Draft date: 2019.3.29

Descriptive Statistics and Graphing

<Material provided by>
AMED “International Journals Project”

Contents

Introduction

In this module, you will learn about ways of using graphs to visually present data. Presenting data in graphs enables you to communicate the point of a study or information that would be hard to understand if summarized using numerical values alone in a way that the reader can understand more easily.

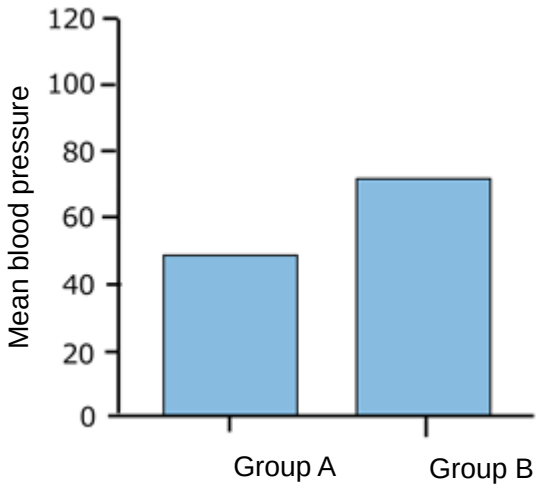
Learning Objectives

- Name each graph and explain its significance.
- Explain the meaning of error bars and how to use them.

Types of Graph and How to Read Them

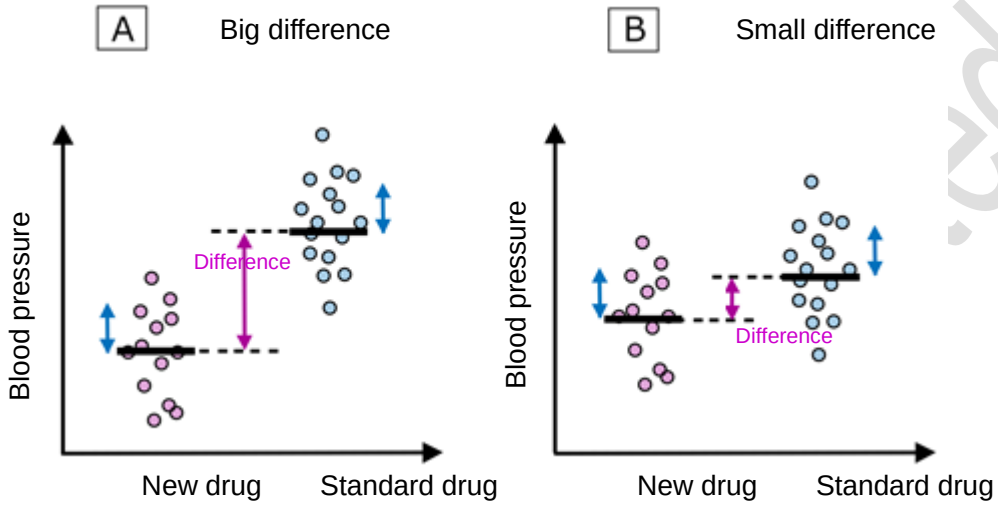
Bar Charts and Error Bars

The type of graph that most of you almost certainly have seen at least once is the bar chart. This graph is often used when comparing numerical values between groups. The most commonly used bar chart is a graph in which the height of the bar represents the mean. From the graph below, we can see that mean blood pressure in Group A is higher than in Group B.

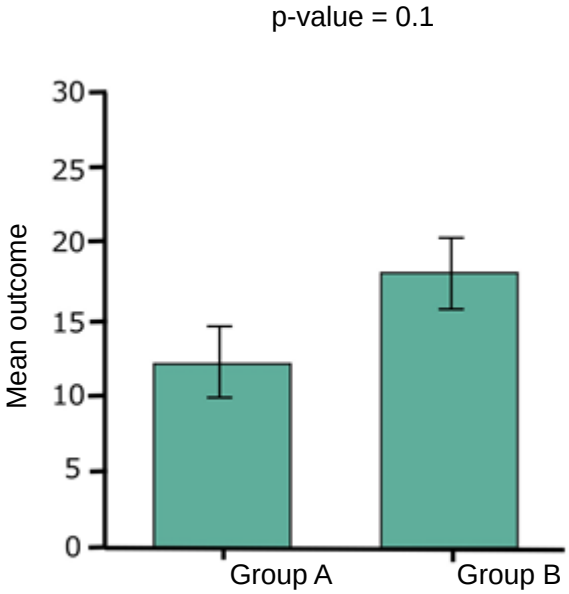


However, this graph does not tell us whether the difference in mean blood pressure between Group B and Group A is statistically significant. It is easier to find statistically significant differences when the difference in data that you wish to detect between groups (such as the

differences in mean blood pressure between groups) is greater than the variation in the data. Consequently, when comparing mean values for outcomes between multiple groups, it is important to present not only the bars representing the mean, but also the variation in the data. Error bars are often used to indicate variation in data.

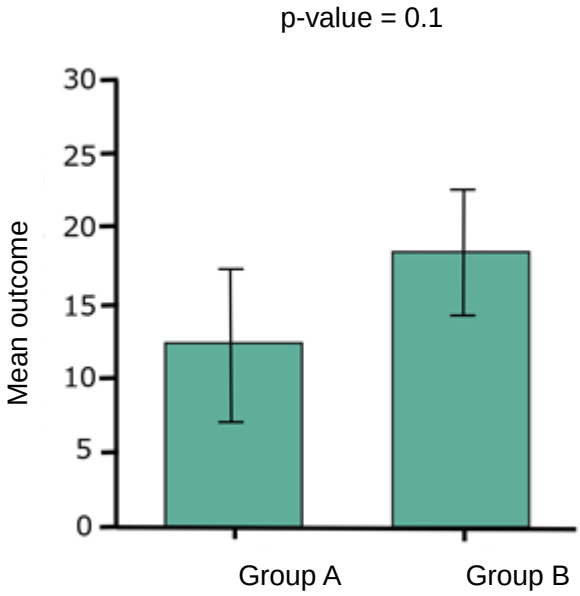


For error bars, either “standard deviation,” “standard error,” or “95% confidence interval” is generally used.



Bar chart with error bars showing standard error

The graph above has error bars showing ± one standard error of the mean. This is used in a great many articles and graphs like this are often used to argue that there is a statistically significant difference because the two error bars do not overlap. However, this is incorrect. In fact, although the two error bars in the graph above do not overlap, the p-value is 0.1, which means that it does not show a statistically significant difference. The only graphs that permit the judgment that there is a statistically significant difference because the error bars do not overlap are those using a confidence interval as the error bar. In this example, the upper and lower limits of the 95% confidence interval correspond to the mean ± 2 × standard error.

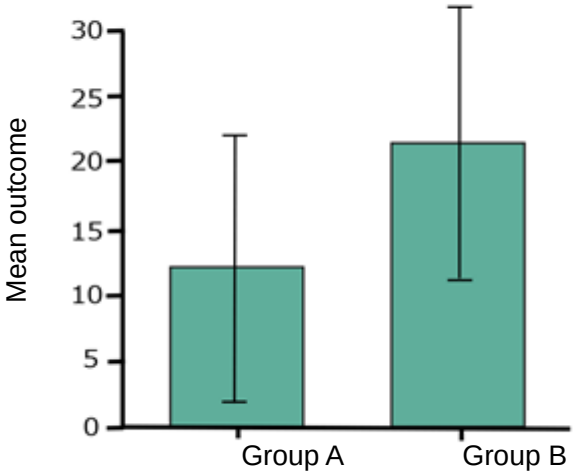


Bar chart with error bars showing confidence intervals

The graph above has error bars presenting the 95% confidence intervals added to a graph showing the mean. Looking at the results, we can see that the two confidence intervals overlap. In this case, if the hypothesis is tested at a significance level of 5%, we cannot say that there is a statistically significant difference.

The graph below has error bars showing the mean ± standard deviation. If the data follow normal distribution, the mean ± standard deviation signifies a range containing around two-thirds of the data. Although the two error bars in the graph below overlap, there is a statistically significant difference. While a standard deviation error bar indicates the variation in the data gathered, it is not useful for checking whether there is a statistically significant difference.

p-value = 0.1



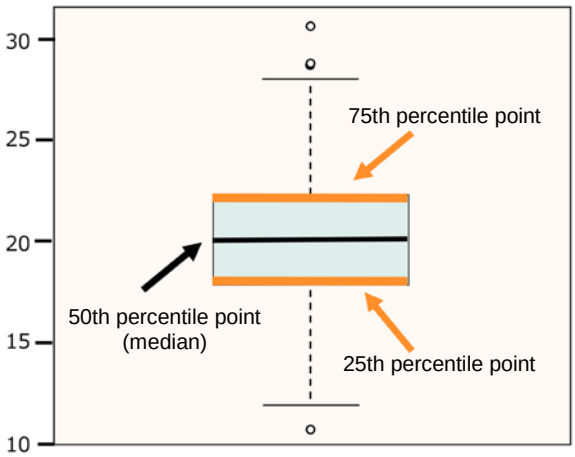
Bar chart with error bars showing standard deviations

Indicating What the Error Bar Shows

As stated above, what the graph shows differs according to which indicator of variation is used as the error bar. Accordingly, when using bar charts in articles to show the mean of data, it is vital to specify what the error bars used in the chart present. The guidelines of international journals such as Nature stipulate that graphs should have error bars wherever possible and that those error bars must be defined.

Box-and-whisker Plots with a Variety of Information

So far, this module has explained bar charts as a means of plotting information with mean values. However, you have also seen that among the other ways of summarizing data are the median and the interquartile range (see Module 8). Box-and-whisker plots are a way of graphically presenting the median and the interquartile range.

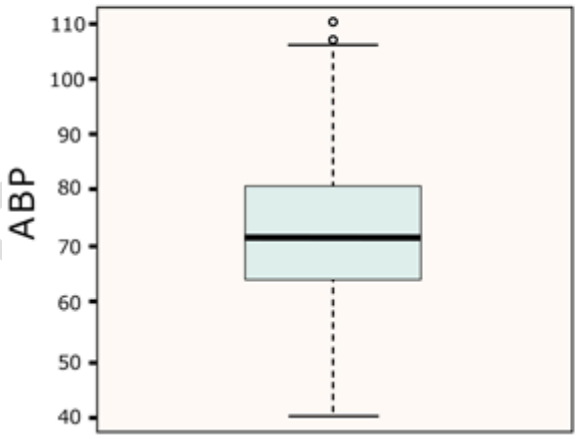


In a box-and-whisker plot, “whiskers” like error bars extend vertically above and below a “box” with a

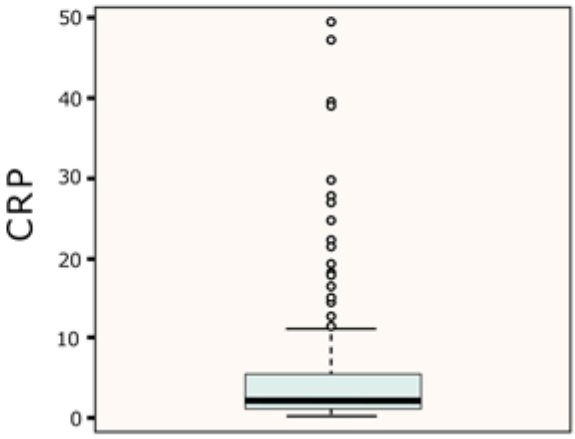
horizontal line inside. The line at the very bottom of the box indicates the 25th percentile point, the line at the very top the 75th percentile point, and the line inside the box the 50th percentile point (median). In other words, the upper and lower values of the box indicate the interquartile range. The whisker above the box is usually drawn from “the red line at the top, which corresponds to the lid of the box, for a length 1.5 times the vertical length of the box up to the highest datum that exists within this range.” Similarly, the whisker below the box is drawn from “the red line at the bottom, which corresponds to the base of the box, for a length 1.5 times the vertical length of the box down to the lowest datum that exists in this range.” The value, “1.5,” is determined differently by each software package. You need to check it. Data that fall outside the whiskers can be indicated as outliers using dots, which allows a box-and-whisker plot to show the location of extreme data values.

If the median is located just about in the center of the box and the outliers are symmetrically located, we can hypothesize that the distribution of the data is close to normal distribution. However, if the median line is extremely close to either end of the box, it suggests that the distribution of the data is skewed. Using the mean in a case like this gives rise to misunderstandings, as explained in Module 8. Thus, there tends to be a preference for using box-and-whisker plots for graphs with a large volume of information, as they not only show the median and interquartile range, but also provide a visual understanding of the extent to which the distribution of the data is skewed. The graphs below show an example in which normal distribution (left) and skewed distribution (right) can be assumed.

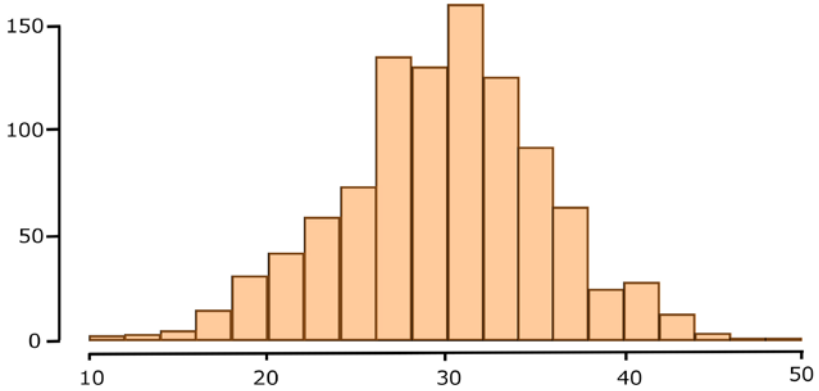
Distribution permits normal distribution to be assumed



Skewed distribution does not permit normal distribution to be assumed

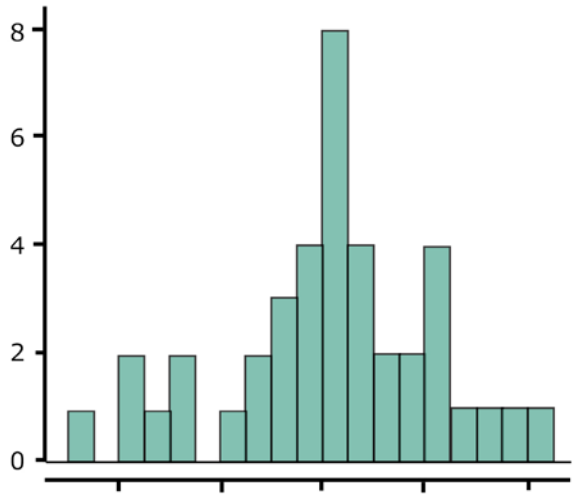


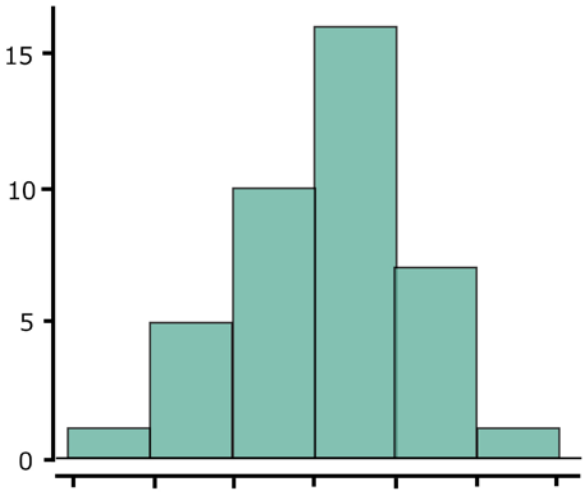
Histograms



Histograms are graphs that are ideal for investigating the distribution of continuous variables such as age. The horizontal axis shows values of the data you are interested in while the vertical axis shows the number of times each value of the data appeared or the frequency of values within a fixed range. The graph on the right shows the distribution of age among 500 people. The number of people of each age is counted and shown as the height on the vertical axis. Thus, in this case, there are 30 people aged 12 (in other words, those aged between 12 years 0 days and 12 years 364 days), 70 aged 15, and 160 aged 20. Histograms have the advantage of enabling us to gain a visual understanding of the distribution of continuous data. If a histogram is symmetrical on either side of the peak and shaped like a bell, as in the graph above, we conclude that it follows normal distribution.

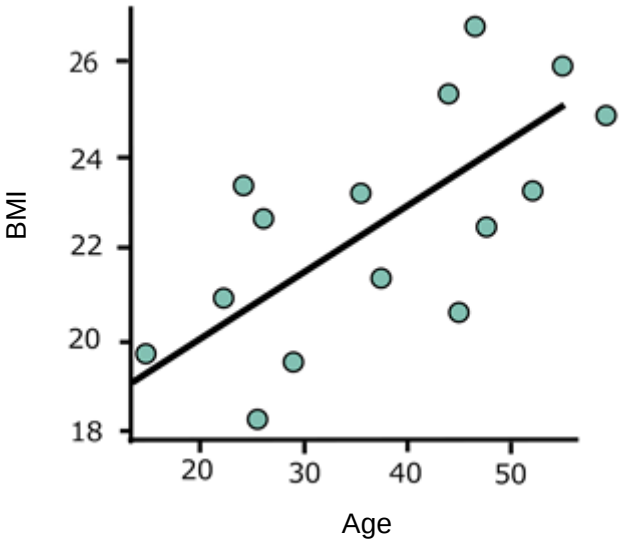
The range containing the values of individual variables can be made wider or narrower as needed in histograms. For example, in the graph below on the left, the data is divided into ranges of a single year. But it looks rather different when divided into five-year ranges (ages 16-20, 21-25, 26-30), as in the graph below on the right. Caution is required when determining the width of the range.





Scatter Diagrams

As the name suggests, scatter diagrams present data as points scattered across the diagram. Scatter diagrams are handy graphs for viewing the relationship between two continuous variables. For example, the graph on the right is a scatter diagram showing age and BMI. We can see the relationship that BMI increases with age.



This module was prepared by “Ethics Education Program on the Research Reliability Standards of International Medical Journals” (or “AMED International Journals Project”) with funding support from “Research and Development of Material/Program for Research Integrity Education” of the Japan Agency for Medical Research and Development. Please visit the APRIN website for the names of the experts who prepared and/or reviewed this module.

A number of checklists for submitting articles to international journals relevant to this module are provided here.
(Some have been translated loosely to aid understanding)

(1) Nature
(http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/2010128212513557501.pdf; visited on 2019.01.25)
(2) New England Journal of Medicine
(http://www.nejm.org/page/author-center/manuscript-submission#electronic; visited on 2019.01.25)
(3) Science (http://www.sciencemag.org/authors/science-editorial-policies; visited on 2019.01.25)
(4) The EMBO Journal (http://emboj.embopress.org/authorguide#statisticalanalysis; visited on 2019.01.25)
(5) JAMA (http://jamanetwork.com/journals/jama/pages/instructions-for-authors; visited on 2019.01.25)

(1) Nature

- Any distorted effect sizes (e.g. by truncation of y axis) are clearly labelled and justified
- Error bars are present on all graphs, where applicable
- All error bars are clearly labeled

(3) Science

- For continuous variables, distributions should be described using graphical displays such as scatterplots, boxplots, or histograms or by reporting measures of central tendency (e.g., mean or median) and dispersion (e.g., SD, interquartile range).

(4) The EMBO Journal

- Discussion of statistical methodology can be reported in the materials and methods section, but figure legends should contain a basic description of n, P and the analytical technique applied.
- Graphs must include clearly labeled error bars. Authors must state whether a number that follows the ± sign is a standard error (s.e.m.) or a standard deviation (s.d.)