TEACHING ARTICLE

EMJ SERIES ON STATISTICS AND METHODS
Part IV: Presenting and Summarizing Data Using Graphical Tools

Sanni Ali, DVM, MSc, PhD[1,2], Sileshi Lulseged, MD, MMed[3], Girmay Medhin, MSc, PhD[4]

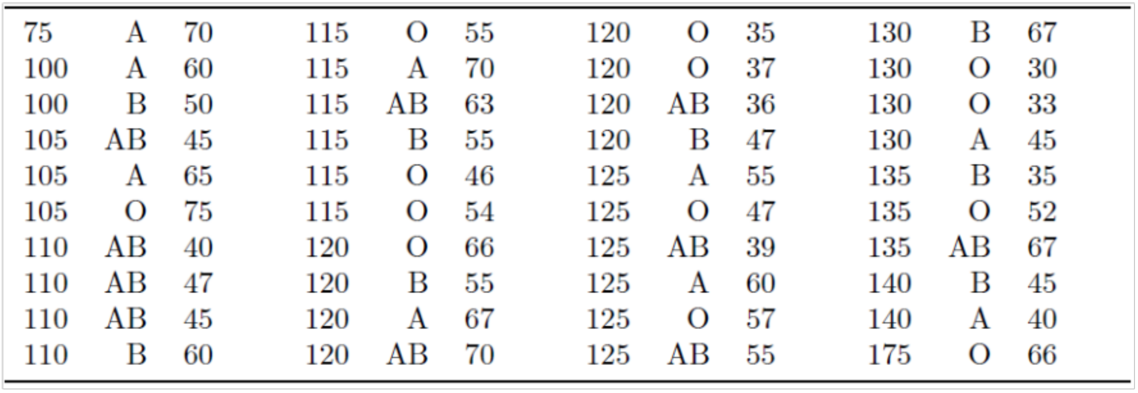
ABSTRACT

Graphs are a way to present data in visual form and if properly prepared can be a powerful way in which to convey statistical information. Graphs are a useful tool for displaying many types of data, and one of the easiest ways to see relationships between variables and/or compare numbers. To ensure that they are easy to interpret, graphs need to be presented in a way that enables them to stand-alone. They should be clutter free and use appropriate titles, legends, axis titles and footnotes. There are a range of different types of graphs that can be used. Care should be taken to ensure that the type of graph chosen is appropriate for the type of data that is being plotted. This article presents a brief overview of the most common types of graphs- It provides guidelines on how to create meaningful, easy to read and well-formatted graphs that are commonly used, including histograms, frequency polygon, line graphs, scatter plots, bar charts, and pie charts. It also highlights the design and presentation of components of a graphs - titles, axis labels, legends and footnotes, and appropriate representation of both axes, scale and error.

Displaying data

"A good picture is said to be worth a thousand words, but a good statistical portrait may sometimes be even more valuable, because summary for a large group may reflect more than a thousand items of dataset. On the other hand, a few well-chosen words (or numbers) of description can often be better than an unsatisfactory portrait". (1) Statistical information can be summarized and communicatednumerically using summary measures or visually using tables, graphs or charts. In Part I, we presented measures of location and spread. In this teaching article, we describe graphical methods to summarize and display huge information that exists in a dataset using , as example, records of systolic blood pressure (in mm Hg), blood groups and ages of 40 individuals presented in Table 1.

Table 1: Systolic blood pressure measurements (in mmHg), blood groups and ages (in years) of 40 Individuals

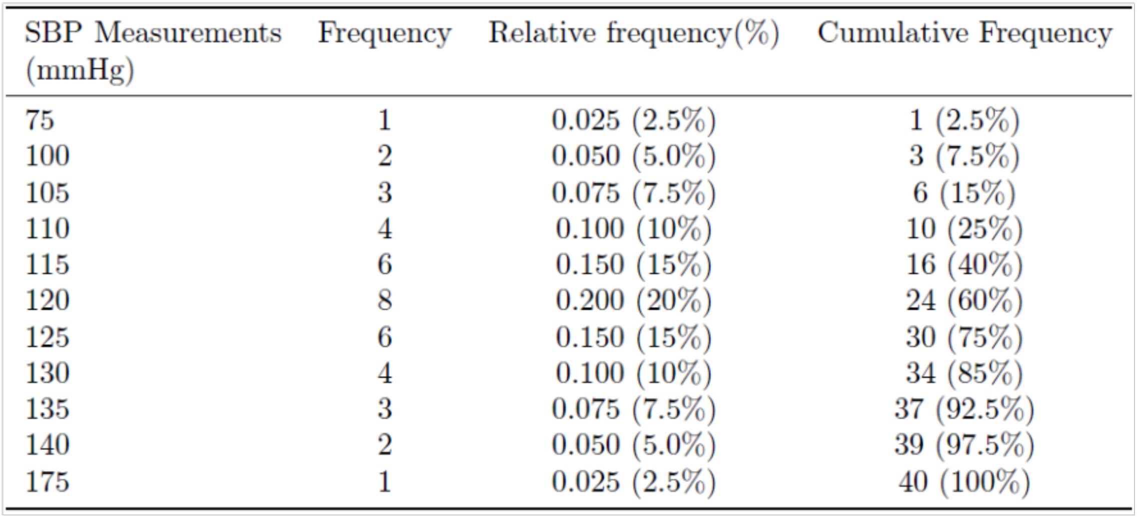


75	A	70	115	O	55	120	O	35	130	B	67
100	A	60	115	A	70	120	O	37	130	O	30
100	B	50	115	AB	63	120	AB	36	130	O	33
105	AB	45	115	B	55	120	B	47	130	A	45
105	A	65	115	O	46	125	A	55	135	B	35
105	O	75	115	O	54	125	O	47	135	O	52
110	AB	40	120	O	66	125	AB	39	135	AB	67
110	AB	47	120	B	55	125	A	60	140	B	45
110	AB	45	120	A	67	125	O	57	140	A	40
110	B	60	120	AB	70	125	AB	55	175	O	66

Frequency tables

Categorical (nominal and ordinal) variables can be easily summarized by counting the number of observations in each category. These counts are called frequencies, often presented using relative frequencies as the proportions or percentages of the total number of observations. In contrast, cumulative frequency tells us how the count or percentage of observations accumulate starting from the lowest value, as the values increase, up to and including a certain value or interval, the cumulative frequency in the final interval has a cumulative percentage of 100% (Table 2). Frequencies and relative frequencies are either tabulated using frequency tables or illustrated using bar charts or pie charts.

Table 2: Frequency table for systolic blood pressure measurements of the 40 Individuals



SBP Measurements (mmHg)	Frequency	Relative frequency(%)	Cumulative Frequency
75	1	0.025 (2.5%)	1 (2.5%)
100	2	0.050 (5.0%)	3 (7.5%)
105	3	0.075 (7.5%)	6 (15%)
110	4	0.100 (10%)	10 (25%)
115	6	0.150 (15%)	16 (40%)
120	8	0.200 (20%)	24 (60%)
125	6	0.150 (15%)	30 (75%)
130	4	0.100 (10%)	34 (85%)
135	3	0.075 (7.5%)	37 (92.5%)
140	2	0.050 (5.0%)	39 (97.5%)
175	1	0.025 (2.5%)	40 (100%)

Graphs

A graph is a visual representation of data that can greatly help description, exploration or summarization of large amount of complex information in a quick, clear and simplified way (2). Graphs also enhance readability of a research report by showing patterns and trends in data. It is an effective way of communicating data in a pictorial form particularly when precise numeric details are not required and when a trend or comparison or relationship between data values can be demonstrated.

There are many different graph types to choose from; a critical issue is to ensure that the graph type chosen is the most appropriate for the data and the design and presentation of the graph help the reader better interpret the data (2). The most common types of graphs include histograms, frequency polygon, line graphs, scatter plots, bar charts, and pie charts. Appropriate design and presentation of graphs should include 1) the different components such as titles, axis labels, legends and footnotes, 2) appropriate representation of both axes, scale and error, 3) a visual style that is uncluttered, easy to interpret, and clearly shows trends or differences in the data.

Bar charts

Bar charts, also called bar diagrams, are the simplest and the most commonly used graph types to display and compare frequencies or relative frequencies or other measures (e.g. mean) for categorical data. They are easy to create and straightforward to interpret. They have several variations including horizontal bar charts, grouped or component charts, and stacked bar charts (2). A bar chart is constructed such that one axis represents the categories being compared and the other axis represents the size of each category. The lengths of the different bars are proportional to the size (proportion or percentage) of the category they represent. In standard bar chart, the x-axis (the horizontal axis) represents the different categories, hence, it has no scale (2). In order to emphasize the fact that the categories are discrete, a gap is often left between the bars on the x-axis.

The y-axis (the vertical axis), on the other hand, has a scale indicating the units of measurement.

Bar charts are useful for displaying data that are classified into nominal (such as blood group) or ordinal (such as stages of cancer) categories. With nominal data, arranging the categories so that the bars grade sequentially from the largest category to the smallest category or vice versa helps the reader to interpret the data. However, this is not necessary for ordinal data because the categories already have an obvious sequence. Bar charts are also useful for displaying data that include categories with negative values, because it is possible to position the bars below and above the x-axis.

The bars can be drawn either vertically or horizontally depending upon the number of categories and length or complexity of the category labels. Horizontal bar charts (Figure 1A) are normally drawn so that the bars are vertical which means that the taller the bar, the greater the value or the larger the category. However, it is also possible to draw vertical bar charts (Figure 1B) so that the bars are horizontal which means that the longer the bar, the larger the value or the larger the category. Horizontal bar chart is a particularly efficient way of displaying data (means or percentages) when the different categories have long titles that would be difficult to label below a vertical bar, or when there are a large number of different categories (usually more than eight) and there is insufficient space to fit all the columns required for a horizontal bar chart across the page.

Clustered or grouped bar charts are a way of showing information about two or more subgroups of the main categories on one graph. Plotting multiple categories on one graph increases the amount of information that can be shown, although care must be taken to avoid over-complicating the graph. For the sake of description, we describe a grouped bar chart for the frequency of blood groups (sub-groups) within each of the four age categories (main category) from Table 1 (Figure 1C). A separate bar represents each of the sub-groups (e.g. blood groups) and these are usually colored or shaded differently to distinguish between them. In such cases, a legend or key is usually provided to indicate what sub-group each of the shadings or colors represent. The legend is often placed in the plot area or may be located below the chart. In using grouped bar charts, one has to ensure that the chart does not contain too much information making it complicated to read and interpret. Grouped bar charts can be drawn as both horizontal and vertical charts depending upon the nature of the data to be presented.

Stacked bar charts are similar to grouped bar charts in that they are used to display information about the sub-groups that make up the different categories (Figure 1D). In stacked bar charts the bars representing the subgroups are placed on top of each other to make a single column, or side by side to make a single bar. The overall height or length of the bar shows the total size of the category whilst different colors or shadings are used to indicate the relative contribution of the different sub-groups. Stacked bar charts can also be used to show the percentage contribution of different sub-groups to each separate category. In this case the bars representing the individual categories are all of the same size.

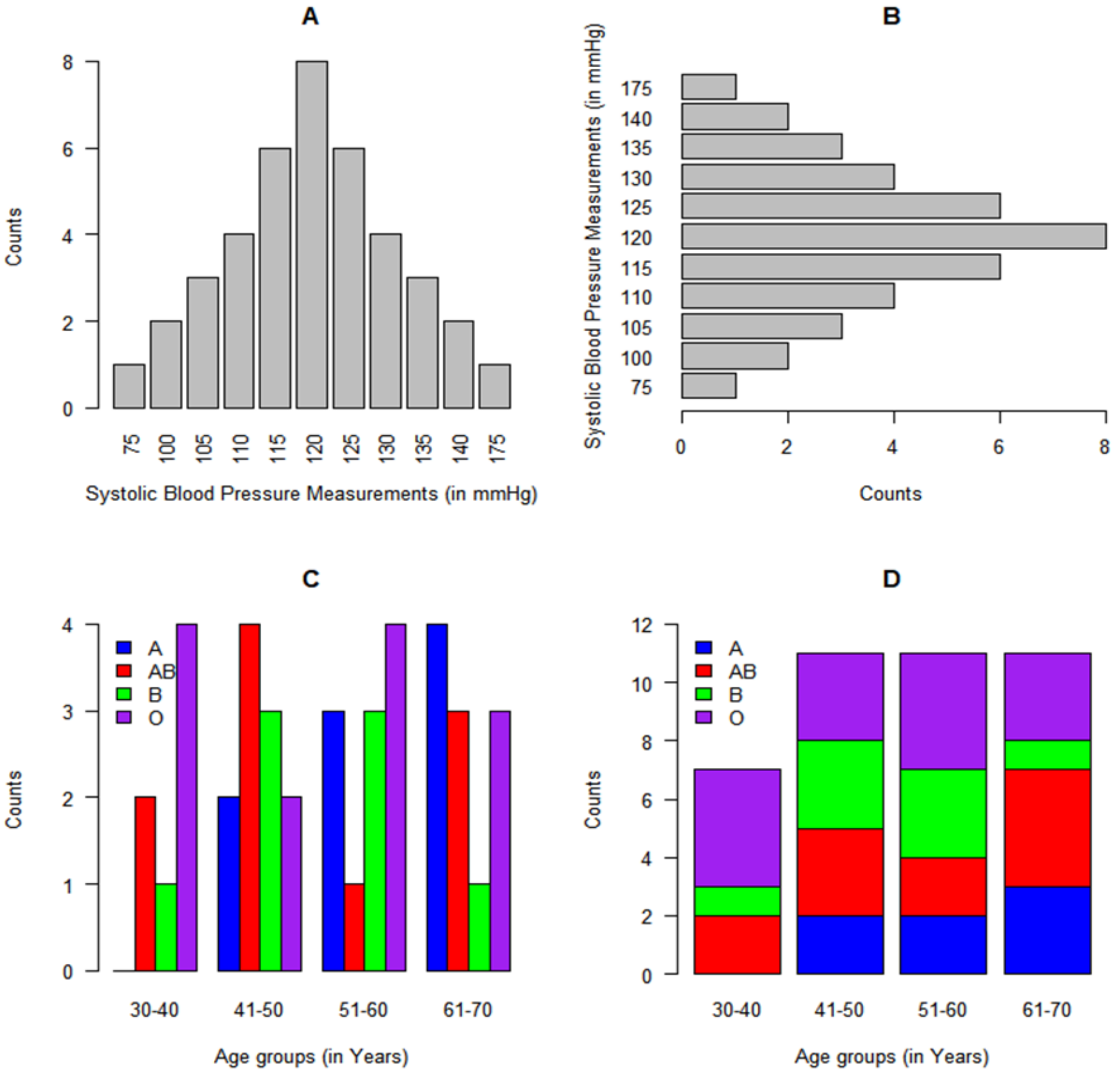


Figure 1: Bar plots of systolic blood pressure measurements (in mm Hg), blood groups and ages of 40 individuals in Table 1. Horizontal bar plots (A), vertical bar plots (B), clustered or grouped bar charts of blood groups with in age categories (C), and stacked bar charts of blood groups with in age categories (D)

Pie charts

Pie charts are a visual way of displaying how data are distributed between different categories. Hence, pie charts should only be used for displaying nominal data having small number of categories (around six or fewer) with some variation in size (2). In pie chart, the circle is divided so that the areas of the sectors are proportional to the frequencies or percentages which are usually provided next to the corresponding slice of pie (Figure 2).

Pie charts are not recommended when the number of categories is large. If it is used in that situation it becomes cluttered and difficult to distinguish between the relative sizes of the different sectors making interpretation difficult. Pie charts provide a good visual representation of the data when the categories show some variation in size.

The different sectors of the pie chart are usually arranged clockwise in order of magnitude. A category of data that does not contain a unique category of data but summarizes several, for example “others” is represented by a slice of the sector and is usually displayed last. This helps to avoid detraction from the named categories of interest. It is also helpful to color or shade the different slices so that they grade from dark to light tones as we move from the first to the last slice. In using two or more pie charts to compare two sets of data where the categories are the same or similar, it is helpful to maintain the same ordering and coloring of slices in the second pie chart as the first in order to facilitate comparison. Some design elements such as 3D effects (Figure 2 right) could also detract from the message that the researcher want to convey for example by producing optical effects that makes it hard to compare different categories.

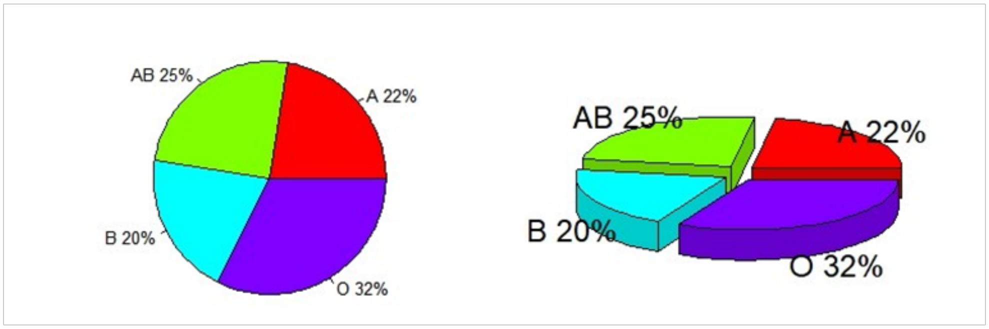


Figure 2: Pie charts of blood groups of observation in Table 1, 2-D (Left) and 3-D (Right).

Histograms

Histograms are a special form of bar chart that presents numeric data and its distribution; the data often represent continuous rather than discrete categories. This means that in a histogram, unlike bar chart, there are no gaps between the columns (bars) representing the different categories or a range of data. The width of the bars in histograms can be different (with uneven sized categories) or equal (with even sized categories); equal sized categories is recommended to avoid distorted view of the data (3).

In a bar chart the length of the bar indicates the size of the category, but in a histogram (with even sized categories) it is the area of the bar that is proportional to the size of the category (2, 4). This difference is due to the fact that in a histogram both the x-axis and y-axis have a scale, whereas in a bar chart only the y-axis (in horizontal bar charts) or the x-axis (in vertical bar charts) has a scale. In a histogram with uneven sized categories, the vertical axis measures the number of values that fall within each interval or, if desired, the percentage of values that fall within each interval. The horizontal axis displays the limits that are used for each interval. For each interval a rectangular column centered on the midpoint is drawn rising from the horizontal axis (Figure 3A).
A histogram is the most commonly used graphical method of displaying the frequency of continuous data divided into suitable intervals or ranges. Dividing the data will help reduce the number of intervals since the data may have a large number of possible values resulting in complex histogram. When the number of columns is very large, it will be difficult to interpret the information. Histograms can also be used for ordinal and discrete data.

Frequency polygon

Frequency polygons are a natural extension of the (relative) frequency histograms. They differ in that, rather than drawing bars, each class is represented by one point (mid-point of the width of the bars) and these are joined together by straight lines. The method is similar to that is used in producing a histogram: (1) Produce a (percentage relative) frequency table, (2) Draw the axes: the x-axis needs to contain the full range of the classes used and the y-axis needs to range from 0 to the maximum (percentage relative) frequency, (3) Plot points and pick the mid-point of the class interval on the x-axis and go up until you reach the appropriate (percentage) value on the y-axis and mark the point, and (4) Join adjacent points together with straight lines (Figure 3B).

A histogram is the most commonly used graphical method of displaying the frequency of continuous data divided into suitable intervals or ranges. Dividing the data will help reduce the number of intervals since the data may have a large number of possible values resulting in complex histogram. When the number of columns is very large, it will be difficult to interpret the information. Histograms can also be used for ordinal and discrete data.

Frequency polygon

Frequency polygons are a natural extension of the (relative) frequency histograms. They differ in that, rather than drawing bars, each class is represented by one point (mid-point of the width of the bars) and these are joined together by straight lines. The method is similar to that is used in producing a histogram: (1) Produce a (percentage relative) frequency table, (2) Draw the axes: the x-axis needs to contain the full range of the classes used and the y-axis needs to range from 0 to the maximum (percentage relative) frequency, (3) Plot points and pick the mid-point of the class interval on the x-axis and go up until you reach the appropriate (percentage) value on the y-axis and mark the point, and (4) Join adjacent points together with straight lines (Figure 3B).

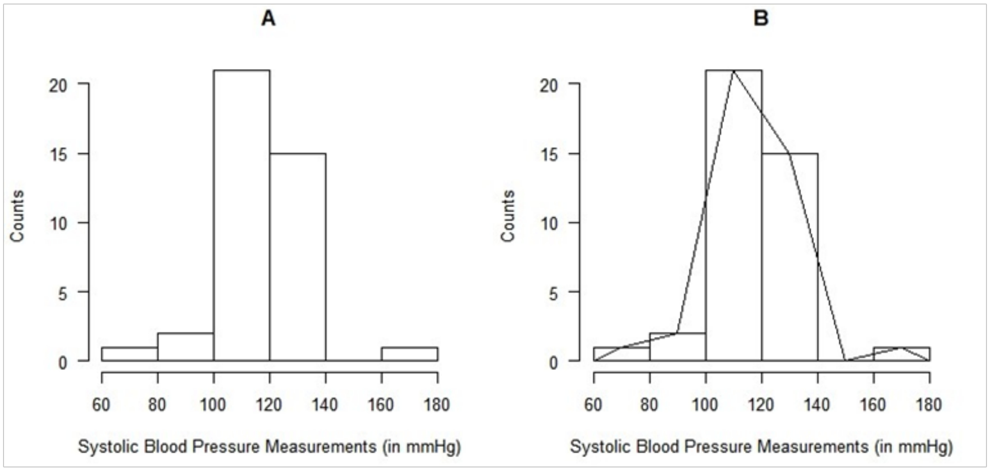


Figure 3: Histograms (A) and frequency polygons overlaid with histograms (B) systolic blood pressure measurements (in mm Hg) of systolic blood pressure measurements (in mm Hg).

Scatter plots

Scatter plots illustrate the relationship between two numerical characteristics (continuous variables) plotted on the x and y axes. In a scatter plot a dot represents each individual and is located with reference to the x-axis and y- axis, each of which represent one of the two measurements. It is usually possible to identify whether there is any association between the two measurements by analyzing the pattern of dots that make up a scatter. Regression lines can also be added to the graph to better visualize and explain the pattern or the relationship. Figure 4 shows scatter plot of age (x-axis) and systolic blood pressure measurements (y-axis) only for descriptive purpose. As can be seen from the plot, the systolic blood pressure is constant with an average of about 120 mmHg irrespective of age as adult blood pressure measurement does not change with age.

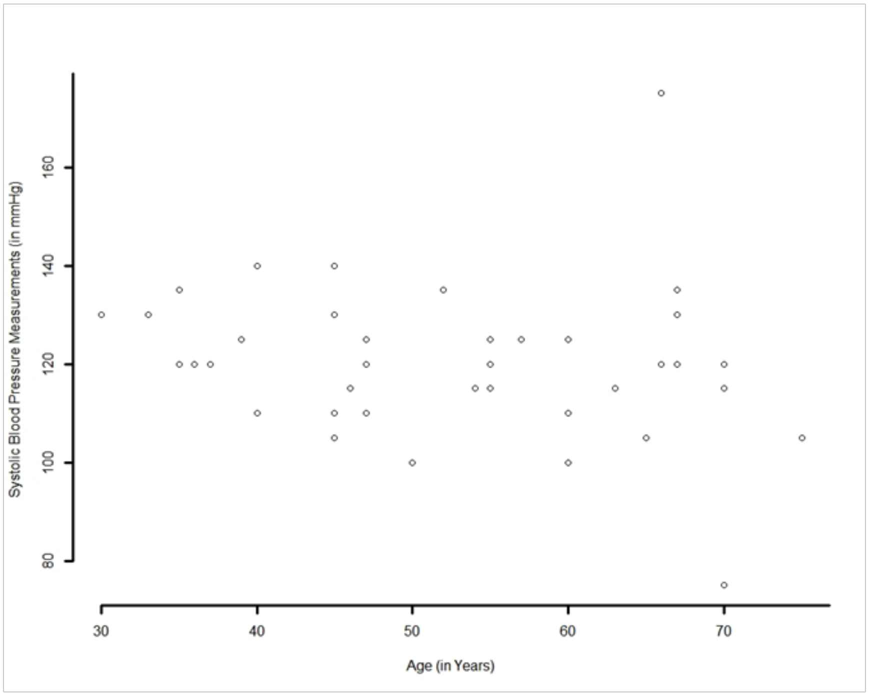


Figure 4: Scatter plots of age (in years) and systolic blood pressure measurements (in mm Hg) of systolic blood pressure measurements (in mm Hg).

Box plots

Box and whisker plots (i.e. Box plots) are another graphical method for displaying and summarizing data. They are particularly useful in highlighting differences between groups. Box plots use some of the key summary statistics (also called the five number summary): the minimum value, the three quartiles, and the maximum value. Box plots display only a summary of the data, and some of the information about the shape of the distribution is obscured. In box plot, the rectangle (”box”) is drawn with ends at the 3rd quartile and 1st quartile (i.e. the length of the box representing the interquartile range) divided by a line at the mid-point representing the median (the second quartile). Lines projecting out from the box (“whiskers”) are added to display more extreme parts of the distribution of values (3). Extreme observations are indicated by points beyond the lines on both ends (Figure 5).

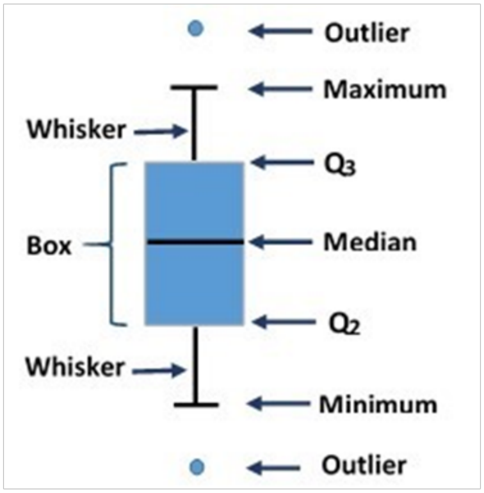


Figure 5: Box and whisker plots using the five number the minimum value, the three quartiles, and the maximum value.

Line graphs

Line graphs are usually used to summarize time series data (i.e. data collected over several intervals) displaying how one or more variables vary over a continuous period of time. These type of graphs are particularly useful for identifying patterns and trends that might exist in the data such as temporal effects, large changes and turning points. They are also appropriate for displaying data that are measured over other continuous variables such as blood pressure measurements. In a line graph, the x-axis often represents time at which measurements are taken whilst the y-axis has a scale that indicates the measurement. Several line graphs can be plotted on the same line chart. This type of graph is particularly useful for analyzing and comparing trends in different datasets.

Graph formatting:

An appropriately formatted graph should, as required, include a title, axis labels, legends and footnotes and representation of axes, scale and error. It should have a visual style that is clutter-free, easy to interpret, and appropriately represent trends or differences in the data.

Title: The title is usually placed at the center, either above or below the graph and sequentially numbered and referenced in the narrative of the manuscript. The title of the graph should explain what the x- and y-axes represent. Both the x-axis (horizontal) and y-axis (vertical) should be labelled, and the labels should be brief and explain exactly what each aspect of the graph is showing. The units of measurement (e.g. meters, mm Hg, etc.) should also be included.

Legend: The legend provides a key to the various data plotted on a graph. For example, if you have used colors or shading, the legend should explain what the colors and/or shading represent.

Footnotes: The footnote further explain the data; for example, for a sample survey, include a footnote describing the sample that is being represented in the graph, and the number of respondents in the sample (n).

Axes and scales: The vertical axis often starts at zero (0), so that the range of the axes does not distort the data and allow for misinterpretation. An exception to this rule is when there are negative values, in which case the scale would start at less than zero.

Range of error: Results represented in graphs are often estimates. They may not be precise figures, but may fall within a range with a certain level of confidence, covered in Part II of the series (5).

REFERENCES

1. Feinstein AR. Principles of Medical Statistics. Chapman & Hall/CRC, 2002.
2. Altman DG, Bland JM. Presentation of numerical data. BMJ 1996;312:572. https://doi.org/10.1136/bmj.312.7030.572 (Accessed September 2018)
3. Kirkwood BR, Sterne AC. Essential medical statistics. John Wiley & Sons, 2010.
4. Riffenburg RH. Statistics in Medicine (3rd ed). Elsevier, 2012.
5. Ali, S, Lulseged S, Medhin G. EMJ Series on Statistics and Methods. Ethiop Med J 2018, Vol. 56(4):290.