Understanding data for medical statistics

Sonali Sarkar

Department of PSM, Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER), Puducherry, India

INTRODUCTION

Statistics remains an enigma for physicians, even though the knowledge is essential not only for research but also for understanding and interpreting information relevant to the practice of medical science. At the base of all high-end statistics presently in use in the medical field is the understanding of the type of data dealt with. The data generated from the observation of and intervention in the human body are as diverse as all the anatomical, physiological, and biochemical parameters. Added to the complexity of human biology is the environment with which humans interact and, therefore, has very intricate relationship with health. In the endeavor to understand the causes of illnesses and finding appropriate solutions for the same, man has, from time immemorial, generated volumes of information based on data. But with the advent of modern medicine and sophisticated technology both for investigations and treatment of disease conditions, generating evidence for the use of simplest to most complicated procedures has come to be known as evidence-based medicine. Therefore, all medical professionals are expected to have a sound knowledge in the medical statistics to be able to practice their craft efficiently. As the foundation of medical statistics is knowledge of the data generated by day-to-day practice and research, this paper attempts at understanding the data and choosing the appropriate types of statistics applicable from the viewpoint of a medical undergraduate.

Data and variables

Data (plural) are measurements or observations and datum (singular) is a single measurement or observation.[1] Statistics is used in two ways – descriptive, concerned with presentation, organization, and summarization of data, and inferential, concerning generalization from a sample of data to a larger group of subjects.[2] Both types of statistics using data deal with what are called variables. A variable is any characteristic measured that varies from individual to individual. Therefore, blood pressure (BP) is a variable; BP 120/80 mm Hg of a student in a class is a datum and the BP of all the students in the class is the data set.

The variables can be discrete or continuous. A discrete variable is one which can have only one of a limited set of values and can be measured in whole numbers,[2] for example, the number of children and pulse. Continuous variable can take any value within a defined range,[3] for example, height (there can be a measurement between 145 and 146 cm, or if finer measurement is possible, between 1450 and 1451 mm, or even further, fractional values are possible, e.g. 145.5 cm).[3] Most biochemical measurements such as serum cholesterol, albumin, urea, or creatinine are continuous data. Further, depending on the scales of measurement, the variables can be of the following types:

1. Nominal variable is one where the values fall into unordered categories[3] (e.g. gender, religion, color of the eyes, or blood groups). Even though numbers are used to represent the categories, like 0 for males and 1 for females, for performance of statistical procedures in computers, it is important to remember that there is no order in these categories and there cannot be an average of 0.5 genders.

2. Ordinal variable has ordered categories, but the differences between the categories cannot be considered equal. For example, though numbers are used to denote the stages of cancer, the difference in the severity of disease between stages I and II may not be the same as between stages II and III. The numbers only signify the order.[2] Other examples are power of muscles, severity of dyspnea, and class of occupation.

3. An interval variable has all the characteristics of an ordinal variable, and also, the differences or distances between any two values on the scale are equal, but the zero point is arbitrary. Examples are temperature measured as Celsius and the intelligence quotient.

Address for correspondence: Dr. Sonali Sarkar, Department of PSM, Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER), Puducherry, India. E-mail: sarkarsonaligh@gmail.com
Nominal Measured as categories data, summarization and appropriate presentation

For easy interpretation of information from the summarization and presentation of data called quantitative.[4] Table 1. One thing to remember is variables, which can either be discrete or continuous, are otherwise known as qualitative, and interval and ratio quantitative. Nominal and ordinal variables are called categorical variables. In experimental research, one variable is manipulated and effects are observed in another variable. The outcome of interest, which changes in response to the intervention, is the dependent variable and the intervention is the independent variable.[3] Hypertension is dependent on factors like salt intake, exercise, and stress. Therefore, hypertension is a dependent variable and salt intake, exercise, and stress are independent variables. Even in studies analyzing the association of the risk factors with the disease outcomes, risk factors are considered independent variables, the disease being dependent on the changes in their levels.

Summarization and presentation of data

For easy interpretation of information from the data, summarization and appropriate presentation is required. This depends on the type of variables. Data can be summarized in tabular or graphical format. Both quantitative and qualitative data can be presented as frequency distribution tables. The classes can be categories as such, like the number of people of different religions, or combined, as in the categories of education for qualitative data. The classes are sections of a range of values in quantitative data, for example, heights ranging from 150 to 155 cm or age groups of 0–5, 5–10, etc. Among the graphs, pie chart is an excellent way of presenting categorical data, showing comparison of frequencies as proportions. Most commonly, the causes of infant and maternal mortality are presented as pie charts. The limitation is that only one variable can be shown at a time.[4] Bar charts display the frequency distribution of nominal and ordinal data. As the data are discrete, the bars are separated by gaps of equal size. For ordinal data, they are arranged in order. For nominal data, the practice is to arrange them in decreasing frequency from left to right.[4] Figure 1 is an example of a bar graph showing the district wise population distribution of Tamil Nadu. A histogram resembles a bar chart, but has no gaps in between the bars, as the data presented are continuous. When the midpoints of the histograms are joined by straight lines, a frequency polygon is formed. A line graph illustrates the relationship between two continuous quantities and is commonly used to show change in data over time.[3,4] This represents the trend of events with time, for example, the sex ratio in India over the successive census years [Figure 2]. Numerical summarization of data is commonly done by describing the center of the set of data around which most observations are clustered, using what is known as measures of central tendency. Mean is the most frequently used summary measure for both discrete and continuous measurements.[3] But mean is affected by the extreme values. Median is the central value of the distribution such that half the values are less than or equal to it and the other half is greater than or equal to it. Median can be used as a summary

| Table 1: Classifications of variables with examples |
|---------------|-----------------|---------------|
| Types of variables | Description | Examples | Types of data |
| Nominal | Measured as categories without any order | Alive or dead | Qualitative data |
| Ordinal | Categories are ordered, but differences between them not equal | Socio-economic status | Quantitative data |
| Interval | Differences between values are equal, but zero is arbitrary | Intelligence quotient | Quantitative data |
| Ratio | Differences between values are equal with true zero. Ratios of measurements also meaningful | Blood pressure |

Figure 1: Bar graph showing district wise population distribution of Tamil Nadu
measure for ordinal as well as discrete and continuous data. But median is not as sensitive to the value of each measurement as is mean. Therefore, “mean” which has nicer mathematical properties is more useful for further statistical comparison methods. Mode is another measure of central tendency, which is the most commonly occurring value.

**Distribution of data**

On measuring any variable in a large number of individuals, we call the pattern of values obtained a distribution. In biology, most continuous variables in interval or ratio scale are distributed normally when the number of measurements is fairly large (i.e. more than 30), and the frequency polygon tends to get smoother [Figure 3]. A normal distribution is a bell-shaped symmetrical curve where the mean, median, and mode coincide as 50% of the values are above the mean and most of the values are close to the mean. In addition to all these, in a normally distributed data, the mean and variance are not dependent on each other. Even if the mean changes, the proportion of values that lie between the mean ± 1SD (68.2%), mean ± 2SD (95.4%), and mean ± 3SD (99.8%) remains the same in a normal distribution. The name “normal” is a little unfortunate as the distributions, which do not fit into this shape, are in no way abnormal. Contrary to the belief, many variables in medical science may not follow normal distribution, for example, the number of children a family can have or the timing of patients’ arrival at the emergency department. Lack of symmetry in the frequency distribution is called skewness. A frequency distribution that has a long tail extending to the right is known as positive skewness or skewed to the right and the one that has a long tail extending to the left is known as negative skewness or skewed to the left [Figure 4]. But in the history of statistical methods, the first techniques of inference that were developed and the ones most commonly used, including z-test, Student’s t-test, analysis of variance (ANOVA), correlation, and regression, are based on assumptions that the data follow a normal distribution. If the assumption of normality is violated, interpretation and inference may not be reliable or valid. Therefore, it is necessary to assess the normality of data before proceeding with these statistical procedures. Normality of data can be assessed either visually by use of normal plots, the numerical methods which include the skewness and kurtosis coefficients, or by using significance tests. Though visual inspection of data is less reliable, it is preferable that normality be assessed both visually and through normality tests. The normality tests are supplementary to graphical assessment. The frequency distribution (histogram), stem-and-leaf plot, boxplot, P–P plot (probability–probability plot), and Q–Q plot (quantile–quantile plot) are used for checking the normality visually. The most common tests used for the assessment of normality are Kolmogorov–Smirnov (K–S) test, Lilliefors corrected K–S test, Anderson–Darling test, and Shapiro–Wilk test. If the test is significant, the distribution is non-normal. But these tests should be used cautiously for small samples (30 and below), as they have less power for small sample sizes. Of all these tests, Shapiro–Wilk test is the most powerful test for all types of distribution and sample sizes, whereas Kolmogorov–Smirnov test is the least powerful test.

**Application of type and distribution of data**

Further, statistical techniques used in inferential statistics, to be dealt with later, are classified as parametric and non-parametric. Parametric tests involve measurement of population parameters (mean). They require the measurements to be at least on interval scale [Figure 5]. The assumptions are that data are approximately normally distributed, drawn from a normally distributed population having the same variance (σ²). The most commonly used parametric tests are z-test, Student’s t-test, ANOVA, and correlation.

Figure 2: Line graph showing trend of sex ratio in India since 1901

Figure 3: Frequency polygon showing systolic blood pressure with increasing n
and regression. In situations where the assumptions of parametric tests are not fulfilled, in other words,
a. When the variables measured are ordinal or nominal,
b. They do not involve population parameters, but are probability distributions, and
c. The data do not follow a normal distribution, non-parametric tests should be performed.\(^{[12]}\) Some common non-parametric tests are Mann–Whitney U, Wilcoxon signed-rank test, and Spearman’s Rank test.

**CONCLUSION**

Knowledge of data and variables is crucial for performing any statistical procedure. Summarization, presentation, and further inferential statistics all depend upon the types of variables in the study and their distribution. There are different ways of classifying variables. Choice of statistical test is made based on the knowledge about quantitative or qualitative data. All parametric procedures are based on the assumption that the data are normally distributed. Therefore, it is important to assess for normality of data before performing these tests. Assessment of normality can be done visually and also through use of tests. Shapiro–Wilk test is the most powerful of the commonly used tests, but should be interpreted with caution when used for small samples. There are a set of procedures called non-parametric tests for data which do not fulfill the criteria of normality.

**ACKNOWLEDGMENT**

The author thanks Dr. Mahalakshmy T, Assistant Professor in the Department of PSM, JIPMER for giving suggestions to improve the quality of the paper.

**REFERENCES**

1. Gravetter FJ, Wallnau LB. Statistics for Behavioural Sciences. 5th ed. Australia: Wadsworth Thomson Learning; 2000. p. 7, 583-605, 637-56.
2. Norman GR, Steiner DL. Biostatistics: The Bare Essentials. 2nd ed. London: BC Decker Inc; 2000. p. 2-5.
3. Pagano M, Gauvreau K. Biostatistics. 2nd ed. Australia: Duxbury, Thomson Learning; 2000. p. 7-11, 38-43.
4. Driscoll P, Lecky F, Crosby M. An Introduction to everyday statistics - 1. J Accid Emerg Med 2000;17:205-11.
5. Bland M. An Introduction to Medical Statistics. 3rd ed. Oxford: Oxford University Press; 2000. p. 46-62.
6. Altman DG, Bland JM. Statistics notes: The normal distribution. BMJ 1995;310:298.
7. Hill AB, Hill ID. Bradford Hill’s Principles of Medical Statistics. 12th ed. New Delhi: BI Publications Pvt. Ltd; 1993. p. 81.
8. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. JOSMA 2011;2:21-33.
9. Field A. Discovering Statistics Using SPSS. 3rd ed. London: SAGE Publications Ltd; 2009. p. 822.
10. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: A guide for non-statisticians. Int J Endocrinol Metab 2012;10:486-9.
11. Elliott AC, Woodward WA. Statistical Analysis Quick Reference Guidebook with SPSS Examples. 1st ed. London: Sage Publications; 2007.
12. Siegel S, Castellan NJ. Nonparametric Statistics for the Behavioral Sciences. 2nd ed. New York, NY: McGraw-Hill, Siegel and Castellan; 1988. p. 35.
13. Six Sigma Material. Data Classification. Available from: http://www.six-sigma-material.com/Data-Classification.html. [Last accessed on 2014 Apr 04].

How to cite this article: Sarkar S. Understanding data for medical statistics. Int J Adv Med Res 2014;1:30-3.

Source of Support: Nil, Conflict of Interest: None declared.