OPEN ACCESS

Ology Press

Conceptual Paper

Biostatistics and Epidemiology International Journal

# Processing of data and analysis

Dr Balkishan Sharma

Sri Aurobindo Medical College & PG Institute, India

**Correspondence:** Dr Balkishan Sharma, PhD, Associate Professor (Biostatistics), Department of Community Medicine, Sri Aurobindo Medical College & PG Institute, Indore (MP), India, Email bksnew@rediffmail.com and bksnew@gmail.com

**Received:** February 06, 2018 | **Published:** February 20, 2018



## Introduction

Health research is essential in developing evidence-based interventions that will make a difference in mitigating health problems, promoting health and ultimately improving the quality of life. Data collected and compiled from experimental work, records and surveys should be accurate and complete. They must be checked for accuracy and adequacy before processing further. So far they lie in masses, are scattered in the records and in other words they are mixed and unsorted.

The processing of data and further analysis may be break up into three stages: (1) data management, (2) explanatory data analysis and (3) statistical analysis (testing and modeling). However, before proceeding for statistical analysis, researcher must consider the following raised points:

- ✓ Understand the process involved in data processing.
- ✓ Use computers to perform data processing.
- ✓ Distinguish between qualitative and quantitative data.
- ✓ Understand probabilities and their applications.
- ✓ Interpret summary statistics, graphical presentation and contingency tables.
- ✓ Commonly presented in the health literature.
- ✓ Carry out exploratory data analysis.
- ✓ Understand the process involved in estimations and hypothesis testing.
- ✓ Interpret the functions of confidence intervals and p-values.
- ✓ Understand statements in published articles relating to statistics.
- ✓ Use computers to perform some statistical analysis.

Exploratory data analysis involves examination of the data errors and describing data using summary statistics and graphical techniques (descriptive statistics). As data errors are often detected at this stage, the process of exploratory data analysis and data cleaning are typically iterative. The aim of this process is, for the researcher to gain familiarity with, and understanding of the data, in order to determine the approach to take, and methods to use in further statistical analyzes.

## Data-processing techniques

Once the fieldwork has been completed, the information that has been gathered must be centralized and input to the computer. Data entry is tiresome work, but it isnecessary to accord it some attention because of inevitable errors in reading and entering the information, especially if it is not carried out by people accustomed to using a keyboard for data-entry. The computerized data can be stored in tabular form in a spreadsheet (e.g. Microsoft Excel, Lotus 123) or more compactly and efficiently in a relational database management system (e.g. Access, Oracle, dBase).

## Methods of statistical processing

Statistics offers a range of methods, the choice of which will depend on four factors: (1) The type of variables: qualitative or quantitative; (2) The status of the variables: explanatory or dependent; (3) The number of variables: one, two or multiple and (4) the type of analysis: exploratory (descriptive) or confirmatory (inferential).

## Scope and purpose

Data analysis is the process of developing answers to questions through the examination and interpretation of data. The basic steps in the analytic process consist of identifying issues, determining the availability of suitable data, deciding on which methods are appropriate for answering the questions of interest, applying the methods and evaluating, summarizing and communicating the results.

Analytical results underscore the usefulness of data sources by shedding light on relevant issues. Data analysis also plays a key role in data quality assessment by pointing to data quality problems in a given survey. Analysis can thus influence future improvements to the survey process.

## Quality indicators

Main quality elements: relevance, interpretability, accuracy, accessibility. An analytical product is relevant if there is an audience who is (or will be) interested in the results of the study. For the interpretability of an analytical article to be high, the style of writing must suit the intended audience. Sufficient and relevant details must



**Citation:** Sharma B. Processing of data and analysis. *Biostatistics Epidemiol Int J.* (2018);1(1): 3-5.
DOI: 10.30881/beij.00003

be provided so that another person if allowed access to the data could replicate the results.

For an analytical product to be accurate, appropriate methods and tools need to be used to produce the results. For an analytical product to be accessible, it must be available to people for whom the research results would be useful.

## Analysis of data

Data analysis converts data into information and knowledge, and explores the relationship between variables. Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. According to Shamoo & Resnik[1] various analytic procedures "provide a way of drawing inductive inferences from data and distinguishing the signal (the phenomenon of interest) from the noise (statistical fluctuations) present in the data."

Understanding of the data analysis procedures will enable you to appreciate the meaning of the scientific method which includes testing of hypotheses and statistical significance in relation to research questions. There are a number of issues that researchers should be cognizant of with respect to data analysis. Some of the key considerations in analysis and selection of the right test of significance are as follows:

- ✓ Having the necessary skills to analyze.
- ✓ Distinguishing data types.
- ✓ Distinguishing different types of Statistical tests.
- ✓ Identify the selection of a right test.
- ✓ Determining statistical significance.
- ✓ Distinguishing between Parametric and Non-Parametric test with their applying criteria.
- ✓ Distinguishing between Correlation and Regression.
- ✓ Drawing unbiased inference.
- ✓ Inappropriate subgroup analysis.
- ✓ Lack of clearly defined and objective outcome measurements.
- ✓ Partitioning 'text' when analyzing qualitative data.
- ✓ Reliability and Validity.
- ✓ Extent of analysis.

Table 1 consists of common statistical tests. All tests which are described below are provided in the book titled *Intuitive Biostatistics* by Harvey Motulsky[2] and were performed by InStat, except for tests marked with asterisks. Tests labeled with a single asterisk are briefly mentioned in this book, and tests labeled with two asterisks were not mentioned at all.

**Table 1** Common statistical tests

| | ***Type of Data*** | | | |
|---|---|---|---|---|
| ***Goal*** | **Measurement (from Gaussian Population)** | **Rank, Score, or Measurement (from Non-Gaussian Population)** | **Binomial (Two Possible Outcomes)** | **Survival Time** |
| **Describe one group** | Mean, SD | Median, Interquartile range | Proportion | Kaplan Meier survival curve |
| **Compare one group to a hypothetical value** | One-sample *t* test | Wilcoxon test | Chi-square or Binomial test** | - |
| **Compare two unpaired groups** | Unpaired *t* test | Mann-Whitney test | Fisher's test (chi-square for large samples) | Log-rank test or Mantel-Haenszel* |
| **Compare two paired groups** | Paired *t* test | Wilcoxon test | McNemar's test | Conditional proportional hazards regression* |
| **Compare three or more unmatched groups** | One-way ANOVA | Kruskal-Wallis test | Chi-square test | Cox proportional hazard regression** |
| **Compare three or more matched groups** | Repeated-measures ANOVA | Friedman test | Cochrane Q** | Conditional proportional hazards regression** |
| ***Goal*** | **Measurement (from Gaussian Population)** | **Rank, Score, or Measurement (from Non-Gaussian Population)** | **Binomial (Two Possible Outcomes)** | **Survival Time** |


**Citation:** Sharma B. Processing of data and analysis. *Biostatistics Epidemiol Int J.* (2018);1(1): 3-5. DOI: 10.30881/beij.00003

(Table 1 continuous..)

| | | | | |
|---|---|---|---|---|
| **Quantify association between two variables** | Pearson correlation | Spearman Correlation | Contingency Coefficients** | - |
| **Predict value from another measured variable** | Simple LR or Non LR | Nonparametric Regression** | Simple Logistic Regression* | Cox proportional hazard regression* |
| **Predict value from several measured or binomial variables** | Multiple LR* or Multiple Non LR** | - | Multiple Logistic Regression* | Cox proportional hazard regression* |

*briefly mentioned in *Intuitive Biostatistics*[2]

**not mentioned in *Intuitive Biostatistics*[2]

Developments in the field of statistical data analysis often parallel or follow advancements in other fields to which statistical methods are fruitfully applied. Data is known to be crude information and not knowledge by itself. The sequence from data to knowledge is: *from Data to Information, from Information to Facts, and finally, from Facts to Verification of Truth*. Data becomes information, when it becomes relevant to your research problem. Information becomes fact, when the data can support it. Facts are what the data reveals. However, the decisive instrumental (i.e., applied) knowledge is expressed together with some statistical degree of confidence.

Lastly, I do hope that this in-depth knowledge will create an awareness to promote the use of statistical thinking and techniques to apply them to make educated decisions whenever there is variation in data.

## References and further reading


1. Shamoo AE, Resnik DB. Responsible Conduct of Research. Third Edition. Oxford, New York: Oxford University Press. 2015;360p.
2. Motulsky H. Chapter 37: Choosing a test. In: Intuitive Biostatistics. *Oxford University Press Inc*. 1995.
3. Bland M. An Introduction to Medical Statistics. Fourth Edition. Oxford, New York: Oxford University Press. 2015;448p.
4. Daniel W. Biostatistics: A Foundation for Analysis in the Health Sciences. 4th ed. New York: Wiley. 1987.
5. http://www.emacpd.org/sites/default/files/resource_center/3.Data%20Processing%20Module.pdf
6. Korn EL, Graubard BI. Analysis of health surveys. *John Wiley & Sons*. 2011;323.
7. Lehtonen R, Pahkinen E. Practical methods for design and analysis of complex surveys. *John Wiley & Sons*. 2004.
8. Pagano M, Gauvreau K. Principles of Biostatistics. 2nd ed. *Duxbury Press*. 1990.
9. Sharma B. Right choice of a method for determination of cut-off values: A statistical tool for a diagnostic test. 2014.
10. Shamoo AE. Principles of research data audit. *Taylor & Francis*. 1989.
11. Sharma K. Chapter X. In: Nursing research and statistics. Haryana, India: Elsevier Health Sciences. 2011.
12. Shephard RJ. Ethics in exercise science research. Sports Medicine. 2002;32(3):169−183.
13. Silverman S, Manson M. Research on teaching in physical education doctoral dissertations: a detailed investigation of focus, method, and analysis. *Journal of Teaching in Physical Education*. 2003;22(3):280−297.
14. Smeeton N, Goda D. Conducting and Presenting Social Work Research: Some Basic Statistical Considerations. *The British Journal of Social Work*. 2003;33(4):567−573.
15. Thompson M. Theory of Sample Surveys. London: Chapman & Hall. 1997.