Review Article

Statistical methods in clinical studies: An overview

K. V. S. Sarma,[1] Alladi Mohan,[1] Sai Sarada Vedururu[2]

[1]Department of Medicine, Sri Venkateswara Institute of Medical Sciences, [2]Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh, India

Abstract

Basic knowledge of statistics is essential for the proper design of clinical studies, data handling methods, appropriate use of analytical tools and the interpretation of the findings. Recognising the importance of the need for competence in biostatistics and research methodology, the Medical Council of India has made it mandatory for the postgraduate medical student to learn research methodology by taking up an online course; and has also made it necessary for faculty in medical colleges to complete the online basic course in biomedical research to get promoted This review focusses on providing an overview regarding various statistical tools commonly used in the design of studies and analysis of data.

Keywords: Confidence intervals, data, multiple variable analysis, parametric tests, patterns, *P* value

Address for correspondence: Dr K. V. S. Sarma, Statistics Consultant, Department of Medicine, Sri Venkateswara Institute of Medical Sciences, Tirupati 517 507, Andhra Pradesh, India.
E-mail: smskvs@gmail.com

Submitted: 16-Aug-2020 **Accepted:** 04-Sep-2020 **Published:** 22-Oct-2021

INTRODUCTION

The importance of statistical data in clinical research needs no emphasis in these days of data-driven clinical health-care decisions. The common man is flooded with numbers, charts and reviews (all statistical data) to describe the health dynamics of the community. Statistical data, as seen today, are not mere numbers but includes text, images and voice (acoustics) or a mixture of all, which reflects an observed fact. In scientific research, particularly biomedical research, one has to deal with data arising under uncertainties attributable to biological variability, sampling fluctuations, changes in the environmental conditions and so on. The effect of these variations has to be properly handled in drawing conclusions. Since most research is based on samples rather than a census of the cohort, the sample size, the method of drawing sample and the analytical methods used foranalysis will impact the findings.

Recognising the importance of the need for competence in biostatistics and research methodology, the Medical Council of India has made it mandatory for postgraduate medical student to learn research methodology by taking up an online course;[1] and has also made it necessary for faculty in medical colleges to complete the online basic course in biomedical research to get promoted.[2]

The results from a sample study serve as estimates for the target group but prone to carry errors which can be evaluated and controlled. The researcher has to report all sources of variability (confounders) and account for them in summarising the findings or making comparisons. It is the science of statistics that plays a key role in the conduct of a clinical study.

Statistics-related errors in biomedical research have been a concern among the research community for quite some

Access this article online

Quick Response Code:

Website: www.jcsr.co.in

DOI: 10.4103/JCSR.JCSR_69_20

How to cite this article: Sarma KVS, Mohan A, Vedururu SS. Statistical methods in clinical studies: An overview. J Clin Sci Res 2022;11:34-9.

time. The pitfalls in statistical reporting have been well explained by various authors and standards are proposed for researchers as well all reviewers of journals. An extensive survey of literature on statistical reporting and guidelines known as the Statistical Analyses and Methods in Published Literature (SAMPL) has been published[3] that suggested under broad categories, some principles for reporting statistical methods. Reporting issues in biomedical research have been discussed and an improved version of SAMPL with a finer resolution of topics has been proposed.[4] It was also observed that well-designed biostatistics courses are essential for medical researchers and help prevent deficient statistical reporting.[2] In another communication,[5] it was observed that failure to check the validity conditions before using tests based on normality of data; and misinterpretation of *P* values are two major pitfalls in handling statistical data. They have categorised the tests as parametric and non-parametric and outlined different tests for categorical and continuous data.

CLARITY IN RESEARCH PROTOCOL

The study protocol shall be very clear, stating the purpose of the study, the main and sub objectives, need for the study, type of the study (descriptive/comparative), time frame (cross-sectional/longitudinal) and the methodology (both domain specific and statistical). It should also specify the chief outcomes to be observed from the study and the list of variables (categorical and continuous). In descriptive studies, the chief interest will be to report the status of study participants under stated health conditions or interventions. Comparison of interventions is not the objective for this type of study. However, within the sample, comparison of outcomes among factors such as age group, gender or socio-economic status can be made.

In comparative studies, the protocol shall clearly mention the interventions or methodologies (like imaging) and whether the objective is to establish the superiority of one over the other. This needs statistical tests of hypotheses. The methodology part shall mention the list of tools to be used, the level of significance, the desired power of the test, the effect size (if applicable) and the *P* value to consider the statistical significance.

The most important aspect is the sample size (n) required for the study. The magic number $n = 30$ frequently used in clinical studies is not always correct. Statistically speaking, the size n depends on the desired precision in the estimation of outcomes in descriptive studies or on the desired power of testing in comparative studies.

Details regarding formulas for sample size have been published.[6,7] Formulae are also available to find the sample size for various purposes such as correlation analysis, predictive models and design of biomarkers.[8] A systematic and practical approach for sample size estimation is available online.[9] Further, excellent Indian resources for sample size calculations include "nMaster 2.0" and "Dx Diagnostic Test software" developed by the Biostatistics Training and Resource Center, Department of Biostatistics, Christian Medical College, Vellore, Tamil Nadu (www.cmc-biostatistics.ac.in).

When the objective is to present the aetiology of patients under observation, a practice is to include all consecutive eligible subjects reporting to the particular clinic during the study period. However, if it is a comparative study one should use statistically designed size.

When randomisation of interventions to subjects is required, it should be specified in the protocol and the method of randomisation has to be mentioned. The mere mentioning of the 'lottery method' is not correct. When blinding is required, the method should be stated along with block size. Online blinding and randomisation tools are also available (https://www.randomizer. org/or http://mahmoodsaghaei.tripod.com/Softwares/randalloc.html).

A QUICK GUIDE FOR DATA PRESENTATION

The data on each characteristic in the study are a variable because it changes from record to record. Those entities which are held constant (fixed-like gender in women health studies) need not be a part of the datasheet. Characteristics represented as observed facts such as sex, presence/absence of a condition and level of pain are categorical and attaching numerical codes to categories is a good option when data is manually entered into spreadsheets. For such variables, we can only count the cases of interest, convert them into proportions or percentages to make a summary. Such data are called qualitative data and the variable is called a categorical variable.

When the characteristic of interest is a measurement like blood pressure or glucose level, it is measured as a continuous variable (like the blood glucose). Such variables are reported using statistical summaries like mean (average) or median, along with a measure of variability known as standard deviation (SD) or the Inter-Quartile Range (IQR).

Occasionally, we use measures known as percentiles to determine ad-hoc standards for clinical parameters.

The first step in the analysis is to visualise the data with suitable charts. As a rule, continuous data are presented by histogram and categorical data by the bar chart. If the histogram shows nearly bell shape (symmetric around the mean), the data are approximately normally distributed (should also have skewness close to 0 and kurtosis close to 3). A more rigorous way of cheking normality is by using the Kolmogorov–Smirnov test. Categorical data are usually skewed (unsymmetrical) but shows bell shape in large samples. The presence of outliers (extreme values) is one reason for the lack of symmetry and it often leads to higher SD than the mean.

A quick summary of methods for presenting statistical data is shown in Table 1.

Table 1: Guidelines for primary understanding on data

Data type	Data pattern (distribution)/ presentation	Summary
Categorical	Any pattern (skewed or symmetric)	Count, percentage, minimum, maximum
Continuous	Normal (to be verified by Kolmogorov–Smirnov test; *P*>0.05 for normality)	Mean, SD, 95% CIs for normal data Median and IQR for nonnormal data
Categorical or continuous	Any pattern (skewed or symmetric) Regrouping into non-overlapping categories required	Count, percentage, minimum, maximum Two-way cross tabulation to study Associations
Bivariate categorical	Each variable shall have few categories only (not to use continuous variables)	Body of the table contains count or percent or both
Categorical or continuous	Bar/pie for categorical data Line chart for time series Histogram for continuous data Box plot to compare distributions	Charts with caption/ title Data labels on the chart Display vital results on the chart 3D charts to display depth

CI=Confidence interval, IQR=Inter-quartile range, SD: Standard deviation, 3D=Three dimensional

DATA COMPARISON AMONG INDEPENDENT GROUPS (UNPAIRED COMPARISON)

Comparison of data summaries within the sample among various categories of the confounding factors (like sex or baseline conditions) is necessary to confirm that significant differences do not occur among the groups within the sample. This is done by testing for the statistical significance of the difference in summaries due to groups at some level of significance, say α =0.05 or confidence level 95%. The test result is indicated by *P* value and reported as significant when $P < 0.05$.

The American Statistics Association[10] stated in 2016 that 'Scientific conclusions and business or policy decision should not be based on whether the *P* value passes a specified threshold.'. The importance of reporting the confidence interval or effect size in addition to reporting the exact *P* value has been presented.[11]

Statistical tests are conducted for comparing two or more independent proportions, two or more means, tests for the correlation coefficient, regression coefficients, sensitivity of a procedure, testing the level of agreement between methods, etc., Every test has its own *P* value and needs to mention about the test used in the context.

Many statistical tests assume that the summary value under comparison (like mean or proportion) follows normal distribution, with large sample size and the *P* value is based on the Z-statistic of the standard normal distribution. These tests are known as parametric tests. The Student's *t*-test and analysis of variance (ANOVA) are small sample tests for comparing means. With large samples, the *t*-test and the Z-test produce the same result. For comparing two independent proportions, the Chi-square test is used when the sample is small and Z-test can be used with large samples. An outline of tools used for comparing data across groups and among different samples is shown in Table 2.

COMPARISON OF PAIRED AND REPEATED MEASURES DATA

In the case of continuous variables, satisfying normality, when data are observed from the same subject of the sample before and after a condition, we get a paired comparison. Such data exhibits a correlation between the values before and after the condition, and hence, independent sample *t*-test for means cannot be used. The tool used in this context is the paired *t*-test. For similar conditions with categorical data the McNemar's test is used.

Sometimes, observations are repeatedly collected from the patient at specific time points. The anaesthesiologist takes measurements such as mean arterial pressure and heart rate at baseline and three or more times at equal intervals. It is like the extension of a paired comparison situation, but all the observations on the same patient form a multivariate data and the variation within subject (patient) is to be addressed. This is done by applying repeated measures

Table 2: Guide lines for comparison of data sets

Variable/nature of data	Data pattern/ distribution	Number of groups	Test(s) applicable	Values to report
Categorical	Dichotomous	Two	*Z*-test for proportions Chi-square test/Fisher's exact test	Observed difference *P-value*
Count data as contingency table	Dichotomous/ polychotomous	Two	Chi-square Fisher's exact test	*P-value* Measure of association
Continuous	Normal	Two	Student's *t*-test for means (unpaired)	Mean, standard deviation for each group Mean difference Effect size *P-value*
Continuous or ranked data	Not normal	Two	Mann–Whitney U-test	Mean, standard deviation for both groups Mean difference Effect size *P-value*
Continuous	Normal	Three or more	One factor ANOVA Multiple comparison test	*F*-value, *P* value of ANOVA Means plot (optional) *P*-values of multiple comparisons
Continuous	Not normal	Three or more	Kruskal–Wallistest Multiple comparison test	*Z*-value, *P* value Means plot (optional) *P*-values of multiple comparisons
Continuous with two or more factors	Normal	Two or more for each factor	Multifactor ANOVA	*F*-value and *P* value for each factor
Continuous with two or more factors	Not normal	Two or more for each factor	Transform data to normality. Statistician's assistance recommended	

ANOVA=Analysis of variance

ANOVA. A systematic approach for this tool has been described.[12] A summary of the tools and reporting style is presented in Table 3.

CONFIDENCE INTERVALS

The sample mean is basically a summary value claimed as a point estimate (single value) for the target group, while the variation within the sample is given by the SDs. The precision of this estimate is presented as an interval which holds the true value of the population with some confidence, usually 95%. Such an interval is called the confidence interval (CI) and conveys more information than the point estimate. For normally distributed data, the 95% CI is given by mean ± 1.96 × SE where SE = s/ sqrt (n) denotes the standard error. For instance, the mean HbA1c is 5.8, with 95% CI as (5.1, 6.8) says that the true average will be between 5.1 and 6.8 with 95% chance. This principle is used in determining the sample size for estimating a parameter in descriptive studies. CIs are also used while presenting other metrics like odds ratio, regression coefficients, sensitivity, specificity, etc.

In the recent two decades, there is growing awareness on the use of CI not only to report the precision of the estimate but also to support the findings when the *P* value of the result is significant. The width of the interval measures the precision of the result[13,14] and smaller width indicates higher precision. We can use CI to understand the significance of a result in a way different from using *P* value. Suppose H_0: Mean of Group A = mean of Group B denotes the null hypothesis, according to which the difference in means of two independent groups is 0. If the result is significant at $P = 0.05$, then the CI, in general, should not contain zero.[14]

CORRELATION AND REGRESSION

The strength of the relationship between two continuous variables Y and X is measured using Pearson's correlation

Table 3: Guidelines for analysing data with repeated measures

Variable/nature of data	Data pattern (distribution)	Number of groups	Test	Values to report
Continuous	Normal	Two	Paired *t*-test	Mean and standard deviation Mean difference *P*-value
Continuous with repeated observations	Normal	Three or more	Repeated measures ANOVA	*P*-value for duration *P*-value for groups
Continuous	Not normal	Two/three or more	Transform data to normality. Statistician's assistance recommended	

ANOVA=Analysis of variance

coefficient denoted by '*r*' with values between −1 and +1. It is to be calculated only if the two variables have a meaningful relationship; else leads to a phenomenon known as non-sense correlation,[15,16] like the correlation between the body mass index and the monthly income of persons.

The correlation coefficient is meaningful only when the scatter plot of (X, Y) shows an approximately linear trend (upwards or downwards). A value $r = 0.80$ indicates a stronger relationship than $r = 0.15$. When the scatter diagram does not show any specific trend/pattern the value of r trends to be very low. When the data are in the form of ranked pairs as in the case of interpreting a medical image by independent experts, we use another measure called Spearman's correlation (ρ). A guide to the proper use of the correlation coefficient in medical research has been published.[17] Specific applications of correlation coefficient in thoracic surgery have also been described.[18]

Correlation coefficient is used for continuous variables only. In the case of categorical variables, the relationship is measured in terms of strength of association calculated using Chi-square value and several such measures are available. Table 4 shows a summary methodology for reporting relationships.

MULTIVARIABLE ANALYSIS

When each variable is independently studied, the analysis is known as univariate analysis in contrast with the study of several variables together, known as multivariable analysis or multivariate analysis. The latter approach is important while comparing the profiles of patients in terms of inter-correlated parameters like the lipid profile. Such a study conveys more information than the univariate approach because the dependencies are accounted for, using correlation or association measures. It is important to note that many real-life problems are multivariate in nature involving all types of variables. MANOVA and repeated measures analysis are also multivariate methods and one has to use for example for example statistical software like the Statistical Package for the Social Sciences (SPSS) (IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM CorpSYSTAT) and R for the analysis (https://www.r-project.org/about.html).[13] Multiple linear regression, logistic regression with multiple explanatory variables, multidimensional receiver operator characteristic curves are also applications of multivariate analysis and such methods accommodate both categorical and continuous variables. Variables which are found significant in univariate analysis are included in multivariate analysis.

Image processing in radiology deals with software-supported multivariate methods[19,20] to assess a health condition or for disease prediction.

We conclude with the observation that clearly written protocol, unbiased selection of subjects into the study, choosing the proper tool for the analysis are three vital components in clinical research to arrive at meaningful conclusions. Data analytics is the key to the success of a clinical study, which is possible with the software-supported data analytical tools.

Financial support and sponsorship

Nil.

Table 4: Methodology for reporting relationship measures

Measure/tool	Data pattern	Purpose	Test(s)	Values to report
Pearson's correlation coefficient (*r*)	Continuous data Normal	Strength of relationship	*t*-test	*r*- value with CI *P*-value of *r* Scatter diagram (optional)
Spearman's correlation coefficient (ρ)	Ranked/ordinal data	Strength of relationship	*Z*-test	ρ value with CI *P*-value of ρ Scatter diagram (optional)
Several correlations	Normal	Understanding several correlations at a time	*t*-test	Correlation matrix *P*-values Matrix scatter plot
Regression	One continuous response variable Several independent predictors	Prediction of average outcome	*t*-test *F*-test	Model R^2 and *P* value Regression coefficients Final model
Logistic regression	Binary outcome and one or more independent predictors	Prediction of outcome	Several tests	Model R^2 *P*-value Regression coefficients Odds ratio with CI Final model

CI=Confidence intervals

Conflicts of interest

Alladi Mohan is an Editor of Journal of Clinical and Scientific Research. The article was subject to the journal's standard procedures, with peer review handled independently of this faculty and their research groups.

REFERENCES

1. Board of Governors in Supersession of Medical Council of India. Notification. New Delhi, the 11th December, 2019. No. MCI-18(1)/2019-Med./171700.
2. Board of Governors in Supersession of Medical Council Of India. Notification New Delhi, the 12th February, 2020 No. MCI-12(2)/2019-Med. Misc./189334.
3. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: The "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. Int J Nurs Stud 2015;52:5-9.
4. Indrayan A. Basic statistical reporting for articles published in biomedical journals: The "Statistical Analyses and Methods in the Published Literature" or the SAMPL guidelines. Int J Nurs Stud 2015;52:5-9.
5. Ahmed S, Dhooria A. Pitfalls in statistical analysis– A reviewers' perspective. Indian J Rheumatol 2020;15:39-45.
6. Wayne DW, Chad CL. Biostatistics: Basic Concepts and Methodology for the Health Sciences. 10th ed. Delhi, India: Wiley Series in Probability and Statistics; 2014.
7. Indrayan A, Malhotra RK. Medical Biostatistics. 4th ed. Boca Raton: CRC Press; 2018.
8. Sarma KV. Sample Size Determination in Clinical Research– Perceptions and Practice. Special Proceeding of the 21st Annual Conference of Society for Statistics and Computer Applications held at Sri Venkateswara Agricultural College (Acharya N. G. Ranga Agricultural University), Tirupati, January 29-31, 2019. p. 93-9.
9. Berkowitz J. Sample Size Estimation. Available from: http://www.columbia.edu/~mvp19/RMC/6_SampleSize.htm. [Last accessed on 2020 Aug 03].
10. Wasserstein RL, Lazar NA. The ASA statement on *P* values: Context, process and purpose. Am Stat 2016;70:129-33.
11. Indrayan A. Attack on statistical significance: A balanced approach for medical research. Indian J Med Res 2020;151:275-8.
12. Sarma KV, Vishnu Vardhan R. Multivariate Statistics Made Simple – A Practical Approach. Boca Raton: CRC Press; 2019.
13. Flechner L, Tseng TY. Understanding results: *P* values, confidence intervals, and number need to treat. Indian J Urol 2011;27:532-5.
14. Schober P, Vetter TR. Confidence intervals in clinical research. Anesth Analg 2020;130:1303.
15. Yule GY. Why do we sometimes get nonsense correlations between time series? a study in sampling and the nature of time series. J R Statist Soc 1926;89:1-63.
16. Aldrich J. Correlations genuine and spurious in Pearson and Yule. Stat Sci 1995;10:364-76.
17. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J 2012;24:69-71.
18. Gust L, D'journo XB. The use of correlation functions in thoracic surgery research. J Thorac Dis 2015;7:E11-5.
19. Jang YH, Park S, Park YU, Kwack KS, Jeon SW, Lee HY. Multivariate analyses of MRI findings for predicting osteomyelitis of the foot in diabetic patients. Acta Radiol 2020;61:1205-12.
20. Lacson R, Odigie E, Wang A, Kapoor N, Shinagare A, Boland G, *et al*. Multivariate analysis of radiologists' usage of phrases that convey diagnostic certainty. Acad Radiol 2019;26:1229-34.