Statistics Corner

Types of biological variables

Shreemathi S. Mayya, Ashma D Monteiro, Sachit Ganapathy

Department of Statistics, Manipal University, Manipal-576104, Karnataka, India

Correspondence to: Dr. Shreemathi S. Mayya, Associate Professor (Sr. Scale). Department of Statistics. Manipal University, Manipal-576104, Karnataka, India. Email: shreemathi.mayya@manipal.edu.

Abstract: Identification and description of variables used in any study is a necessary component in biomedical research. Statistical analyses rely on the type of variables that are involved in the study. In this short article, we introduce the different types of biological variables. A researcher has to be familiar with the type of variable he/she is dealing with in his/her research to decide about appropriate graphs/diagrams, summary measures and statistical analysis.

Keywords: Biological variables; discrete variables; continuous variables; categorical variables

Submitted Apr 19, 2017. Accepted for publication May 09, 2017.
doi: 10.21037/jtd.2017.05.75
View this article at: http://dx.doi.org/10.21037/jtd.2017.05.75

Introduction

Research question is the initial and integral step in any research work. Depending on the research questions to be answered and the data available, researchers decide about the statistical methods to be used for analysis. Researchers have to be acquainted with the variety of variables involved in their study to choose appropriate diagrams/graphs and summary measures for presentation, and valid statistical tests for the analysis of data.

Information collected about a sample of subjects (often patients) comprises characteristics which vary among the subjects. Any characteristic, which varies from individual to individual is called a variable (1). The characteristics such as age, sex, height, weight, body mass index (BMI), blood group, body temperature, blood glucose level, blood pressure, heart rate, number of teeth, severity of disease (mild, moderate, severe) etc. are some of the examples for biological variables in research. A basic distinction in the nature between these variables is their quantitative or qualitative (categorical) measurements (1,2).

Quantitative variables

Quantitative variables are those characteristics which can be a count or measured numerically. They can be continuous or discrete. Continuous variable can theoretically take infinitely many values in a given range. This means that, we can always find an intermediate value between any two values, however close they are. For example, in a given range of 5–10 cm length one can write infinitely many values like 5, 5.1, 5.12, 5.01, 5.003 cm… etc, depending on the extent of accuracy decided by the researcher. Height of a person, weight, age, arm length, blood pressure, temperature, glucose level are some of the examples for continuous variable. Here the obtained measurements can take any value in a given range.

Discrete variable (discontinuous variable) can take only specified number of values in a given range. For example, number of children per family in a given range of 0–5 can be 0, 1, 2, 3, 4 and 5. No more values in this range can be written. Number of visits to hospital in a year, number of children in a family, number of admitted patients in a hospital ward, number of missing teeth etc. are some of the examples for discrete variables. Discrete variables are usually counts.

Qualitative variables

Qualitative (categorical) variables are those characteristics which are not numerically measurable. These variables are either nominal (no natural ordering) or ordinal (ordered categories). Usually, for the purpose of data entry and analysis using software, categories are coded assigning

© Journal of Thoracic Disease. All rights reserved.

numerical values.

Nominal variables allow for only classification or categorization based on some distinctively different characteristic, but we cannot rank order those categories. Typical examples of nominal variables are sex, religion, blood group, symptoms of disease, cause of death etc. Numerical values assigned to different categories are useful for the purpose of identification only (e.g., 1= male, 2= female). When a qualitative variable has only two categories (alive/dead, male/female, diabetic/non-diabetic), it is called a binary or dichotomous variable. Nominal variables are summarized through counting (frequency) and expressing proportion of each category (percentage).

Ordinal variables allow us to rank order the categories in terms of which category has less and which category has more of the quality represented by the variable, but the distances between categories are not known. A typical example of an ordinal variable in medicine is the stages of a diseases (stage I to stage IV). For example, we know that "stage I" is less severe than "stage II" of a disease but we cannot tell the exact difference between the two stages. Socioeconomic status of families (low, middle and high socio-economic status), BMI category (underweight, normal, overweight, obese), disease condition (deteriorated, same, improved), Pain score etc. are a few examples for ordinal variables. Numerical values assigned for various categories are useful for identification as well as rank ordering (e.g., 1= low, 2= middle and 3= high income group). Ordinal variables are summarized through counting (frequency) and expressing proportion of each category (percentage).

Categorizing a continuous variable

Quantitative variables are often converted to categorical ones using "Cut-points". Instead of presenting the mean fasting glucose level of male and female subjects, one may prefer to present the proportion of diabetics in male and female population using a fasting glucose level of 110 mg/dL as the cut-point to categorize the subjects as diabetic/ non-diabetic. However, categorizing a continuous variable lead to loss of information (3). For example, while categorizing, subjects with fasting glucose level of 85 and 109 mg/dL are treated as equal and classified as non-diabetic. Similarly, subjects with glucose level 111 and 150 mg/dL are classified as diabetic. The difference in the values will not be noticed while presenting only the number of diabetic and non-diabetic cases.

Dealing with Likert type data

Likert scale is developed with a principle of measuring attitudes by asking people to respond to a series of statements about a topic, in terms of the extent to which they agree with them (4). A statement (Likert item) such as: "It's important for all biologists to learn statistics" can be asked to be rated as 1= strongly disagree, 2= disagree, 3= neither agree nor disagree, 4= agree, or 5= strongly agree or sometimes on seven values instead of five, including "very strongly disagree" and "very strongly agree". Variables measured on Likert item are a type of ordinal variables. Likert scale is the result of adding together the scores on several Likert items. Likert scale may be treated as a continuous variable. Descriptive and inferential statistics depend on the distribution of scores, symmetric or skewed.

Presentation of data

Qualitative variables

Qualitative data (nominal or ordinal variable) may be presented in the form of frequency tables. We count the number of subjects/units in each category of the variable along with percentage and present the numbers and percentages in a table. E.g., we summarize Blood group distribution of 100 subjects in the form of a table showing blood group and corresponding frequency along with percentages. If we have the data for two categorical variables, data may be presented in the form of a contingency table showing frequency and percentages.

As ordinal variables are also categorical variables with a pre-determined order, the descriptive measures such as frequency and percentage has to be reported when the number of categories are few. In addition, median, inter-quartile range along with maximum and the minimum value is considered appropriate for summarizing ordinal variables.

Nominal data and ordinal data with limited number of categories can also be presented in a diagrammatic form, such as a bar chart and pie chart. In a bar chart, length of the bars represents the frequency or relative frequency of each category of the variable. Usually the bars are of equal width and there is a space between them. A pie chart is essentially a circle divided into segments with the area of each proportional to the observed frequency in each category of the variable. Total area represents the total frequency.

Quantitative variables

Mean and standard deviation are appropriate summary measures for continuous variables with symmetrical distributions. Median and inter-quartile range are to be computed to summarize quantitative variables with skewed distributions. Range is informative if used as a supplement to standard deviation or inter quartile range. Discrete variables may be summarized and analyzed either as a continuous variable or as an ordinal variable depending on the number of distinct values.

Quantitative data can be represented graphically by means of a histogram. Histogram is useful to decide about the shape of the distribution, symmetrical or skewed. But, with small samples, histogram may not be useful to identify the shape. As a rule of thumb, if the mean is smaller than twice the standard deviation the data are likely to be skewed for variable with positive values (5). Quantitative data can also be displayed as stem & leaf plots, dot plots, box & whisker plots and scatterplots, depending on the situation (6).

Analysis of data

Type of the variables decides the type of statistical analyses to be performed, parametric or non-parametric. Parametric methods, such as *t*-tests, ANOVA, Pearson's correlation, and regression, require the assumption that the data follow a normal distribution and that variances of the distributions are equal. Frequently used nonparametric methods are Mann-Whitney or Wilcoxon rank sum test, Wilcoxon signed rank test and rank correlation. Non-parametric methods, make no assumptions about the distribution of the data; they use the rank order of observations rather than actual measurements (7). Chi-square test (or Fisher's exact test if the numbers are very small) is the most often used method to compare categorical data. Failure to pay attention to assumptions and their implications can lead to increase in type I or type II errors.

We analyze data from similar studies, completely differently depending on the type of variable involved. For example, let us say that our target population is 50+ age group in a certain population and we have measured the variable systolic blood pressure in a sample of 40 male and 40 female subjects, and our null hypothesis is "Male and female population have the same systolic blood pressure". We would compare the mean blood pressure in males and females with a two-sample *t*-test (parametric test). If the variable is converted to hypertension status (hypertensive/normal), it is a nominal variable, and we would compare the hypertension frequencies in males and females with a Chi-square test (non-parametric test). We find smaller P value for *t*-test compared to chi-square test. An important message that we try to convey here is that, statistical tests will have more power for a continuous variable than the corresponding nominal or ordinal variables (2). In other words, to achieve the same power as that of a parametric test, non-parametric tests require larger sample size than a parametric test. Therefore, one may categorize the data for the purpose of presentation (e.g., hypertensive/normal), but not for statistical analysis (3).

Detailed discussion of various tests is out of the scope of this article. Campbell & Swinscow (2) have summarized the tests suitable for various types of variables in a single table. For computation procedure and more details about various parametric tests, researchers may refer some standard text books (1,3,8). For a good discussion of a number of nonparametric tests readers may refer Siegel and Castellan (9) and Conover (10).

Conclusions

The type of descriptive and analytical measures to be used in data summarization and analysis, all depend on the type of variables. Therefore, to obtain the relevant measures for dataset at hand, we recommend the researchers to study the characteristics of data (categorical, quantitative) and shape of the frequency distribution (symmetrical bell shaped, skewed) before deciding about the descriptive measures, graphs and diagrams, and statistical tests suitable for the presentation and analysis of data.

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Daniel WW. editor. Biostatistics: A foundation for analysis in the health sciences. 6th ed. New York: John Wiley & Sons, 1995.
2. Campbell MJ, Swinscow TD. editors. Statistics at Square One. 11th ed. Oxford: Wiley-Blackwell, 2009.
3. Altman DG, Bland JM. The cost of dichotomizing continuous variables. BMJ 2006;332:1080.

4. McDonald JH. editor. Handbook of biological statistics. Baltimore, MD: Sparky House Publishing, 2009.
5. Altman DG, Bland JM. Detecting skewness from summary information. BMJ 1996;313:1200.
6. Freeman JV, Walters SJ, Campbell MJ. editors. How to display data. Oxford: Blackwell, 2008.
7. Altman DG, Bland JM. Parametric v non-parametric methods for data analysis. BMJ 2009;338:a3167.
8. Bland M. editor. An Introduction to Medical Statistics. 3rd ed. Oxford University Press; 2000.
9. Siegel S, Castellan NJ. editors. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill, 1988.
10. Conover WJ. editor. Practical nonparametric statistics. 3rd ed. New York: John Wiley, 1998.

Cite this article as: Mayya SS, Monteiro AD, Ganapathy S. Types of biological variables. J Thorac Dis 2017;9(6):1730-1733. doi: 10.21037/jtd.2017.05.75