Basic Statistical Tools in Research and Data Analysis

Gopal Prasad Sedhai*

ABSTRACT

Statistical methods involved in carrying out the study consists of planning, designing, data collection, analyze the data, drawing suitable interpretation and reporting the findings. For each research work, statistical methods are used for the analysis and interpretation of the data. In the selection of a proper statistical method for data analysis, the researcher has a concrete idea about the assumption and conditions of the statistical methods. The results and inferences are precise only if proper statistical tests are used. This article will try to update the reader with the basic research tools that are utilized while conducting various studies. The two most well-known statistical methods use in data analysis are descriptive statistics and inferential statistics. Descriptive statistics summarize and organize the characteristics of a data set using various measures such as mean, median, variance, etc. while inferential statistics help to come to conclusions from sample data using statistical tests such as t-test, F -test, etc. The selection of a proper statistical method depends on the objective of the research study, type of data, type of distribution and nature of the observations. Methods that use distributional assumptions are called parametric methods because we estimate the parameters of the distribution assumed for the data. The parametric statistical methods are used to compare the means while methods that do not require us to make distributional assumptions about the data, such as the rank methods, are called non-parametric. Nonparametric methods are used to compare other than means like proportions, medians, etc. This article covers a brief outline of parametric and nonparametric methods, the assumption in these methods and the selection of appropriate statistical tests for analysis and interpretation of the survey data.

***Keywords**: Inferential statistics, parametric tests, non-parametric methods, regression analysis, hypothesis*

INTRODUCTION

Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability or dispersion. Measures of central tendency include

**Lecturer, Department of Statistics, Prithvi Narayan Campus, Pokhara*

Email: gopal.sedhai@prnc.tu.edu.np

ISSN: 2392-4519 (Print)

the mean, median and mode [1] while measures of variability include the standard deviation or variance, the minimum and maximum values of the variables, kurtosis and skewness. [2]

Statistical inference is the process of using data analysis to infer properties of an underlying probability distribution. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.[3]

A researcher should have good knowledge to select the appropriate statistical method because the result of the wrong selection of the statical method carries serious problems in the implementation of the finding. Various statistical methods are available for a specific situation and condition to analyze the data. The assumptions and conditions of different statistical methods are different. So, in a selection of statistical methods for data analysis, good knowledge of the assumptions and conditions is essential and the proper statical method can be selected in data analysis[4]. Likewise, the type and nature of the data and objective of the research also play a very important role in the data analysis procedure. Hence, a particular statistical method is used for a particular objective. Nowadays, various statistical software such as SPSS, SAS, Stata, R, etc. is available in data processing and analysis. This software easily performs statistical analysis but using this software is very difficult for a person with a nonstatistical background. Two main statistical methods are used in data analysis called descriptive statistics and inferential statistics. A descriptive statistic is a summary statistic that quantitatively describes or summarizes features from a collection of information with main indexes mean and variance [5]. while in inferential statistics, you take data from samples and make generalizations about a population by using various statistical tests such as t-test (paired and unpaired), F- test, etc. [6]

Factors Influencing Selection of Statistical Methods

The selection of an appropriate statistical method depends on the following things: Aim and objective of the study, the Type and distribution of the data used, the Nature of the observations, distribution of data and paired/unpaired [7].

1. *Objective of the study (Type of analysis)*

The aim and objective of the study influence the statistical test. There are three main purposes of study comparison, the relation between two variables and prediction.

A comparison may be done by comparing the mean when the data are numerical continuous that follow a normal distribution, comparison of median if the data are numerical that do not follow normal distribution and comparison of proportion can be done in two variables of the same group or two variables of different groups [8].

Relation between two variables also known as correlation which measures the relationship between two variables of the same group such as the relation between body weight and blood pressure while in prediction (regression), the value of one variable is predicted or estimated based on another variable. For example, suppose we are interested in evaluating the effectiveness of a company training program, we have to measure the performance of a sample of employees before and after completing the program, and analyze the differences using a paired sample t-test.

2. *Type of variables*

There are two types of variables one is numerical and another is categorical. Numerical variables are of two types discrete and continuous. Numerical data have a numeric value. The numeric- discrete variable does not take any fractional values such as the number of students in a class, the number of accidents in a particular road, etc. but the numeric -continuous variable takes the fractional values such as blood pressure, body weight, etc. Categorical data has no numerical values. Numerical continuous data usually summaries the mean but numerical discrete data does not follow the normal distribution and usually summaries the median and categorical data summaries the proportion.

3. *Number of groups and data sets*

There may be one, two, or more than two groups. One group may have two data sets or more than two data sets. The one group with two datasets has one intervention data and one post-intervention data. For example, measuring the effect of a drug in a group of patients before and after the drug. A group with more than two data sets has one intervention data and more than two post-intervention data. For example, the effect of a drug on a group of patients measures in five different time intervals after the drug. Two groups with two data sets have two intervention data and two post-intervention data. For example, the effect of a drug on two different groups of patients with high blood pressure after and before the drug is given. Same way, there may be three different groups of patients using three types of drugs given and measure the blood pressure to identify the more effective drug.

4. *Paired and unpaired data*

The selection of proper statistical tests also depends on whether the data is paired or unpaired. In unpaired design, there are two or more than two data sets that are different from each other and have no related or called independent. An independent sample t-test is used to test the means of two unpaired data [10]. The paired design consists of one group with two data sets related to each other. For example, preformation of employers before and after conducting the training program. Paired sample t-test is used to test the paired data.

5. *Distribution of data*

Another important factor that affects the proper selection of statistical tests is whether the data follows normal distribution or not [11]. There may be three types of distribution classified in statistical tests known as normal, non-normal and dichotomous distributions. Data of body weight, blood pressure, sugar level, etc. are followed a normal distribution. Rank Data does not follow the normal distribution while data having two alternatives (yes/ no answered questions) is dichotomous distribution. If a continuous variable follows a normal distribution, the mean is the representative measure while for nonnormal data, the 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.

Parametric and Non-parametric Tests

A population is defined as a set of all the observations of a study or experiment. To understand the nature of the population, some quantitative measurements of the population are needed that is defined as population parameter. Thus, the population parameter is a characteristic of a population by which the natures of a population can be estimated. It is mainly used in inferential statistics. The value of the population parameter is not known in advance. It is usually estimated using the sample statistic values obtained from the sample).

Parametric tests are those that make assumptions about the parameters of the population distribution from which the sample is drawn. This is often the assumption that the population data are normally distributed. Non-parametric tests are "distribution-free" and, as such, can be used for non-Normal variables [9].

All types 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. Parametric tests are generally used on the assumption that the variable is normally distributed and continuous. The non- perimetric test is used when data is continuous with the nonnormal distribution. It may be used when a noncontinuous variable with any type of distribution 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.

Table 1

Parametric vs Nonparametric Test

	Parametric	Nonparametric
Assumed Distribution	Normal	Nonnormal
Typical Data	Ratio or Interval (Mostly for dependent variable)	Nominal and ordinal (Independent variable and dependent variable)
Usual Central Measures	Mean	Median
Benefits	Can draw any conclusions	Simplicity, less affected by outliers
Type of Tests	T-test (independent sample or paired sample), ANOVA, etc.	Mann Whitney U test, Chi-square test, etc.
Tests		
Independent measures, 2 groups	Independent measure t-test	Mann – Whitney test
Independent measures, > 2 groups	One-way independent measures of ANOVA	Kruskal Wallis test
Repeated measures, 2 conditions	Match paired t-test	Wilcoxon test

Selectin of Parametric and Nonparametric Test

All types of t-tests, F- tests are considered parametric tests. 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 the F test (oneway 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 [9]. Similarly, the Pearson correlation coefficient, linear regression is also considered parametric methods, is used to calculate using the mean and standard deviation of the data. For the above parametric methods, counterpart nonparametric methods are also available. For example, MannWhitney U test and the Wilcoxon test are used for student's t-test while the 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 a nonparametric method of the Pearson correlation and linear regression, respectively.[9] Parametric and their counterpart nonparametric methods are given in Table 2.

Table 2

Parametric and their Alternative Nonparametric Methods

	Comparison				Association (Correlation between two Variables)	Regression (Prediction of from another)
	2 data sets		> 2 data sets			
	Paired	Unpaired	Paired	unpaired		
Normal Distribution (mean)	Paired t-test	Unpaired t-test	Repeated measures ANOVA	One way ANOVA	Pearson's correlation	Linear Regression
Non- Normal (Median)	Wilson Sped Rank	Wilson Rank Sum test (U- test)	Friedman test	Kruskal Wallis H	Spearman's rank correlation	Non-parametric Regression
Dichotomous Data (Proportion)	McNew's test	Chi-Square test	Cochran's Q test	Chi-square test	Contingency Coefficient	Logistic Regression

Related Pairs of Parametric and Nonparametric Tests

Nonparametric tests are a shadow world of parametric tests. In the table below, linked pairs of statistical hypothesis tests are given

Table 3

Related Pairs of Parametric and Nonparametric Tests

Parametric tests of means	Nonparametric tests of medians
One - sample t-test	1-sample Sign, 1-sample Wilcoxon
Two - sample t-test	Mann-Whitney test
One -Way ANOVA	Kruskal-Wallis, Mood's median test
Factorial DOE with a factor and a blocking variable.	Friedman test

Advantages of Parametric Tests

Advantage 1: Parametric tests can provide trustworthy results with distributions that are skewed and nonnormal. Many people aren't aware of this fact, but parametric analyses can produce reliable results even when the data are continuous and non - normally distributed. We just have to be sure that the sample size meets the requirements for each analysis in the table below.

Table 4

Advantages of Parametric Test

Parametric analyses	Sample size requirements for non-normal data
1-sample t-test	Greater than 20
2-sample t-test	Each group should have more than 15 observations
One-Way ANOVA	For 2-9 groups, each group should have more than 15 observations For 10-12 groups, each group should have more than 20 observations

The parametric tests can be used with nonnormally distributed data.

Advantage 2: Parametric tests can provide trustworthy results when the groups have different amounts of variability. Indeed, nonparametric tests don't require data that are normally distributed. However, nonparametric tests have the disadvantage of an additional requirement that can be very hard to satisfy. The groups in a nonparametric analysis typically must all have the same variability (dispersion). Nonparametric analyses might not provide accurate results when variability differs between groups.

Conversely, parametric analyses, like the 2-sample t-test or one-way ANOVA, allow to an analysis of groups with unequal variances.

Advantage 3: Parametric tests have greater statistical power. In most cases, parametric tests have more power. If an effect exists, a parametric analysis is more likely to detect it.

Advantages of Nonparametric Tests

Advantage 1: Nonparametric tests can be used to analyze the data of all scales (ordinal data, ranked data, and outliers).

Advantage 2: Nonparametric tests are valid if the sample size is small and the data do not follow the normal distribution.

Advantage 3: Nonparametric tests are easier to compute and have fewer assumptions.

Advantage 4: Nonparametric tests assess the median rather than mean which can be better for some study areas.

Minimum Sample Size Required for Statistical Methods

The minimum sample size is essential to test the significant difference between the means, medians, ranks, proportions, etc. at a minimum level of confidence (usually 95%). In case of lack of the sample size that is required, our study cannot detect the given difference and the result would be statistically insignificant. We can detect the difference significantly only on the sufficient sample size [7].

Impact of Wrong Selection of the Statistical Methods

A wrong selection of the statistical method not only creates some serious problem during the interpretation of the findings but also affect the conclusion of the study. There are specific statistical methods for every situation. Failing to select an appropriate statistical method, our significance level as well as their conclusion is affected.[10] Due to incorrect practice, we detected a statistically significant difference between the groups although actually difference did not exist.

CONCLUSIONS

The selection of the appropriate statistical methods is very important for quality research. A researcher must know the basic concepts of the statistical methods used to conduct research that produces reliable and valid results. In different situations, different statistical methods are used and 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 and carries in the evidence-based practices. So, an adequate knowledge of statistics and the appropriate use of statistical tests are important for improving and producing quality research. However, it is extremely difficult for academicians to learn the entire statistical method. Hence, 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. Many types of software are 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 the 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.[14]

Conflict of Interest

There are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcomes.

REFERENCES

Hippel, Paul T. von (2005). *"Mean, Median, and Skew: Correcting a Textbook Rule". Journal of Statistics Education.* ***13*** *(2). doi:10.1080/10691898.2005.11910556*

Duncan Cramer (1997) Fundamental Statistics for Social Research Routledge. ISBN 9780415172042 (p 85)

Upton, G., Cook, I. (2008) Oxford Dictionary of Statistics, OUP. ISBN 978-0-19-954145-4. https://en.wikipedia.org/wiki/Statistical_inference#cite_note-Oxford-1

Nayak BK, Hazra A. How to choose the right statistical test?. Indian J Ophthalmol 2011;59:856.

Mann, Prem S. (1995). Introductory Statistics (2nd ed.). Wiley. ISBN 0-471-31009-3

Mishra P, Mayilvaganan S, Agarwal A. Statistical methods in endocrine surgery journal club. World J Endoc Surg 2015;7:213.

Mishra P, Pandey CM, Singh U, Gupta A. Scales of measurement and presentation of statistical data. Ann Card Anaesth 2018;21:419-22.

Kendall, M.G.; Stuart, A. (1969) *The Advanced Theory of Statistics, Volume 1: Distribution Theory, 3rd Edition*, Griffin. ISBN 0-85264-141-9 (Ex 12.9)

David J. Sheskin (27 August 2003). *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition*. CRC Press. pp. 7–. ISBN 978-1-4200-3626-8. Retrieved 25 February 2013.

Mishra P, Pandey CM, Singh U, Keshri A, Sabaretnam M. Selection of appropriate statistical ethods for data analysis. Ann Card Anaesth 2019;22:297-301. https://statisticsbyjim.com/glossary/factors/ Investopedia, Descriptive Statistics Terms

Bajwa SJ. Basics, common errors and essentials of statistical tools and techniques in anesthesiology research. J Anaesthesiol Clin Pharmacol. 2015 Oct-Dec;31(4):547-53. doi: 10.4103/0970-9185.169087. PMID: 26702217; PMCID: PMC4676249.

https://statisticsbyjim.com/hypothesis-testing/