A simple guide to analyse data: Descriptive statistics in quantitative research

E Libhaber

School of Clinical Medicine, Health Sciences Research Office,
Faculty of Health Sciences, University of the Witwatersrand,
South Africa

Email:
elena.libhaber@wits.ac.za

E Libhaber https://orcid.org/0000-0002-7043-4002

DOI: https://doi.org/10.24170/22-3-7664

STATISTICAL ANALYSIS

Statistical data analysis can be divided into two big domains: descriptive statistics and inferential statistics. The process of statistical inference in drawing conclusions about the entire population is based on the information from a sample.[1] Notably, "the inferential statistics process fails if the sample is not representative of the population".[2] To achieve statistical inference, the first step of data analysis before hypothesis testing is descriptive statistics, which is directly linked to the research questions or the study objectives. This guide assumes that sampling and sample size were adequately considered before data analysis. The various steps involved in the analysis of descriptive data are discussed below.

EARLY STAGE: COLLECTING DATA

REDCap (Research Electronic Data Capture) is a tool used to build and manage online surveys and databases.[3] This software is widely applied in clinical research scenarios and basic sciences, reducing data entry errors. The data captured with REDCap can be exported directly to Excel or any statistical package (SPSS, STATA, SAS, R, etc.). For simple data entry and quick data management, an Excel spreadsheet serves its purpose, and it can perform primary statistical analyses. Data saved in Excel can also be exported to any statistical software (Statistica, SPSS, STATA, SAS, R, GraphPad, etc.).

All variables for each participant should be stored in a single, merged dataset to ensure compatibility in analysis. A frequent hurdle in inferential statistics arises when demographic and clinical data are captured in different formats and/or placed on different sheets. Consequently, the demographic data cannot be analysed together with the rest of the data. An example is a comparison between male and female patients with idiopathic cardiomyopathy, or predicting ischaemia in patients with severe hypertension by adjusting for age and sex, which cannot be performed unless the clinical and demographic data are merged.

It is advised not to use a statistical programme to capture data initially. The decision of which statistical software to use depends on the nature of the analysis.

DETECTING MISSING DATA

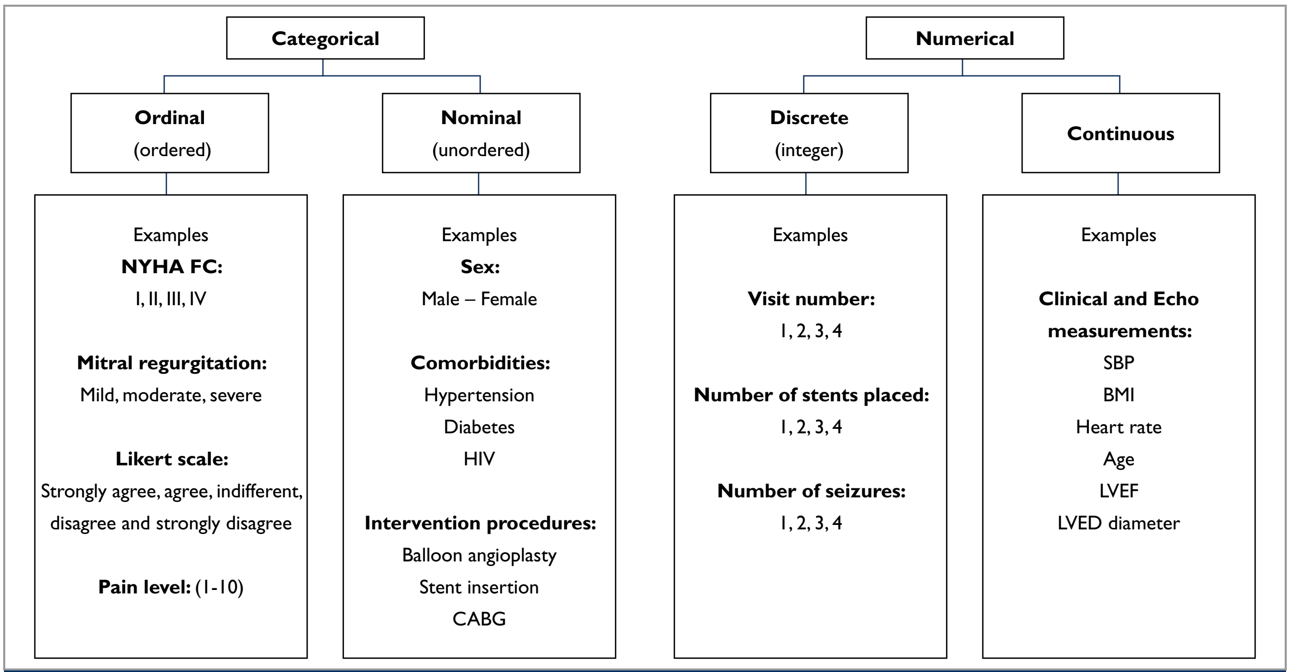
Missing data is a frequent problem in medical research. Excluding individuals with missing observations can yield substantial bias in the results and reduce statistical power. Different methods have been introduced to handle missing data. The most recommended method is multiple imputation (MI).[4] This method is already included in standard statistical programmes, such as STATA, SPSS, and R. However, the primary approach to identifying missing data is to run a descriptive analysis.

All outputs of descriptive statistics, regardless of the statistical software used, will show the number of observations per variable (n). Thus, a first reading of the results should be to check the size of the study sample (N) and n per group if more than one group is included. The purpose is to verify if the numbers concur with the initial sample size calculations, stated as such in the methods section of the protocol.

An easy way to detect missing data is to examine the number of observations; whether key variables, such age and sex, have consistent numbers of observations, which may reveal missing data patterns, e.g. age (n = 135), sex (n = 135), heart rate (HR) (n = 132), body mass index (BMI) (n = 123), and ejection fraction (EF) (n = 132). Subsequently, the identification of patients who exhibit missing observations may assist with the retrieval of the omitted data. If variables such as sex and age are missing, those participants will have to be excluded from the analysis, unless MI techniques are applied.

REPORTING DATA

In the description of the summarised data in the results, variables must be classified according to the scale in which they were measured: nominal (categorical and ordinal) or numerical (discrete [integer], continuous) (Figure 1).[5] Categorical variables are presented as frequencies (n) and percentages (%). Continuous data following an approximately normal distribution (symmetric) are summarised as the mean and standard deviation (SD) (e.g. age (SD) 29(6)), otherwise median and interquartile range (IQR) (75%–25% quartile) (e.g. BMI 26 [17–52]).[6]



Categorical

Ordinal (ordered)

Examples
NYHA FC: I, II, III, IV
Mitral regurgitation: Mild, moderate, severe
Likert scale: Strongly agree, agree, indifferent, disagree and strongly disagree
Pain level: (1-10)

Nominal (unordered)

Examples
Sex: Male – Female
Comorbidities: Hypertension, Diabetes, HIV
Intervention procedures: Balloon angioplasty, Stent insertion, CABG

Numerical

Discrete (integer)

Examples
Visit number: 1, 2, 3, 4
Number of stents placed: 1, 2, 3, 4
Number of seizures: 1, 2, 3, 4

Continuous

Examples
Clinical and Echo measurements: SBP, BMI, Heart rate, Age, LVEF, LVED diameter

Figure 1: Variable types according to the scale of measurement
Modified and adapted from " Petrie and Sabin, Medical biostatistics at glanc"[5]

It is essential to depict data graphically to better guide one with the inferential statistics (e.g. by detecting skewness to the right or left, outliers, etc.). Graphic methods, such as histograms, box plots, normality plots, and/or a statistical test (e.g. the Shapiro–Wilk test), can be used to assess normality. The conventional summary statistics of reporting discrete data (e.g. number of stents) are frequency and percentages, or median and IQR (e.g. parity 2 [1–4]).[6]

PRESENTING DATA IN A TABLE

Descriptive tables are usually sufficient to summarise sample characteristics; there is no need to supplement the information with a pie chart or bar graph. There are cases where a figure highlighting a particular variable, such as "multiple comorbidities", could better visually summarise data than a table.

Herewith, some recommendations to draw up a comprehensive table (examples provided in Tables Ia and Ib):

- Identify the type of variables (categorical or numerical) to select the summary statistics adequately.
- Ensure that the information in the table supports the objectives of the study.
- Always display the "N" for the total sample and the "n" for each group (if more than one group) in the title of the table, on the first row, or the second row.
- Always write the units of the variables.
- Always clarify if continuous variables are expressed as means (SD) or median (IQR) next to the variables' name or in the footnotes.
- Do not display percentages without frequencies if the sample/group size is not shown in the table, especially when the sample size is small or varies between groups.
- Ensure that percentages in columns add up to 100%; otherwise, include a sentence in the footnotes of the table (e.g. "Column percentages do not add to 100% as patients can have more than one condition.").
- Write out abbreviations with the full wording. If the name of a variable is too long and there is not enough space for full words, list the abbreviations in the footnote.
- If the study design is cross-sectional and includes more than one group of patients p-values of comparisons (t-tests or chi-square tests) between the characteristics can be added as an extra column.
- If the study design is longitudinal (e.g. EF measurements at baseline, 6 months, and 1 year), the table should include only baseline data (see Table Ia).
- Symbols indicating the significant differences could be placed next to the results and the p-value in the footnotes (see Table Ib).
- It is not advisable to collapse categorical data before inferential analysis.
- Never duplicate results in figures and tables; it is redundant.

Printing out the table will help to visualise the data and guide the researcher for subsequent analyses. A thorough descriptive analysis will lead to the selection of appropriate statistical tests, such as an independent or paired t-test, chi-square test, ANOVA, Mann–Whitney U test, and Kruskal–Wallis in inferential statistics.

TABLE Ia: Baseline clinical and echocardiographic characteristics.*

Variables	Total (*n* = 120)
Age (years)	38.7 ± 12.8
Sex, female, n (%)	60 (50)
Body mass index (kg/m^2)	27.9 ± 5.8
Systolic blood pressure (mmHg)	122 ± 17
Diastolic blood pressure (mmHg)	77 ± 10
Heart rate (bpm)	77.2 ± 12.6
Ejection fraction (%)	62.5 ± 8.1
LV mass index (g/m^2)	61.1 ± 18.0
Left atrial volumes	
Max-LAVi (ml/m^2)	19.7 ± 5.9
Min-LAVi i (ml/m^2)	7.7 ± 3.2

Data reported as mean ± standard deviation, unless otherwise noted.

LV: left ventricular, max-LAVi: maximum left atrial volume index, min-LAVi: minimum left atrial volume index.

** Table extracted partially and adapted from Meel, et al.*[7]

TABLE Ib: Clinical and demographic characteristics of hypertensive patients with normal and low ejection fraction.*

Variable	HTNEF group	HTLEF group
Number of patients	41	41
Age (years)	55.5 ± 8.4	55.1 ± 9.0
Women/men (n)	22/19	22/19
Body mass index (kg/m^2)	30.2 ± 4.9	28.9 ± 4.5
NYHA functional capacity		
I	26 (63%)	12 (29%)**
II	15 (37%)	14 (34%)
III		15 (37%)
Duration of hypertension (years)	15.6 ± 8.4	12.2 ± 6.4
Duration of heart failure (years)	1.8 ± 1.0	3.3 ± 1.6**
Systolic blood pressure (mmHg)	141 ± 14	156 ± 8
Diastolic blood pressure (mmHg)	84 ± 12	89 ± 11
Heart rate (bpm)	75 ± 12	81 ± 10**
Medications		
Furosemide	41 (100%)	41 (100%)
ACE inhibitors or ARBs	41 (100%)	41 (100%)
β-blockers	28 (68%)	41 (100%)

Data are expressed as mean ± standard deviation or number (percentage).

HTNEF: hypertensive with normal ejection fraction, HTLEF: hypertensive with low ejection fraction, ACE: angiotensin-converting enzyme, ARB: angiotensin receptor blocker, NYHA: New York Heart Association.

** Table extracted partially and adapted from Maharaj, et al.*[8]

*** p-value < 0.05.*

REFERENCES

1. Pagano M, Gauvreau K. Principles of biostatistics. New York: Cengage Learning; 2000. p. 196.
2. Altman DG. Practical statistics for medical research. 2nd ed. London: Chapman and Hall/CRC; 1999. p. 490.
3. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019;95:103208. https://doi.org/10.1016/j.jbi.2019.103208.
4. Heymans MW, Twisk JWR. Handling missing data in clinical research. J Clin Epidemiol 2022;151:185-188. https://doi.org/10.1016/j.jclinepi.2022.08.016.
5. Petrie A, Sabin C. Medical statistics at a glance. Wiley: London; 2009. p. 14.
6. Azibani F, Pfeffer TJ, Ricke-Hoch M, et al. Outcome in German and South African peripartum cardiomyopathy cohorts associates with medical therapy and fibrosis markers. ESC Heart Fail 2020;7(2):512-522. https://doi.org/10.1002/ehf2.12553.
7. Meel R, Khandheria BK, Peters F, et al. Left atrial volume and strain parameters using echocardiography in a black population. Eur Heart J Cardiovasc Imaging 2017;18(3):350-355. https://doi.org/10.1093/ehjci/jew062.
8. Maharaj N, Khandheria BK, Libhaber E, et al. Relationship between left ventricular twist and circulating biomarkers of collagen turnover in hypertensive patients with heart failure. J Am Soc Echocardiogr 2014;27(10):1064-1071. https://doi.org/10.1016/j.echo.2014.05.005.