Selection of Appropriate Statistical Methods for Data Analysis

Abstract

In biostatistics, for each of the specific situation, statistical methods are available for analysis and interpretation of the data. To select the appropriate statistical method, one need to know the assumption and conditions of the statistical methods, so that proper statistical method can be selected for data analysis. Two main statistical methods are used in data analysis: descriptive statistics, which summarizes data using indexes such as mean and median and another is inferential statistics, which draw conclusions from data using statistical tests such as student's t-test. Selection of appropriate statistical method depends on the following three things: Aim and objective of the study, Type and distribution of the data used, and Nature of the observations (paired/unpaired). All type of statistical methods that are used to compare the means are called parametric while statistical methods used to compare other than means (ex-median/mean ranks/proportions) are called nonparametric methods. In the present article, we have discussed the parametric and non-parametric methods, their assumptions, and how to select appropriate statistical methods for analysis and interpretation of the biomedical data.

Keywords: *Diagnostic accuracy, parametric and nonparametric methods, regression analysis, statistical method, survival analysis*

Prabhaker Mishra, Chandra Mani Pandey, Uttam Singh, Amit Keshri[1], Mayilvaganan Sabaretnam[2]

Departments of Biostatistics and Health Informatics, [1]Neuro-otology and [2]Endocrine Surgery, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, Uttar Pradesh, India

Introduction

Selection of appropriate statistical method is very important step in analysis of biomedical data. A wrong selection of the statistical method not only creates some serious problem during the interpretation of the findings but also affects the conclusion of the study. In statistics, for each specific situation, statistical methods are available to analysis and interpretation of the data. To select the appropriate statistical method, one need to know the assumption and conditions of the statistical methods, so that proper statistical method can be selected for data analysis.[1] Other than knowledge of the statistical methods, another very important aspect is nature and type of the data collected and objective of the study because as per objective, corresponding statistical methods are selected which are suitable on given data. Practice of wrong or inappropriate statistical method is a common phenomenon in the published articles in biomedical research. Incorrect statistical methods can be seen in many conditions like use of unpaired t-test on paired data or use of parametric test for

the data which does not follow the normal distribution, etc., At present, many statistical software like SPSS, R, Stata, and SAS are available and using these softwares, one can easily perform the statistical analysis but selection of appropriate statistical test is still a difficult task for the biomedical researchers especially those with nonstatistical background.[2] Two main statistical methods are used in data analysis: descriptive statistics, which summarizes data using indexes such as mean, median, standard deviation and another is inferential statistics, which draws conclusions from data using statistical tests such as student's t-test, ANOVA test, etc.[3]

Factors Influencing Selection of Statistical Methods

Selection of appropriate statistical method depends on the following three things: Aim and objective of the study, Type and distribution of the data used, and Nature of the observations (paired/unpaired).

Aim and objective of the study

Selection of statistical test depends upon our aim and objective of the study. Suppose our objective is to find out the predictors of the outcome variable, then regression

Address for correspondence:
Dr. Prabhaker Mishra,
Department of Biostatistics and Health Informatics, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, Uttar Pradesh, India.
E-mail: mishrapk79@gmail.com

analysis is used while to compare the means between two independent samples, unpaired samples t-test is used.

Type and distribution of the data used

For the same objective, selection of the statistical test is varying as per data types. For the nominal, ordinal, discrete data, we use nonparametric methods while for continuous data, parametric methods as well as nonparametric methods are used.[4] For example, in the regression analysis, when our outcome variable is categorical, logistic regression while for the continuous variable, linear regression model is used. The choice of the most appropriate representative measure for continuous variable is dependent on how the values are distributed. If continuous variable follows normal distribution, mean is the representative measure while for non-normal data, median is considered as the most appropriate representative measure of the data set. Similarly in the categorical data, proportion (percentage) while for the ranking/ordinal data, mean ranks are our representative measure. In the inferential statistics, hypothesis is constructed using these measures and further in the hypothesis testing, these measures are used to compare between/among the groups to calculate significance level. Suppose we want to compare the diastolic blood pressure (DBP) between three age groups (years) (<30, 30--50, >50). If our DBP variable is normally distributed, mean value is our representative measure and null hypothesis stated that mean DBP values of the three age groups are statistically equal. In case of non-normal DBP variable, median value is our representative measure and null hypothesis stated that distribution of the DBP values among three age groups are statistically equal. In above example, one-way ANOVA test is used to compare the means when DBP follows normal distribution while Kruskal--Wallis H tests/median tests are used to compare the distribution of DBP among three age groups when DBP follows non-normal distribution. Similarly, suppose we want to compare the mean arterial pressure (MAP) between treatment and control groups, if our MAP variable follows normal distribution, independent samples t-test while in case follow non-normal distribution, Mann--Whitney U test are used to compare the MAP between the treatment and control groups.

Observations are paired or unpaired

Another important point in selection of the statistical test is to assess whether data is paired (same subjects are measures at different time points or using different methods) or unpaired (each group have different subject). For example, to compare the means between two groups, when data is paired, paired samples t-test while for unpaired (independent) data, independent samples t-test is used.

Concept of Parametric and Nonparametric Methods

Inferential statistical methods fall into two possible categorizations: parametric and nonparametric. All type of statistical methods those are used to compare the means are called parametric while statistical methods used to compare other than means (ex-median/mean ranks/proportions) are called nonparametric methods. Parametric tests rely on the assumption that the variable is continuous and follow approximate normally distributed. When data is continuous with non-normal distribution or any other types of data other than continuous variable, nonparametric methods are used. Fortunately, the most frequently used parametric methods have nonparametric counterparts. This can be useful when the assumptions of a parametric test are violated and we can choose the nonparametric alternative as a backup analysis.[3]

Selection between Parametric and Nonparametric Methods

All type of the *t*-test, *F* test are considered parametric test. Student's *t*-test (one sample *t*-test, independent samples *t*-test, paired samples *t*-test) is used to compare the means between two groups while *F* test (one-way ANOVA, repeated measures ANOVA, etc.) which is the extension of the student's *t*-test are used to compare the means among three or more groups. Similarly, Pearson correlation coefficient, linear regression is also considered parametric methods, is used to calculate using mean and standard deviation of the data. For above parametric methods, counterpart nonparametric methods are also available. For example, Mann--Whitney U test and Wilcoxon test are used for student's *t*-test while Kruskal--Wallis H test, median test, and Friedman test are alternative methods of the *F* test (ANOVA). Similarly, Spearman rank correlation coefficient and log linear regression are used as nonparametric method of the Pearson correlation and linear regression, respectively. [3,5-8] Parametric and their counterpart nonparametric methods are given in Table 1.

Statistical Methods to Compare the Proportions

The statistical methods used to compare the proportions are considered nonparametric methods and these methods have no alternative parametric methods. Pearson Chi-square test and Fisher exact test is used to compare the proportions between two or more independent groups. To test the change in proportions between two paired groups, McNemar test is used while Cochran Q test is used for the same objective among three or more paired groups. Z test for proportions is used to compare the proportions between two groups for independent as well as dependent groups.[6-8] [Table 2].

Other Statistical Methods

Intraclass correlation coefficient is calculated when both pre-post data are in continuous scale. Unweighted and weighted Kappa statistics are used to test the absolute agreement between two methods measured on the same subjects (pre-post) for nominal and ordinal data, respectively. There

Table 1: Parametric and their Alternative Nonparametric Methods

Description	Parametric Methods	Nonparametric Methods
Descriptive statistics	Mean, Standard deviation	Median, Interquartile range
Sample with population (or hypothetical value)	One sample *t*-test (n <30) and One sample *Z*-test (n ≥30)	One sample Wilcoxon signed rank test
Two unpaired groups	Independent samples *t*-test (Unpaired samples *t*-test)	Mann Whitney U test/Wilcoxon rank sum test
Two paired groups	Paired samples *t*-test	Related samples Wilcoxon signed-rank test
Three or more unpaired groups	One-way ANOVA	Kruskal-Wallis H test
Three or more paired groups	Repeated measures ANOVA	Friedman Test
Degree of linear relationship between two variables	Pearson's correlation coefficient	Spearman rank correlation coefficient
Predict one outcome variable by at least one independent variable	Linear regression model	Nonlinear regression model/Log linear regression model on log normal data

Table 2: Statistical Methods to Compare the Proportions

Description	Statistical Methods	Data Type
Test the association between two categorical variables (Independent groups)	Pearson Chi-square test/Fisher exact test	Variable has ≥2 categories
Test the change in proportions between 2/3 groups (paired groups)	McNemar test/Cochrane Q test	Variable has 2 categories
Comparisons between proportions	Z test for proportions	Variable has 2 categories

are some methods those are either semiparametric or nonparametric and these methods, counterpart parametric methods, are not available. Methods are logistic regression analysis, survival analysis, and receiver operating characteristics curve.[9] Logistic regression analysis is used to predict the categorical outcome variable using independent variable(s). Survival analysis is used to calculate the survival time/survival probability, comparison of the survival time between the groups (Kaplan--Meier method) as well as to identify the predictors of the survival time of the subjects/patients (Cox regression analysis). Receiver operating characteristics (ROC) curve is used to calculate area under curve (AUC) and cutoff values for given continuous variable with corresponding diagnostic accuracy using categorical outcome variable. Diagnostic accuracy of the test method is calculated as compared with another method (usually as compared with gold standard method). Sensitivity (proportion of the detected disease cases from the actual disease cases), specificity (proportion of the detected non-disease subjects from the actual non-disease subjects), overall accuracy (proportion of agreement between test and gold standard methods to correctly detect the disease and non-disease subjects) are the key measures used to assess the diagnostic accuracy of the test method. Other measures like false negative rate (1-sensitivity), false-positive rate (1-specificity), likelihood ratio positive (sensitivity/false-positive rate), likelihood ratio negative (false-negative rate/Specificity), positive predictive value (proportion of correctly detected disease cases by the test variable out of total detected disease cases by the itself), and negative predictive value (proportion of correctly detected non-disease subjects by test variable out of total non-disease subjects detected by the itself) are also used to calculate the diagnostic accuracy of the test method.[3,6,10] [Table 3].

Advantage and Disadvantages of Nonparametric Methods over Parametric Methods and Sample Size Issues

Parametric methods are stronger test to detect the difference between the groups as compared with its counterpart nonparametric methods, although due to some strict assumptions, including normality of the data and sample size, we cannot use parametric test in every situation and resultant its alternative nonparametric methods are used. As mean is used to compare parametric method, which is severally affected by the outliers while in nonparametric method, median/mean rank is our representative measures which do not affect from the outliers.[11]

In parametric methods like student's t-test and ANOVA test, significance level is calculated using mean and standard deviation, and to calculate standard deviation in each group, at least two observations are required. If every group did not have at least two observations, its alternative nonparametric method to be selected works through comparisons of the mean ranks of the data.

For small sample size (average ≤15 observations per group), normality testing methods are less sensitive about non-normality and there is chance to detect normality despite having non-normal data. It is recommended that when sample size is small, only on highly normally distributed data, parametric method should be used otherwise corresponding nonparametric methods should be preferred. Similarly on sufficient or large sample size (average >15 observations per group), most of the statistical methods are highly sensitive about non-normality and there is chance to wrongly detect non-normality, despite having normal data. It is recommended that when sample size is

Table 3: Semi-parametric and non-parametric methods

Description	Statistical methods	Data type
To predict the outcome variable using independent variables	Binary Logistic regression analysis	Outcome variable (two categories), Independent variable (s): Categorical (≥ 2 categories) or Continuous variables or both
To predict the outcome variable using independent variables	Multinomial Logistic regression analysis	Outcome variable (≥ 3 categories), Independent variable (s): Categorical (≥ 2 categories) or continuous variables or both
Area under Curve and cutoff values in the continuous variable	Receiver operating characteristics (ROC) curve	Outcome variable (two categories), Test variable : Continuous
To predict the survival probability of the subjects for the given equal intervals	Life table analysis	Outcome variable (two categories), Follow-up time : Continuous variable
To compare the survival time in ≥ 2 groups with P	Kaplan--Meier curve	Outcome variable (two categories), Follow-up time : Continuous variable, One categorical group variable
To assess the predictors those influencing the survival probability	Cox regression analysis	Outcome variable (two categories), Follow-up time : Continuous variable, Independent variable(s): Categorical variable(s) (≥ 2 categories) or continuous variable(s) or both
To predict the diagnostic accuracy of the test variable as compared to gold standard method	Diagnostic accuracy (Sensitivity, Specificity etc.)	Both variables (gold standard method and test method) should be categorical (2×2 table)
Absolute Agreement between two diagnostic methods	Unweighted and weighted Kappa statistics/Intra class correlation	Between two Nominal variables (unweighted Kappa), Two Ordinal variables (Weighted kappa), Two Continuous variables (Intraclass correlation)

sufficient, only on highly non-normal data, nonparametric method should be used otherwise corresponding parametric methods should be preferred.[12]

Minimum Sample Size Required for Statistical Methods

To detect the significant difference between the means/medians/mean ranks/proportions, at minimum level of confidence (usually 95%) and power of the test (usually 80%), how many individuals/subjects (sample size) are required depends on the detected effect size. The effect size and corresponding required sample size are inversely proportional to each other, that is, on the same level of confidence and power of the test, when effect size is increasing, required sample size is decreasing. Summary is, no minimum or maximum sample size is fix for any particular statistical method and it is subject to estimate based on the given inputs including effect size, level of confidence, power of the study, etc., Only on the sufficient sample size, we can detect the difference significantly. In case lack of the sample size than actual required, our study will be under power to detect the given difference as well as result would be statistically insignificant.

Impact of Wrong Selection of the Statistical Methods

As for each and every situation, there are specific statistical methods. Failing to select appropriate statistical method, our significance level as well as their conclusion is affected.[13] For example in a study, systolic blood pressure (mean $\pm$ SD) of the control (126.45 ± 8.85, $n_1 = 20$) and treatment (121.85 ± 5.96, $n_2 = 20$) group was compared using Independent samples t-test (correct practice). Result showed that mean difference between two groups was statistically insignificant ($P = 0.061$) while on the same data, paired samples t-test (incorrect practice) indicated that mean difference was statistically significant ($P = 0.011$). Due to incorrect practice, we detected the statistically significant difference between the groups although actually difference did not exist.

Conclusions

Selection of the appropriate statistical methods is very important for the quality research. It is important that a researcher knows the basic concepts of the statistical methods used to conduct research study that produce a valid and reliable results. There are various statistical methods that can be used in different situations. Each test makes particular assumptions about the data. These assumptions should be taken into consideration when deciding which the most appropriate test is. Wrong or inappropriate use of statistical methods may lead to defective conclusions, finally would harm the evidence-based practices. Hence, an adequate knowledge of statistics and the appropriate use of statistical tests are important for improving and producing quality biomedical research. However, it is extremely difficult for a biomedical researchers or academician to learn the entire statistical methods. Therefore, at least basic knowledge is very important so that appropriate selection of the statistical methods can decide as well as correct/incorrect practices can be recognized in the published research. There are many softwares available online as well as offline for analyzing the data, although it is fact that which set of statistical tests are appropriate for the given data and study objective is still very difficult for the researchers to understand. Therefore, since planning of the study to data collection, analysis and finally in the review process, proper consultation from statistical experts may be an alternative option and can reduce the burden from the clinicians to go in depth of statistics which required lots of time and effort and ultimately affect their

clinical works. These practices not only ensure the correct and appropriate use of the biostatistical methods in the research but also ensure the highest quality of statistical reporting in the research and journals.[14]

Conflicts of interest

There are no conflicts of interest.

References

1. Nayak BK, Hazra A. How to choose the right statistical test?. Indian J Ophthalmol 2011;59:85-6.
2. Karan J. How to select appropriate statistical test?. J Pharm Negative Results 2010;1:61-3.
3. Mishra P, Mayilvaganan S, Agarwal A. Statistical methods in endocrine surgery journal club. World J Endoc Surg 2015;7:21-3.
4. Mishra P, Pandey CM, Singh U, Gupta A. Scales of measurement and presentation of statistical data. Ann Card Anaesth 2018;21:419-22.
5. Campbell MJ, Swinscow TDV. Statistics at Square One 11th ed. Wiley-Blackwell: BMJ Books; 2009.
6. Sundaram KR, Dwivedi SN, Sreenivas V. Medical Statistics: Principles And Methods. Anshan; 2010.
7. Altman DG. Practical Statistics For Medical Research. CRC Press; 1990.
8. Barton B, Peat J. Medical Statistics: A Guide to SPSS, Data Analysis and Clinical Appraisal. 2nd ed. Wiley Blackwell, BMJ Books; 2014.
9. Peat J, Barton B. Medical Statistics: A Guide to Data Analysis and Critical Appraisal. John Wiley & Sons; 2008.
10. Armitage P, Berry G, Matthews JNS. Statistical Methods In Medical Research. John Wiley & Sons; 2008.
11. Kim HY. Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. Open lecture on statistics. Restor Dent Endod 2013;38:52-4.
12. Ghasemi A, Zahediasl S. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. Int J Endocrinol Metab 2012;10:486-9.
13. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research: A review of common pitfalls. Swiss Med Wkly 2007;137:44-9.
14. Bajwa SJ. Basics, common errors and essentials of statistical tools and techniques in anesthesiology research. J Anesthesiol Clin Pharmacol 2015;31:547-53.