International Journal of Research Studies in Biosciences (IJRSB)
Volume 5, Issue 5, May 2017, PP 33-37
ISSN 2349-0357 (Print) & ISSN 2349-0365 (Online)
http://dx.doi.org/10.20431/2349-0365.0505005
www.arcjournals.org

How to Choose the Statistical Technique in the Data Analysis

Hari Prasad Upadhyay

Lecturer of Biostatics, Department of Community Medicine, College of Medical Science, Bharatpur-10, Chitwan

Abstract: *The main objective of writing this paper is to provide the algorithm of choosing the statistical tools in the data analysis. Choosing an appropriate test one of the most important task in research. So, that right test will give the valid conclusion and wrong test may give the misleading inference. In order to choose the right statistical test we should be familiar with the different variable and their nature (for parametric, test of normality)*

Keywords: *Statistical too, Parametric, Test of Normality*

1. INTRODUCTION

Nowadays statistics is the one of the most important part in all sectors. Like in the medical sectors it plays the vital role for the inference. With out of the knowledge on the tools and technique of statistics nobody can write the quantitative research paper. There are the various rules in the statistics for the data analysis. Knowingly or unknowingly there is no proper use of the statistics in data analysis in the research writing for the. Most of the researchers use the wrong statistics for the significant results. In statistics there are various rules and condition for the choosing of suitable test statistics. Most of the researcher find the mean and standard deviation but they don't know the concept of normality and rational for choosing of the particular test. Why they calculate only mean and standard deviation, it becomes a grate issue in research. So, the result and inference are not highly valid. In most of the Intuition only few of the researchers follow the rule during the data analysis. In various research works published in biomedical journals journal we can see and observed the use of wrong or inappropriate statistical test used in the data analysis [1,2]. In order to do the statistical calculation and analysis there are lot of software. But the problem is that there software cannot chose the suitable test statistics. So, if we follow the suitable way or guideline for the data analysis our result becomes well and inference become valid. Researcher chose the test statistics depending upon the need and type of objective not by the nature or kind of collected data.

Variable: A variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country of birth, capital expenditure, class grades and eye color and vehicle type are examples of variables. It of two type of variable depending upon relationship.

Dependent Variable: A variable that may depend on the other factor is term as dependent variable. For e.g. Exam score as a variable may change depending on the student's genders.

Independent Variable: A variable that does not depend on the other factor is term as independent variable. For e.g. gender doesn't change depending upon the exam score.

Qualitative Data: Those data which cannot be measure but can be count are called qualitative data. These data are also called the categorical variable. For e.g. number of child vaccinated, number of patients cured and number of person died. In descriptive statistics results obtained are presented in the form of percentage, rates, ratios and proportions. In inferential statistics the method employed in the analysis of such data are Z-test for proportion and Chi- square test.

Quantitative Data: Those data which can be measure but not count are called quantitative data. These data are also called continuous data/variable. For e.g. height, weight and B.P., Hb%. In descriptive statistics depending upon their nature (Normally distributed or not), different, measure of central tendency (mean, median and mode) & dispersion (SD, range) will be use. For normally

distributed data mean and SD can calculate and for not normally distributed data median and range will be calculated. In inferential statistics depending upon their nature (Normally distributed or not), Z-test, t test and Mann Whitney test will be used.

Parametric Test: The branch of statistics which assumes that sample data comes from a population that follows a probability distribution based on a fixed set of Parameter like mean, S.D. [4].One of assumption of the parametric is that the data is normally distributed. Examples of such tests are mean, paired t test, ANOVA etc.

Non-Parametric: Nonparametric test refer to a statistical method wherein the data is not required to fit a normal distribution. Nonparametric statistics uses data that is often ordinal, meaning it does not rely on numbers, but rather a ranking or order of sorts.

In this test there is no assumption about the normality. Examples of non-parametric tests include the various forms of chi square tests, Fisher Exact Probability test, ManneWhitney test, Wilcoxon Signed-Rank test etc.

2. NORMAL DISTRIBUTION

The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean. [5]

Test of Normality of Data: Test of normality of the data is the one of the most important step in the data analysis. It gives the path to choose the suitable test statistics. Many parametric statistical methods such as ANOVA, t-test, Parson Correlation etc. Require that the dependent variable is approximately normally distributed for each category of the independent variable. There are various methods of test of normality such as:

1. Using histogram (with normal curve)
2. Normal Q-Q plot
3. Box plot
4. Skewness and Kurtosis value.
5. Kolmogorov and Shapiro wilk test.

Among these tools we use histogram with normal curve and Kolmogorov and Shapiro wilk test. For small sample size we use Shapiro wilk test and for large sample size we use kolmogrovsmimrov test. If P value is ≥ Alpha, reject alternative hypothesis indicating that data is normally distributed. If p-value is less than α, the data is not normally distributed. If data is normally then we can parametric test for the analysis similarly for not normally distributed data non parametric test will be use.

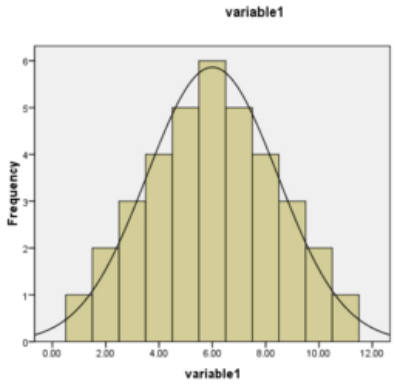

Fig1.*Normally Distributed Data*

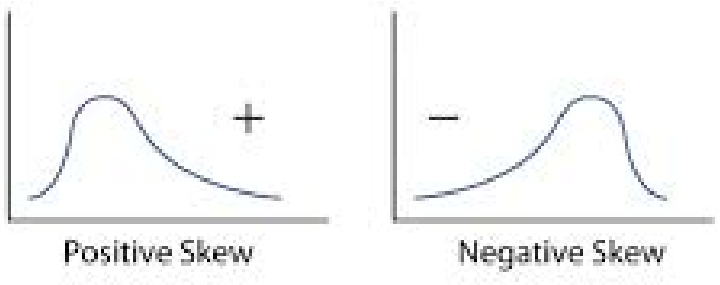

Fig2. *Positively and Negatively Skewed Data*

Paired and Unpaired Data: When we compare two groups then we need to decide whether to use a paired test or unpaired test.Paired observation means if we repeat the measurement on the same subject or individual (before and after an intervention or medication) or measurement on the matched subject. Use an unpaired test to compare groups when the individual values are not paired or matched with one another. For e.g. we take the B.P. of the patients, if his/her B.P. level is abnormal then doctor give the medicine, in the next follow of the same patients doctors again take the B.P., if researcher wants to know whether the given medicine is effective or not at that time after checking normality of the data we used paired t-test. We should select a paired test when values in one group are more

closely correlated with a specific value in the other group than with random values in the other group. It is only appropriate to select a paired test when the subjects were matched or paired before the data were collected. We cannot base the pairing on the data you are analyzing.

3. METHODS OF CHOOSING THE TEST

1. First starts with type or nature of the data whether it is numeric or categorical if it is categorical choose the chi square test. For e.g. let's take the output of SPSS In the chi-square there are five different test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square					
Continuity Correction[b]			.		
Likelihood Ratio					
Fisher's Exact Test					
Linear-by-Linear Association					
N of Valid Cases					

a. 0 cells (.0%) have expected count less than 5.

b. Computed only for a 2x2 table

Rule for 2×2 table in chi-square

a. Continuity Correction: If we have 2×2 table (2 category of each variable, e.g. sex: Male & Female, smoking status: yes, no). In that situation if 0 cells have expected count less than 5 then we take P-value of continuity corrected chi-square.

b. Fisher Exact Test: If we have 2×2 table (2 category of each variable, e.g. sex: Male & Female, smoking status: yes, no). In that situation if some cells have expected count less than 5 then we take P-value of Fisher Exact test.

Rule for more than 2×2 table in chi-square

a. Pearson Chi-Square Test: If we have more than 2×2 table of the categorical variable. In that situation if 0 cells have expected count less than 5 then we take P-value of Pearson chi-square test.

b. Likelihood Ratio: If we have more than 2×2 table of the categorical variable. In that situation if any cells have expected count less than 5 then we take the P-value of likelihood ratio test.

Linear by Linear Association: In any bivariate table in we have ordinal variable in row and column. In that situation we take the P-value of liner by linear association.

2. If we have the numerical variable of only one group which is supposed to be compare with standard value then we use student t- test.
3. If dependent variable is continuous and independent variable have only two categories. Then we check the normality of continuous variable.
    - If the continuous data follows the condition of normally then we use parametric test (Independent t- test).
    - If the data is not normally distributed then we use non parametric test (Mann-Whitney test)
4. If dependent variable is continuous and independent variable have more than two categories. Then we check the normality of continuous variable.
    - If the continuous data follows the condition of normally then we use parametric test (One way ANOVA "Mean comparison").
    - If the data is not normally distributed then we use non parametric test (Kruskal walls test "Median comparison")
5. If we have the parried observation data (Before and after) then check the normality of paired data
    - If the paired data follows the condition of normally then we use (paired t-test)

- If the data is not normally distributed then we use non parametric test (Wilcoxon test)

6. If the objective is to know the correlation between dependent and independent variable. Then we check the normality of dependent variable.
    - If the dependent variable follows the condition of normally then we use parametric test (Pearson correlation).
    - If the dependent variable is not normally distributed then we use non parametric test (Spearman rank correlation)

For e.g. age is a dependent variable (X) and blood pressure is independent variable (Y), then we check the normality of dependent variable and correlation coefficient is obtain which can be interpret in the following ways:

Interpretation of correlation coefficient [2]

- r=-1 :there is perfect neg. correlation between X and Y
- r=1,: there is perfect pos. correlation between X and Y
- r=0, there is no correlation between X and Y
- r=0 to±0.3,there is negligible correlation between X and Y
- r= ± 0.3 to ± 0.49 there is week correlation between X and Y
- r= ± 0.5 to ± 0.69,ther is average correlation between X and Y
- r= ± 0.7 to ± 0.89 ,there is high correlation between X and Y
- r= ± 0.9 to ± 1 ,there is very strong correlation between X and Y

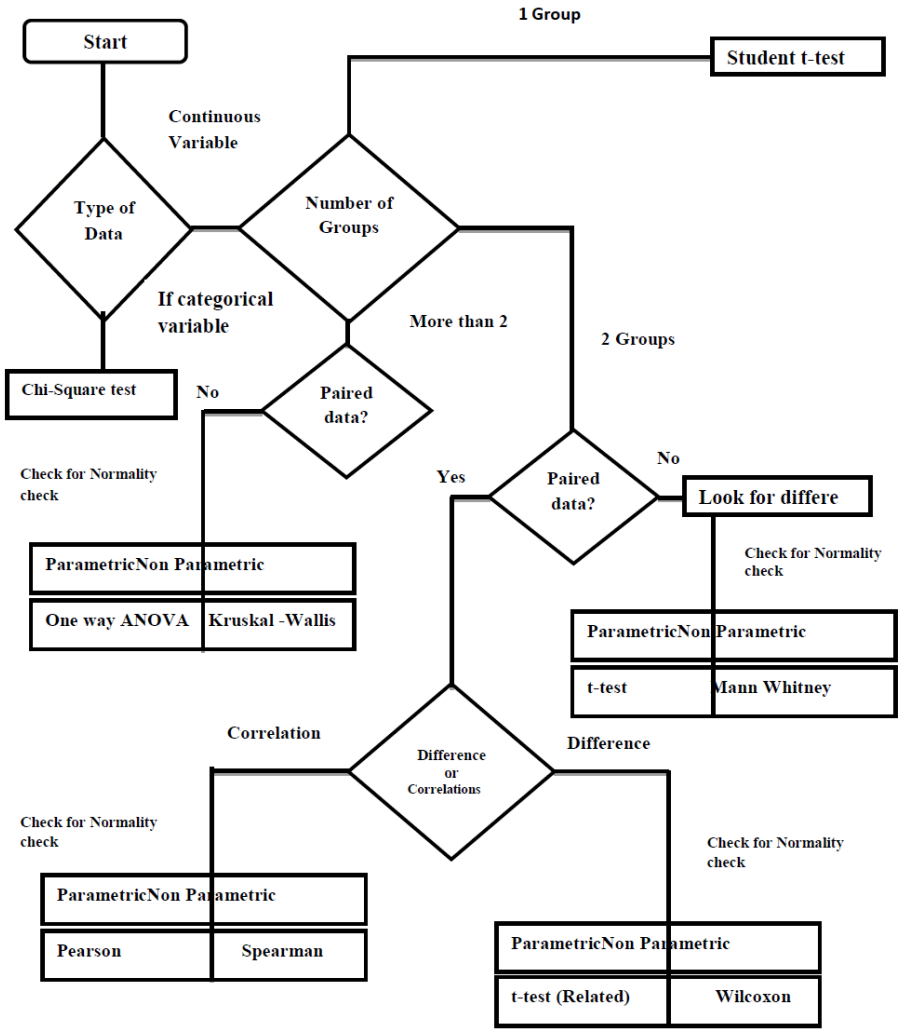

Chart Showing the Way of Choosing Test [2]

REFERENCES

[1] Karan J, Goyal JP, Bhardwaj P, Yadav P. Statistical reporting inIndian pediatrics. Indian Pediatr. 2009; 46:811e812.

[2] Gujarati, D. N. (2009). *Basic econometrics*, New Delli, Tata McGraw-Hill Education.

[3] Lang T. twenty statistical errors event you can find in biomedical research articles. Croat Med J. 2004; 45:361e370.

[4] Geisser, S.; Johnson, W.M. (2006) *Modes of Parametric Statistical Inference*, John Wiley & Sons, ISBN 978-0-471-66726-1

[5] Freedman, D. (2000) *Statistical Models: Theory and Practice*, Cambridge University Press, ISBN 978-0-521-67105-7