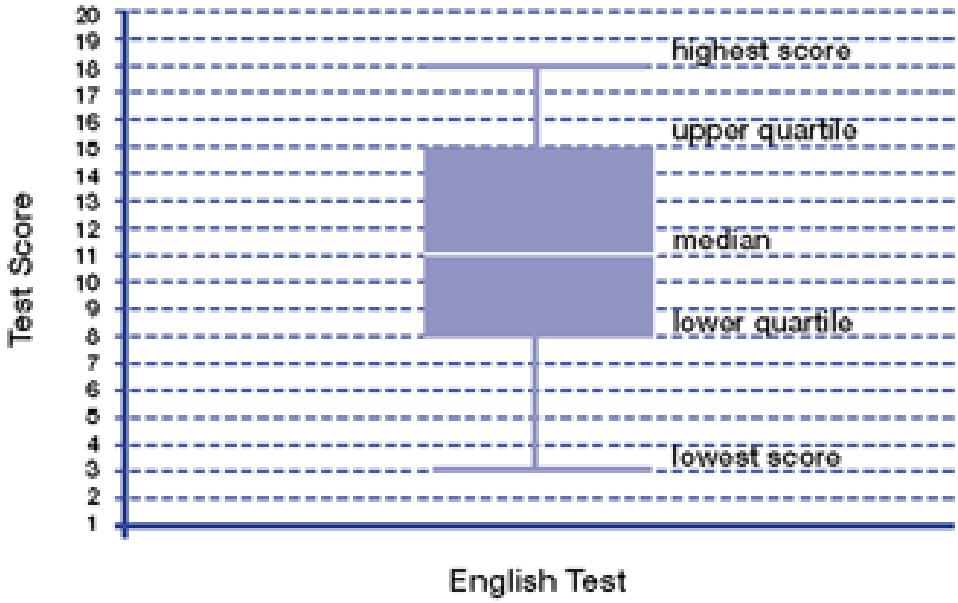



*What one can say from boxplot or box and whisker plot?*

1. One can get an idea of interquartile range from the length of the box. The larger the size of the box, greater is the range.
2. The whiskers, i.e. the length of line from Q1 to the lowest value and from Q3 to the highest value give an idea of the spread of the data beyond Q3 and below Q1 respectively. In the above example, the long tail or whisker to the right indicates the spread of data is more beyond Q3.

Boxplot type of graphical statistics has one drawback. It pays much importance on two extreme values (minimum and maximum). Suppose in the above example had one of the students scored 10 marks, and then the tail would have been elongated to the left (i.e. much beyond the value 60. This could create confusion. In such cases outlier boxplot is useful. Here, the effect of extreme values is minimized in the following way. The lower whisker is drawn to a further minimum point determined by using the formula Q1- 1.5 (Q3-Q1); similarly the upper whisker is extended beyond the maximum value by using the formula Q3+1.5 (Q3-Q1). The observations lying below and above these two points are represented with a symbol *asterisk (*)*.

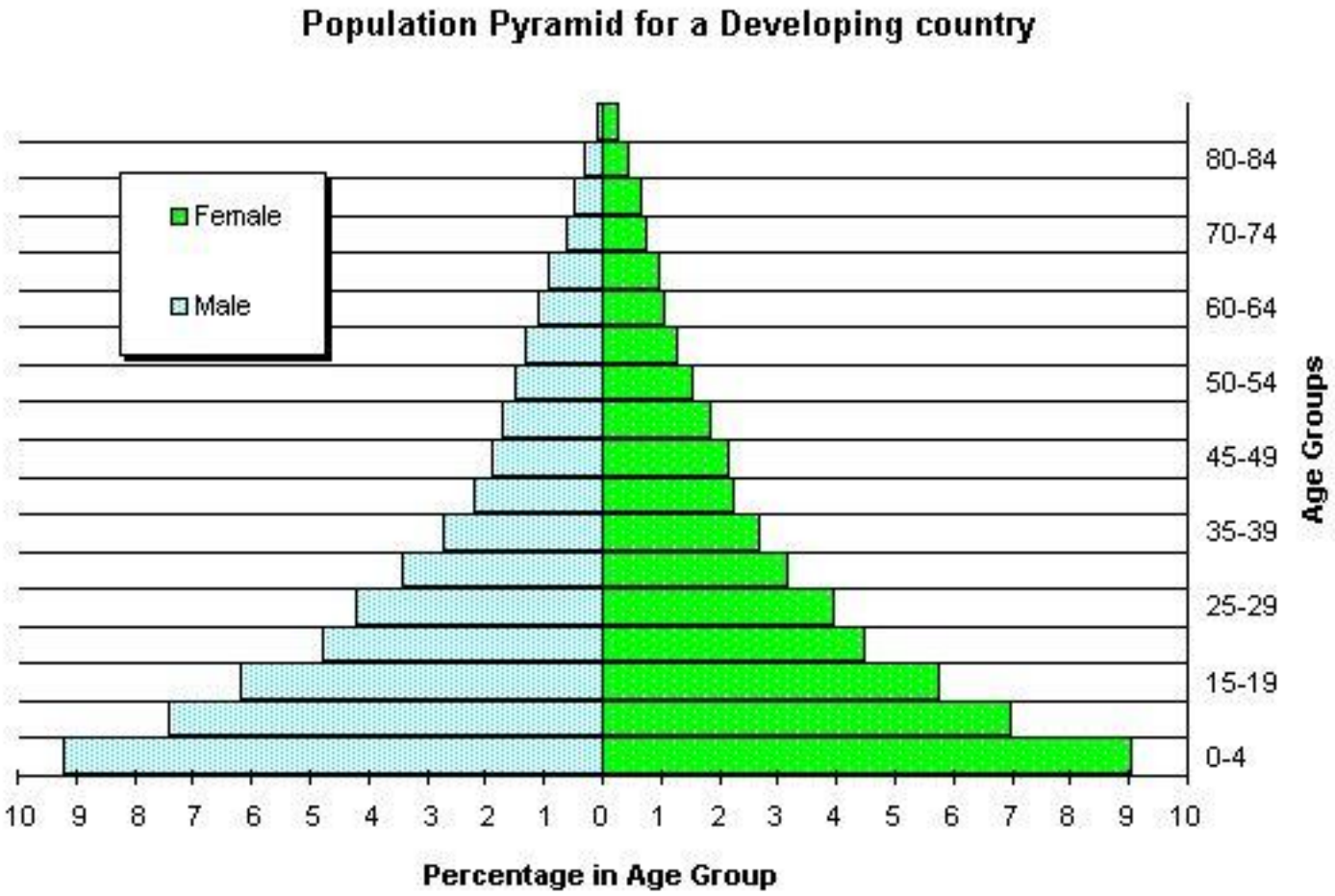
Boxplot graph may be of two types horizontal and vertical. The earlier is an example of horizontal type and the following one is of vertical type. This vertical type of boxplot has been prepared from test score in English language of a class.



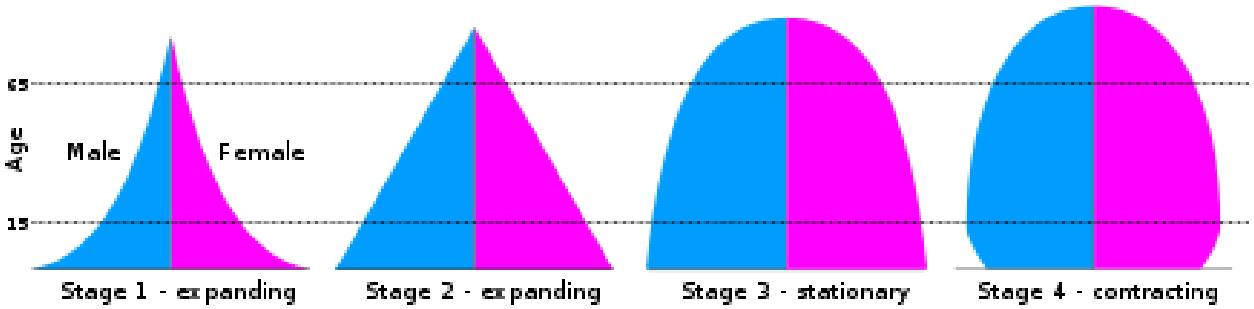

**Population Pyramid**

Population pyramid is the graphical representation of the age-sex structure of a population, i.e. distribution of a population according to sex for different age groups. It is just two histograms back to each other, one representing the age-structure of the female population and the other of the male population. Age groups (in years) like, (0-4), (5-9), (10-14) and so on are placed in ascending order on the vertical axis and the proportions or percentages of population are plotted against each age group and sex on the horizontal sides. By convention, the male population is presented on the right and the female on the left side of the vertical axis. An ideal population pyramid has a broad base which gradually tapers towards the apex, which means that highest population at the lowest age group with a gradual decrease in the frequency of population towards successive higher age groups. A pyramid of this type signifies a growing population. However, the shape of a population pyramid varies with the age-sex structure of a population. A dent in the population pyramid at a particular age group means lesser number of population at that age group; the reasons may be large scale migration, death due to war or epidemic, etc. of

that particular age group. A dent at the lowest age group of population pyramid signifies a trend of decreasing population; the reason may be recent adoption of family planning practice by the people or large scale death of children because of any recent epidemic. The following figure represents an ideal type of population pyramid, with a broad base with gradual tapering on both the sides.




The other types of population pyramids are as follows-

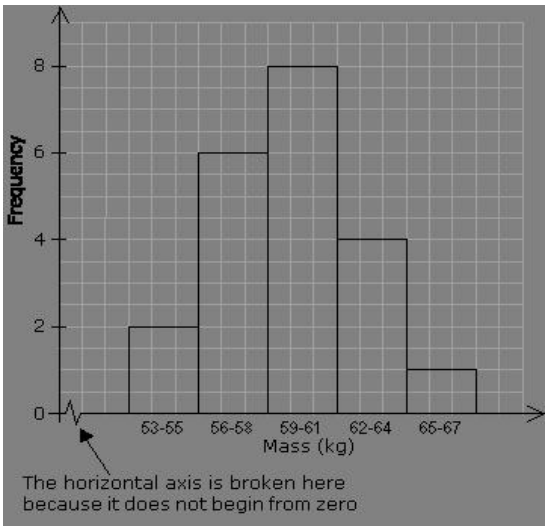



The first two population pyramids are of ideal types having an expanded base and a gradually tapering side. The third type of pyramid is like a bee-hive, i.e. the sides are not converging like the first two. This means that the proportion of representation of population in successive age groups is not decreasing like the previous two types. The fourth and the last type represent a pyramid of contracting type. Here, the upper portion of the pyramid is like that of the third (stationary) type, but the base is constricted. This indicates that (1) the proportion of children in this population is less compared to the successive older age groups, The reasons may be recent adoption of family planning practice by the people or large scale death of children because of any recent epidemic.

## Rules for good graphs

1. Always choose the correct type of graph as per your data. See, whether the data are discrete or continuous
2. Label the graph clearly. A graph should have an appropriate title so that the title becomes self explanatory. The X and Y axis should be labeled (for eg., the units of measurements should be mentioned)
3. In a graph, the X and Y axis meet at a point marked as zero (0). But it may happen that the scale of division of X and Y axis might not start from zero. In such cases, the axis that is not starting from zero is kept broken, implying that the scale is not from zero. The following figure demonstrates the number of individual who fall in various class intervals for weight. The Y axis depicts the frequency and it starts from zero. On the other hand, X

axis showing the class intervals for weight starts from the range (53-55) kg. Had the X axis been shown as continuous, then it would have meant that distance between the point zero and the first class interval (53-55) kg represents the class interval (0-53) kg; and this is absurd. The X axis is broken suggesting that the scale of X axis is not starting from zero.




4. The graph itself should be neat and clear
5. Do not forget to mention the source, if you are using any published data.

**CONCLUSION**

We can say that in this academic discourse, we learn that graphical presentation gives a portrayal of the frequency distribution. Namely, graphical presentation, types vary with respect to data types and also the purpose. These pictorial presentations will help you depict the distribution of as well as the dispersion of data.