Statistics Corner

How to describe univariate data

Stefania Canova[1], Diego Luigi Cortinovis[1], Federico Ambrogi[2]

[1]Department of Medical Oncology, San Gerardo Hospital Monza, Monza 20900, Italy; [2]Department of Clinical Sciences and Community Health, Medical Statistics, Biometry and Bioinformatics, University of Milan, Milan 20133, Italy

Correspondence to: Stefania Canova. Department of Medical Oncology, San Gerardo Hospital Monza, via Pergolesi 33, Monza 20900, Italy. Email: s.canova@asst-monza.it.

Abstract: Univariate analysis has the purpose to describe a single variable distribution in one sample. It is the first important step of every clinical trial. In this short review, we focus on this analysis, the methods that authors should use to report this type of data, information that they should not miss and mistakes that they must avoid.

Keywords: Univariate; variable; graph; survival

Submitted Apr 24, 2017. Accepted for publication May 09, 2017.
doi: 10.21037/jtd.2017.05.80
View this article at: http://dx.doi.org/10.21037/jtd.2017.05.80

Introduction

A variable is any characteristic that can be observed or measured on a subject. In clinical studies a sample of subjects is collected and some variables of interest are considered. Univariate descriptive analysis of a single variable has the purpose to describe the variable distribution in one sample and it is the first important step of every clinical study.

Variables

Authors should identify the type and number of examined variables, as well as missing data for each variable.

Variables can be categorical or numerical.

Categorical or qualitative data can be binary, nominal or ordinal. Binary variables are characterized by only two possible categories, for example male/female, dead/alive.

When there are more than two categories/classes, it is important to distinguish between nominal variables, such as blood group, and ordinal variables, such as disease stage.

Categorical data should be presented not only giving percentages for each class, but also absolute frequencies.

Numerical or quantitative data can be broadly divided into discrete or continuous. Discrete variables arise mainly from counts, such as the number of words in a sentence, the number of components of a family, while continuous variables arise mainly from measurements, such as height, blood pressure or tumour size. Such variables are continuous as, in principle, any value (in the admissible range of measurement) can be taken, while discrete variables can take only certain numerical values. For continuous variables, the only limitation arises from the accuracy of the instrument of measurement. Discrete variables are sometimes treated as continuous, when the number of possible values is very large. Numerical variables can be transformed in categorical by grouping values into two or more categories to simplify the comprehension of results (but not in general the analysis). Categorization of numerical variables results in loss of information, especially with two groups, and should be done with caution.

Authors should always specify how categorization was obtained, in particular how the choice of cut-points was made, if on the basis of previous analyses or arbitrarily by the authors (using median and quartiles for example). In absence of previous analyses, theoretical or clinical arguments should justify categorization to avoid biases and to obtain reliable results (1).

Researchers should avoid arbitrary cut-points and should prefer categorization into at least three groups avoiding dichotomization.

Frequency distribution and central tendency

A variable can be described by its frequency distribution

that reports the absolute (or relative to the total) number of times a specific value/class of a variable is observed in the sample. Continuous variables should be divided in classes for this purpose. For ordered nominal variables and for numerical variables, the cumulative frequencies can also be computed. Instead of tables, graphs can be used to describe the distributions. Pie charts, where each slice represents the proportion of observations of each category, are useful for nominal data (without ordering), while bar charts can be used for ordinal categorical data or for discrete data. Histograms must be used for continuous data.

Another useful possibility is a box-whisker plot which is composed of a box representing upper and lower quartiles, a central line indicating the median, while the whiskers represents extreme centiles, with extreme values showed above and below the whiskers.

Due to space limitations, tables reporting summary values for each distribution are often used to describe the variables considered in the study. Before summarizing the distribution with few numbers, it is always necessary to look at the whole distribution. If the shape of the distribution is approximately symmetric (like for the Gaussian distribution), the mean and the standard deviation (SD) can be used, reporting the results as mean (SD), and avoiding the ±. If the shape of the distribution is skewed, it is better to use the median and the quartiles. A general recommendation could be to report, in every case, mean, median, SD and the quartiles. Mean, median and mode are very similar in case of symmetric distributions. In case of skewed distributions, median is less influenced by extreme observations.

Another summary measure is the mode that is the most frequent observation. This is rarely useful for numerical variables, whereas it is the only measure to be used with categorical variables. When describing categorical variables in tables, not only percentages for each class, but also their absolute frequencies, should always be reported

SD should not be confused with standard error (SE).

SE is a measure of the dispersion of the sample means around the population mean and is used for inferential (not descriptive) purposes. SE is the ratio between the SD and the square root of the sample size (n) (2).

SD is especially useful when the distribution is approximately Gaussian, as in the Gaussian case about 95% of observations are included within two SD of the mean (3).

Rounding numbers

The general rule is to present summary statistics at no more than one decimal place than the raw data (4). In the case of percentages, it is often enough to approximate at one decimal place. Rounding should be done only in the final report, not during analysis, to maintain precision and not to lose information.

According to one commonly used rule, excess digits are removed if the first one in excess is less than five. In case the first excess digit is more or equal to five, the last not in excess is increased by one. Be aware that computers output always contains spurious results that should be rounded according to the original accuracy of the measurements.

Time to event data

In many clinical studies, the time to onset of an event is of interest. Censored data refer to subjects included in the analysis but for whom the event of interest has not yet been observed when the study is closed (3). For example, in survival studies censored data include both patients still alive at the end of follow-up and patients lost during follow-up.

When reporting the number of events, it is advisable to avoid calculating the percentage with respect to the total number of subjects unless all subjects were followed-up for the same amount of time.

The completeness of follow-up is an indicator of study quality. Therefore, researchers should report the number of subjects lost to follow-up in addition to the follow-up range (minimum and maximum). The Kaplan-Meier method is suitable to describe the distribution of such a variable taking correctly into consideration the follow-up time and censored observations (5).

Authors should graphically report the number of subjects at risk. Moreover, they should indicate censoring times and confidence intervals, as well as which software was used to perform analyses.

Conclusions

When reporting study results, authors should keep in mind the advice of the International Committee of Medical Journal Editors (1991): “Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results.”

Acknowledgements

None.

Footnote

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

References

1. Naggara O, Raymond J, Guilbert F, et al. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. AJNR Am J Neuroradiol 2011;32:437-40.
2. Altman DG. editor. Practical statistics for medical research. London: Chapman and Hall, 2003.
3. Machin D, Campbell MJ, Walters SJ. editors. Medical Statistics. England: John Wiley & Sons, Ltd., 2007.
4. Altman DG, Bland JM. Presentation of numerical data. BMJ 1996;312:572.
5. Altman DG, De Stavola BL, Love SB, et al. Review of survival analyses published in cancer journals. Br J Cancer 1995;72:511-8.

Cite this article as: Canova S, Cortinovis DL, Ambrogi F. How to describe univariate data. J Thorac Dis 2017;9(6):1741-1743. doi: 10.21037/jtd.2017.05.80

