BIOSTATISTICS SERIES

Statistics Corner: Reporting Descriptive Statistics

Kamal Kishore[1], Rakesh Kapoor[2]

Reality check

Most of the studies in medical research collect raw data. However, raw data need to be processed and presented in meaningful numerical summaries to get information. Data processing can be broadly segmented into descriptive and inferential statistics. There are various descriptive measures to report data. However, the mean and standard deviation are the most popular reporting measures. Therefore, many times, investigator face dilemmas to either report mean and SD or some other appropriate descriptive measures?

- Are there any guidelines to report descriptive measure?
- What are the best descriptive measures to report continuous data?
- What are the best descriptive measures to report ordinal data?
- What are the best descriptive measures to report nominal data?

Journal of Postgraduate Medicine, Education and Research (2020): 10.5005/jp-journals-10028-1364

"Statistics are the tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the science of man"

—Sir Francis Galton

Introduction

Accuracy and precision are the twin hallmark of medical research. However, there are several instances where medical researchers face uncertainties (prescribing medicine A or B, wait or operate). An investigator must make informed and objective decisions in the wake of ambiguity. The statistics play a crucial role in mitigating many uncertainties. The statistics can be broadly segregated into descriptive and inferential statistics. Descriptive statistics helps to simplify and understand the data characteristics. It further facilitates the selection of appropriate inferential statistics. Therefore, the descriptive analysis is the most crucial step in the data analysis practices.

The guidelines such as Consolidated Standard of Reporting (CONSORT), Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), and Strengthening and Reporting of Observational Studies in Epidemiology (STROBE) have been developed and endorsed for statistical and overall reporting of medical research.[1–3] The researchers can refer to guidelines for more detail. The current article discusses the importance of the selection of appropriate descriptive statistics for reporting data. Tabular and graphical approaches are two popular approaches for descriptive data. Both the strategies are essential and supplement each other. However, considering the limitation of space and scope of the article, the current article primarily discusses the tabular approach. We broadly dissect the data into the categorical and continuous domains for discussion.

Categorical Data

Data can be broadly segmented into categorical and continuous data. Categorical data have a measurement scale that consists of a finite and distinct number of categories. They are also known as qualitative data and present characteristics that are not measured numerically. These categories are finite and countable. Categorical data can be further broadly segregated into nominal and ordinal data.

[1,2]Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India

Corresponding Author: Kamal Kishore, Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India, Phone: +91 9591349768, e-mail: kkishore.pgi@gmail.com

How to cite this article: Kishore K, Kapoor R. Statistics Corner: Reporting Descriptive Statistics. J Postgrad Med Edu Res 2020;54(2):66–68.

Source of support: Nil

Conflict of interest: None

Nominal Data

Nominal data have a measurement scale that consists of unordered categories. These categories are only for identification purpose. The data such as cancer type (lung cancer, stomach cancer, liver cancer, or mouth cancer) and blood groups ("O," "A," "B," or "AB") are examples of nominal data. Majority of computer software for the statistical analysis do not understand the text data. Therefore, numerals are assigned to nominal data, such as "O = 1," "A = 2," "B = 3," and "AB = 4," to get relevant information. It is essential to label and identify numerals for nominal data appropriately. Nominal data with two categories, such as presence or absence of disease and smoker or nonsmoker, are known as binary data. The outcome of interest, such as diseased or smoker, is usually assigned a value of 1. The frequency and percentages are preferred tabular display for nominal data. However, the categories with few observations are sometimes meaningfully merged to form a new group.

Example: Assume that there are people with blood groups O = 54 (54%), A = 30 (30%), B = 9 (9%), and AB = 7 (7%) present in the study. There are very few people with blood groups B and AB. Therefore, these categories can be merged and labeled as other = 16*. The table subscript can be used to tell readers that: * → Represent blood group from B (n = 10) and AB (n = 6). This rule can be followed for different categorical variables with low frequencies as per requirement.

A general rule of thumb is to merge categories with less than 10% data. However, investigators need to carefully find a balance between information loss and details to give as per study area,

distribution of data, objectives, and sample size.[4–6] The bar graph and pie chart are used to present a graphical display for nominal data. A pie chart is ideal for categorical data for a single group having five or fewer categories. The categories are proportional to the numerical assessment they represent. It is better to use a clustered bar chart for more than five categories and to compare two or more groups. The researchers should refrain from using three-dimensional (3D) graphs. It gives an ambiguous presentation to the data.

Ordinal Data

The ordinal data have a measurement scale that consists of ordered categories. Ordinal data have mutually exclusive classes, and the same can be arranged in the ascending or descending order. The data such as pain (no pain, mild pain, moderate pain, severe pain, and very severe pain) and Tumor grade (grade I, grade II, grade III, and grade IV) are examples of ordinal data. The numeral assignment to ordinal data, such as "no pain = 1," "mild pain = 2," "moderate pain = 3," "severe pain = 4," and "very severe pain = 5," tells that categories are distinct and are ordered in terms of increasing pain intensity. However, the distance between the two categories is relative and not an absolute measure of pain. Therefore, mean and standard deviation (SD) are not appropriate to report for ordinal data. It is better to use number and percentage for less than five categories and median and either or both interquartile and interdecile deviations for five or more categories.[7] Many investigators report ordinal data as a binary outcome by merging categories. An investigator should refrain from reporting ordinal data as binary due to loss of information. However, many times the categories in ordinal data need to be merged for clinical and practical reasons. The investigators should combine only adjacent categories that are combinable.

Example: Assume that there are people with no pain = 6 (6%), mild pain = 35 (35%), moderate pain = 30 (30%), severe pain = 20 (20%), and very severe pain = 9 (9%) in the study. No pain and severe pain categories cannot be combined despite less than 10% observations in each. Logically, no pain can be merged with mild pain and reported as mild or no pain (41%). Whereas severe and very severe pain can be combined and reported as severe or very severe pain (29%).

A thumb rule of less than 10% can also be applied to ordinal data if required. The bar graph and pie chart can also be used to present ordinal data with a limited number of categories. Ordinal data are frequently shown with clustered and stacked bar graphs to display the percentage of categories in many groups. The stacked chart is ideal for the equivalent sample size.

Continuous Data

The continuous data have a measurement scale that can be meaningfully expressed in numbers. Continuous data are also known as quantitative data. They can be broadly segmented into the discrete and continuous groups. Discrete data are countable and expressed in natural numbers. The data, such as the number of students in classrooms and people in the community, are examples of discrete data. Continuous data are data that can take any value between two data points. The data, such as height and weight, are examples of continuous data. The distribution of data (symmetric and asymmetric) determines the reporting of appropriate descriptive statistics.

Symmetric Data

The descriptive statistics for continuous data can be broadly segmented into measures of location (mean, median, and mode), dispersion (SD, range, interquartile, and interdecile deviation), and shape (skewness and kurtosis). The most common measures of reporting continuous data are mean and SD for both symmetric and asymmetric data. However, mean and SD are not an appropriate measure for skewed data. A thumb rule is to have a small SD. It is challenging to define small SD; thus, a generic ratio of ≥ 2 for mean to SD can be taken as a preliminary indication of small SD. Measures of shape (skewness and kurtosis) are vital and established criteria to determine skewness. Majority of the statistical software also calculates skewness and kurtosis. Measures of shape facilitate appropriateness of mean and SD as a descriptive measure. Despite this, researchers rarely compute, ascertain, and report measures of shape in the manuscripts. The values of skewness and kurtosis are 0 and 3, respectively, for symmetric data. The mean, variance, skewness, and kurtosis are based on the principle of moments and represent a first, second, third, and fourth moment, respectively.

Example: The weight of 10 individuals are measured as 40, 42, 53, 47, 54, 60, 51, 45, 59, and 44 kg, respectively. The mean and median weights are 49.5 and 49 kg, respectively. Whereas, SD, interquartile, and interdecile deviations are 6.98, 5.88, and 9.85 kg, respectively. The data are slightly asymmetric with the positive value of skewness = 0.19. However, mean (SD) is the best measure as skewness value is almost near to 0.

Continuous data are frequently presented with histograms and boxplots rather than qq-plots and pp-plots to assess data symmetry. Boxplots can be used to plot and compare data for more than one group. Initially, asymmetric data are transformed with logarithmic, inverse, and square root transformation to make them symmetric. Subsequently, mean, SD, and appropriate inferential statistics are reported on a converted symmetric scale.

Asymmetric Data

Continuous asymmetric data can be both positively (mean > median > mode) and negatively (mean < median < mode) skewed. A value for $|\text{skewness}| \neq 0$ indicates the degree and direction of nature of asymmetry. The negative and positive value signifies elongated left and right tail, respectively. The higher the absolute value, more is the skewness in the data. The value of kurtosis is <3 and >3 for curve flatter and taller than the bell-shaped curve, respectively. The biological data almost always display some degree of asymmetry. However, a gross violation from symmetry makes mean and SD redundant for asymmetric data. Similarly, many times transformation do not make asymmetric data symmetric. In these kinds of situations, it is better to report median, range, and either or both interquartile and interdecile deviations. Many investigators report mean and SD for nonnormal ($|\text{kurtosis}| \neq 3$) and skewed ($|\text{skewness}| \neq 0$) data, which is not the correct approach. A typical example of skewed data is patient hospital stay.

Example: The hospital stay for 10 patients are 3, 4, 5, 5, 6, 6, 7, 8, 9, and 20 days, respectively. The mean and median stay at the hospital are 7.3 and 6 days, respectively. Whereas, SD, interquartile, and interdecile deviations are 4.81, 1.75, and 7.9 days, respectively. The mean to SD ratio for the given data is <2. Further, data are asymmetric with the positive value of skewness = 1.91 and kurtosis = 5.52. Therefore, it is better to report the median, interquartile, and interdecile deviations rather than mean and SD.

Table 1: General guidelines to report descriptive measures for continuous and categorical data

Data	*Nominal*	*Ordinal*	*Continuous nonskew*	*Continuous skew*
Example	Cancer type: pancreatic cancer, neck cancer, mouth cancer, or lung cancer	Cancer grade: grade I, grade II, grade III, or grade IV	Height of the people coming to OPD	Number of days of patients in IPD
Central tendency*	Number	Number (<5 grade) Median (≥5 grade)	Mean	Median
Dispersion	Percentages	Percentages (<5) Interquartile (≥5)	SD	Range, interquartile, interdecile
Location	—	—	Skewness and kurtosis	Skewness and kurtosis
Transformation	Not recommended to transform nominal (>2) to binary data	Not recommended to transform ordinal to binary data	Not required	Logarithmic, square root, inverse, or other
Graphical approach	Bar and pie chart	Bar, pie, clustered, and stacked bar chart	Histogram, pp-plot boxplot, and qq-plot	Histogram and boxplot

*The measure of central tendency for ordinal data will depend on the number of categories of ordinal variable

Mean and SD are not the appropriate measure for this kind of data. It is better to report median and interdecile deviations along with other measures (such as min–max value, range, and interquartile deviation). Boxplot is a preferred method for graphical display of asymmetric data because it helps in the identification of both skewness and outliers in the data. Researchers can consult Table 1 for general guidelines to report descriptive data.

Conclusion

Many researchers report mean and SD for ordinal, continuous, symmetric, and asymmetric data. However, descriptive statistics describes the characteristics of data. It further lay the foundation of using the parametric or nonparametric test during inferential statistics. Therefore, researchers should calculate, evaluate, and carefully report appropriate descriptive measures without being biased. The investigators need to make informed decisions based on a sound statistical methodology rather than pre-mindset to report mean and SD.

Acknowledgments

We acknowledge Dr Meenakshi Sharma and Dr Rahul Mahajan for their valuable time and inputs to improve the quality of the article.

References

1. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg 2012;10(1):28–55. DOI: 10.1016/j.ijsu.2011.10.001.
2. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700. DOI: 10.1136/bmj.b2700.
3. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the reporting of observational studies in epidemiology (strobe): explanation and elaboration. PLOS Med 2007;4(10):e297. DOI: 10.1371/journal.pmed.0040297.
4. Bondell HD, Reich BJ. Simultaneous factor selection and collapsing levels in ANOVA. Biometrics 2009;65(1):169–177. DOI: 10.1111/j.1541-0420.2008.01061.x.
5. Gertheiss J, Tutz G. Sparse modeling of categorial explanatory variables. Ann Appl Stat 2010;4(4):2150–2180. DOI: 10.1214/10-AOAS355.
6. Petry S, Flexeder C, Tutz G, Pairwise fused lasso [Internet]. 2011. Available from: https://epub.ub.uni-muenchen.de/12164/1/petry_etal_TR102_2011.pdf.
7. LaValley MP, Felson DT. Statistical presentation and analysis of ordered categorical outcome data in rheumatology journals. Arthritis Care Res 2002;47(3):255–259. DOI: 10.1002/art.10453.