Applied statistics in vascular surgery Part VI: Basic plots

Constantine N. Antonopoulos[1], Efthymios D. Avgerinos[2], John, Kakisis[1]

[1]Department of Vascular Surgery, Athens University Medical School, "Attikon" Hospital, Athens, Greece
[2]Department of Vascular Surgery, University of Pittsburgh, UPMC Presbyterian Hospital, Pittsburgh, PA

Abstract:
When working with data, it is essential to incorporate data visualization. By using appropriate illustrations, the researcher can greatly help the readers to better comprehend the statistical analysis. Among the most common types of graphs are bar graphs, histograms, box plots and error plots. A brief description of these basic statistical plots is presented.

INTRODUCTION

Descriptive statistics aims to summarize data involved in a study.[1] For example, a medical student collected data from 1,000 patients with infrarenal aortic aneurysm, who were treated with endovascular repair during a 10-year period. Of course, no one is going to read all this piece of data and if they did, they would not be able to capture any useful information from it. In that case, it becomes quite useful to use descriptive statistics and present key metrics such as mean, mode, median, range, variance, standard deviation, skewness and others, in order to provide a comprehensive way to summarize the available data.[2] Although this is the first step before proceeding with further analysis, numbers are not always perfectly conceivable by the readers. That is where graphs/plots can be of great use. They form a major component of almost all quantitative data analysis, by allowing researchers to provide a visual interpretation of complex data analysis and help readers to better comprehend the text and metrics.[3] Our aim in this paper is to present the most common plots used in vascular surgery literature.

WHICH GRAPH TO CHOOSE? BAR GRAPH VS HISTOGRAM

A good graph can convey information quickly and easily to the reader. The type of data often determines what graph is appropriate to use. One of the most widely used plots is the ***bar graph (also called bar chart)***. It consists of rectangular bars or columns (called bins) that represent the total amount of observations in the data for each category. Data can be displayed in vertical bars, horizontal bars, comparative bars, which show a comparison between values, or stacked bars, which contain multiple types of information.

Author for correspondence:
Constantine N. Antonopoulos
Department of Vascular Surgery, Athens University Medical School, Attikon University Hospital, Athens, Greece
E-mail: kostas.antonopoulos@gmail.com
ISSN 1106-7237/ 2020 Hellenic Society of Vascular and Endovascular Surgery Published by Rotonda Publications

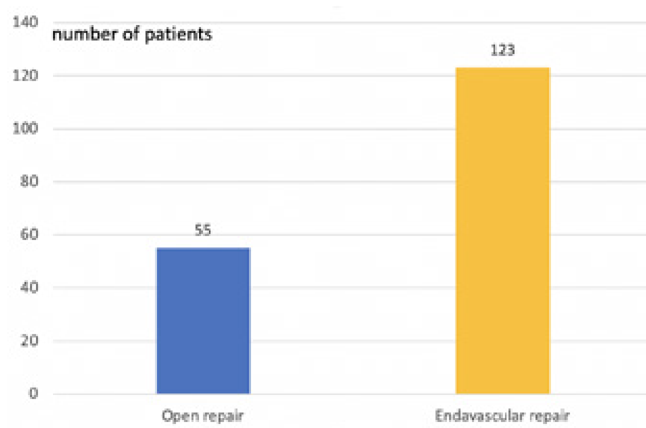

Figure 1. Example of a bar chart. Distribution of patients with abdominal aortic aneurysm according to treatment option

A simple bar graph (Figure 1) has two axes; one axis describes the types of categories being compared (eg. open vs endovascular repair of patients with abdominal aortic aneurysm) and the other has numerical values (eg 55 patients vs 123 patients, respectively) which represent the values of the data. The length of each bar is proportionate to the numerical value or percentage that it represents. A similar-looking graph is the ***histogram***. A histogram (Figure 2) represents the frequency distribution of continuous variables (eg. age of patients with abdominal aortic aneurysm, who were treated with endovascular repair) and it is displayed in such a way that there is no gap between the bars. Importantly, the researcher should know that histogram is used for displaying continuous data, whereas bar graphs display categorical data.[2]

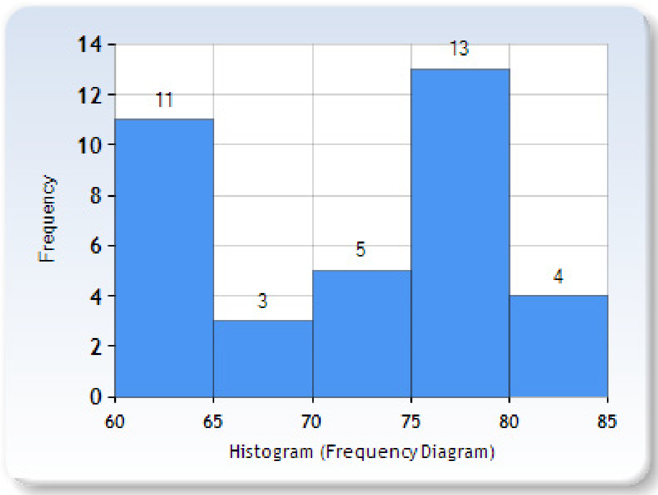

Figure 2. Example of histogram. Age (years) distribution for patients with abdominal aortic aneurysm, who were treated with endovascular repair.

WHEN WE NEED TO ADD ADDITIONAL INFORMATION IN A PLOT; THE BOX PLOT

Although the bar graphs and histograms can illustrate important information about the data being investigated, usually there is much more information about the variables that we may want to know, which cannot be found in a bar chart or a histogram. For example, we may know the age distribution of patients with abdominal aortic aneurysm, who were treated with endovascular repair, but we do not know whether or not the data are skewed, if there are outliers, how tightly the data are clustered around the mean.[4] In that case, the reader should look for a ***box plot*** (also called ***box-and-whisker plots***).

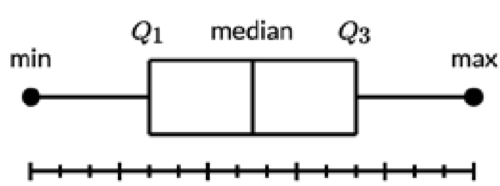

Figure 3. A typical box plot, with minimum (min) value, maximum (max) value, 25th percentile (Q1), 50th percentile (Q2 or median) and 75th percentile (Q3).

A box plot (Figure 3) contains five values: i) the 1st quartile, ii) the median, which is the value that divides the data in half and represents the 2nd quartile, iii) the 3rd quartile, iv) the minimum value and v) the maximum value of the data. Horizontal or vertical lines and a rectangular box are used to illustrate these five values. The one end of the box represents the 1st quartile (Q1, 25th percentile), the other end of the box represents the 3rd quartile (Q3, 75th percentile), while the median or 2nd quartile (Q2, 50th percentile) is represented by a line, somewhere between the ends of the rectangular box. One of the horizontal or vertical lines represent the minimum value, while the other represents the maximum value. The "whiskers" extend from the ends of the box to the smallest and largest data values.[3] Important information provided by the box plot is the graphical illustration of the median, which unlike the mean, it is unaffected by extreme values at one end or the other. Data are more or less symmetrically distributed about the median, if the median falls near the center of the box. In case the median falls near the bottom the data are positively skewed, while if the median is closer to the top of the box, data are negatively skewed. In the data of the example of Figure 2, a box plot (Figure 4) shows that data are not symmetrically skewed, while median is 71 years, with Q1=63 years and Q3=75 years. The difference between the Q3 and Q1 (75 - 63 = 12) is the inter-quartile range (IQR), which contains the middle 50% of the data and can provide an estimation of the dispersion of the data.

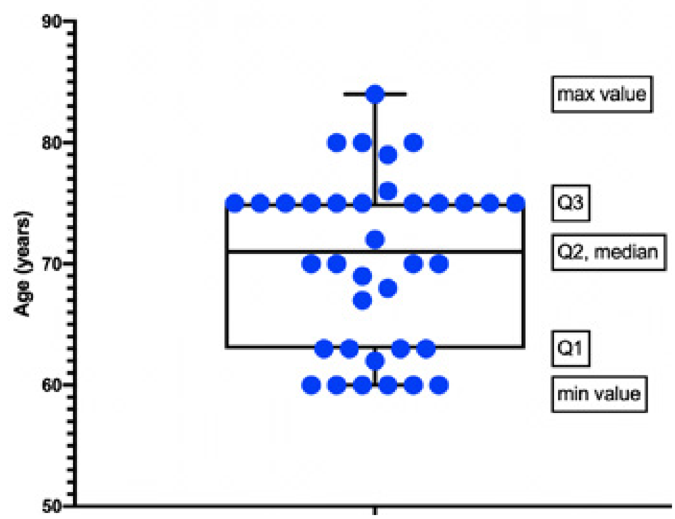

Figure 4. Box plot illustrating data of Figure 2.

ERROR PLOTS

Sometimes, measurement of uncertainty is of interest in medical research. Standard deviation (SD) provides information about how the data is distributed about the mean value and it is a good way to measure this uncertainty.[2,4] In this case, the ***error plot*** can graphically present mean with SD or other measures of uncertainty (eg. 95% confidence intervals or standard error). Figure 5 shows the error plot of data also illustrated in Figure 2. In this example, mean age of patients is 70.4 years, with SD of 7.0 years. Error plots can be applied to other graphs, such as bar graphs in order to provide an additional layer of detail on the presented data. Error Bars are illustrated as cap-tipped lines that extend from the center of the plotted data point and the length of the error lines helps indicate the measure of uncertainty (eg. SD; Figure 6).

Plots give a good, quick picture of the data. As one of the most influencing statisticians, John Tukey, said, "*the researcher should never begin analyzing data before he/she has visualized them in some way*".[2,3]

No conflict of interest.

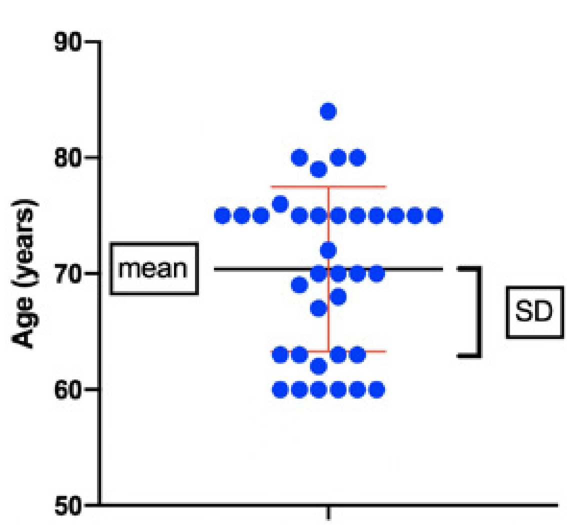

Figure 5. Error plot illustrating data of Figure 2

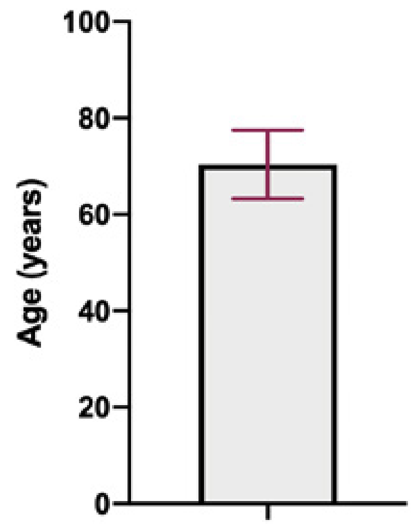

Figure 6. Error plot with bar chart illustrating data of Figure 2

REFERENCES

1 Antonopoulos C, Kakisis J. Applied statistics in vascular surgery Part 2: Correlation analysis and its common misconceptions. HJVES. 2019;1(2):88-90.

2 Tukey JW. Exploratory Data Analysis. Boston,MA: Addison-Wesley; 1977.

3 Streiner DL. Statistics Commentary Series: Commentary No. 24: Box Plots. J Clin Psychopharmacol. 2018;38(1):5-6.

4 Antonopoulos C, Kakisis J. Applied statistics in vascular surgery Part 1: Choosing between parametric and non-parametric tests. HJVES. 2019;1(1):36-7.