

Quality enhancement in health research using biostatistics

S.Valarmathi¹, K.Kanimozhi¹,S.Kalpana¹, Jasmine S Sundar¹, Joseph Maria Adaikalam¹, Parameswari Srijayanth¹, D.Shantharam¹
¹(Department of Epidemiology, The Tamilnadu Dr.MGR Medical University, India)

Abstract: “Absence of evidence is not evidence of absence” –(Altman D G and Bland MJ) ,in the present era where health research is being focused on more evidence based practice, it has become imperative to comprehend the research findings with the support of statistics. Since most researchers these days exacerbate their research findings in an attempt to prove them as statistically significant, dissertation submitted in the department of epidemiology between the years 2006 to 2013 were reviewed. This article, discusses on how biostatistics could be applied to enhance the quality of research in its every sphere thereby avoiding the aberrant research methodology, thus promoting all our pain staking research to give us the pleasure of worth taking.

Keywords : Normality, presentation of data, testing of hypothesis, types of variables, value alignment.

I. Introduction

Statistics explores collection, organization, analysis and interpretation of numerical data. Biostatistics is the application of statistics in the biological and health sciences and it plays a major role in health research. In promotional materials for drugs and other medical therapies it is common to see statistical results which are quoted from research papers. And, number of statistics related to research investigations in medicine are now regularly used in medical literature. Numerical facts being more precise than words in communicating the scientific results, biostatistics does an effective role in exploring the truth and authenticating it as evidence based research. The misuse or inaccurate use of statistical methods may point the research in the wrong path and produce incorrect study results.^[1] The critical aspects of research lies in the statistical analysis, and in the manner in which the findings are prepared and published. Hence, use of statistics is inevitable in medical research and the accurate interpretation of statistical results may pose a challenge for healthcare providers. With the increased availability of statistical software, it is easy to use and misuse the statistical methods, particularly when the user is not aware of the assumption that is to be satisfied before a particular statistical test is used for analysis.^[2] The objective of this paper is to enhance the quality of health research through proper understanding of the statistical concepts and to present the results in the right way.

1 Methods of Preparing Results

Scientific results are presented as a mixture of tables, graphs and diagrams. Always, summaries of data will be given as a support for the study and the raw data will not be given. Statistical analysis could be done in two ways, namely

1. Descriptive Analysis
2. Inferential Analysis

The purpose of descriptive analysis is that, the data will be explained with a single value which will be the most representative value of the samples.

An Inferential analysis enables to arrive at conclusions about the population based on the samples collected. There are different methods for drawing inference which depends on the type of variable and the distribution of the data.

2 Knowledge of Measurement Scales

Statistical analysis of data is done through meaningfully coded variables and in which measurement scales plays a major role . To categorize or to quantify a variable, measurement scales are used. In statistics, we have four types of measurement scales namely Nominal, Ordinal, Ratio and Interval Scales.

A Nominal scale is, making the data into categories without ordering and it also means that there is no distance between categories.

Ordinal data is also making the data into categories but with ordering or ranking and there is distance between categories.

Interval scale is a standard survey rating scale and it has meaningful equidistance between categories.

Ratio scale is the scale which has a true zero point. Interval scale includes negative value whereas ratio scale does not. Table 1 depicts the examples for the scales of measurement.

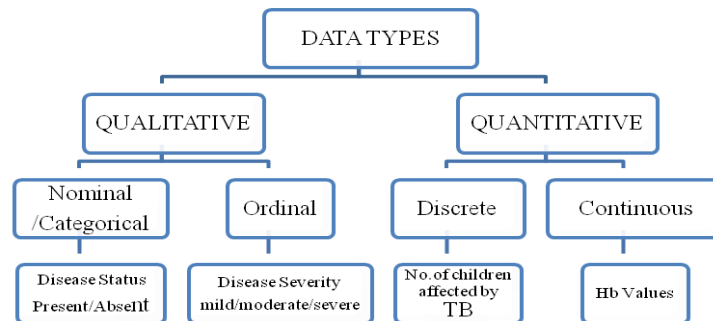
TABLE 1. Examples of measurement scales

Scales of Measurement	Examples	Coding
Nominal	Gender : Male/Female Disease : Present /Absent Blood Group Typing : O,A,B, AB	1/0 1/0 1,2,3,4
Ordinal	Disease Severity : Mild, Moderate, Severe Income Status : Low, Middle, High	1,2,3 1,2,3
Interval	7 point scale : Dissatisfied to Satisfied 5 point scale : Strongly Disagree to Strongly Agree	1,2,3,4,5,6,7 1,2,3,4,5
Ratio	Weight Height Hb values	Actual value or observed number to be recorded and need not be coded.

Answers like “Don’t know” could be coded as 8 and “no response” as 9 which will not fall into any type of measurement scales. Ethically, a researcher could exclude data of doubtful quality.

3 Understanding Data Types

Statistical analysis is the act of transforming data with the aim of extracting useful information in order to infer. Hence, understanding the type of data is essential. Data is represented in the form of variables. A variable is, “a characteristic of a unit being observed “ which could take more than one from a set values. In simple words, variable means, ‘ it varies from unit to unit’.



Always, a study involves two types of variables which are dependent variable and independent variable. Dependent variable is the outcome variable of the study conducted and all the other variables are independent variables. For example, in a study on anemia in girls aged 12-17 years and the study variables are Hb, socio-economic status, BMI, Nutritional status, number of siblings,) etc. Hb, a measure for anemia will be the dependent variable and all the others are independent variables.

4 Reliability of Data

Data is the soul for analysis and so it should be reliable. Data Collection Forms (DCF) or Questionnaire is the backbone of the Research finding. Always DCF should be tested in smaller group (Pilot study – preliminary study) for testing the understandability of the respondent. Reliability of the DCF should be done in order to do an effective research. Cronbach’s alpha should be calculated to check the reliability of DCF especially for the qualitative study. Table 2 gives the idea for proceeding the research with the available DCF.

TABLE 2. Decision Criterion for Cronbach’s Alpha

Cronbach’s Alpha Values	Decision Criteria
$\alpha > 0.9$	Excellent
$0.8 < \alpha < 0.9$	Good
$0.7 < \alpha < 0.8$	Acceptable
$0.6 < \alpha < 0.7$	Questionable
$0.5 < \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable
Source:- http://en.wikipedia.org/wiki/Cronbach's_alpha	

5 Presentation of Tables

A table is the arrangement of rows and columns used to present the results in a concise and well organized manner. A Small table should be clear, well formatted such that the readers could assimilate the data very quickly. Tables can give more precise numerical information than a graph, such as, the number of observations, the mean and some measure of variability of each tabular entry. A well organized and well formatted table should contain table number, title of the table, head note, captions and stubs, body of the table, foot note and source note. Foot note is elaborating the abbreviated headings used in the table. Whenever the tables from other journals, books, newspapers are used, then the source of the table should be mentioned. When presenting the numerical data it should be presented as depicted in table 3.

TABLE 3. Representation of numerical data

Right Presentation	Wrong Presentations	
Right Aligned	Center Aligned	Left Aligned
95.80	95.80	95.80
1046.00	1046	1046
385.58	385.58	385.58

If an dependent and an independent variable are cross tabulated, the headings of the dependent variable should be placed in column and the headings of the independent variable in row.

TABLE 4. Wrong Representation

Variable	Control group (n=66) n(%)	Treatment group (n=83) n(%)
Women	45(68)	54(65)
Mean(SD) age, years	36(7.3)	35(7.0)
Symptomatic	19(29)	26(31)

The right way of representing the cross tabulation is given in table 5 while the wrong way is in table 4. The common errors that could occur in the table are, inaccurate arithmetic values for rows or columns (the numbers do not add up to the total), the abbreviation NA is not explained, the information in the table is repeated word by word in the text and the same is duplicated as graphs.

TABLE 5. Right Representation

Variable	Control group (n=66)	Treatment group (n=83)
Women-n (%)	45(68.0)	54(65.0)
Age In years - Mean(SD)	36(7.3)	35(7.0)
Symptomatic-n (%)	19(29.0)	26(31.0)

The points to be remembered while framing the tables are

1. Prepare the results of the study before framing tables (dummy table)
2. All the tables should be arranged in a logical way
3. Should use minimal decimals in the tables
4. The design and the title of the table should be easy to understand.

Generally, the researchers assume the center alignment of numerical data to be perfect. Out of the 21 dissertations submitted, 20 dissertations had presented their tables which were center aligned and one was aligned perfectly on the right hand side and one of the dissertations had pasted the output derived from the statistical software. However, all the 21 dissertations had followed the right principle of cross tabulation.

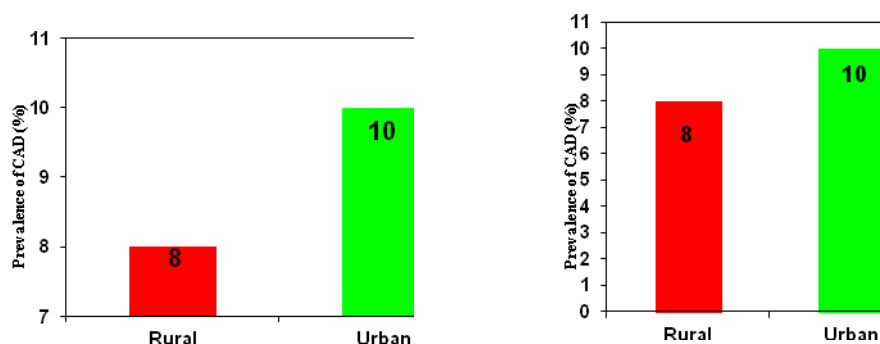
6 Visualization of statistics

Diagrams and graphs are two important methods of presenting data. Visualizing data through graphs and diagrams becomes eye catching. Data which is of utmost importance or that which needs to be highlighted could be presented as graphs or diagrams, in such a way that even a layman could understand the data. The major advantage of diagrams and graphs is that they are impressive and practical. When graphs are drawn, suitable graphs with proper labels, size and legend should be considered.

TABLE 6 . Suitable Diagrams for various types of variables

Type of variables	Suitable diagrams/graphs
DISCRETE NOMINAL ORDINAL	<ul style="list-style-type: none"> • Bar diagram • Pie chart • Pictogram • Spot map
CONTINUOUS	<ul style="list-style-type: none"> • Histogram • Frequency curve • Frequency polygon • Line chart • Scatter diagram

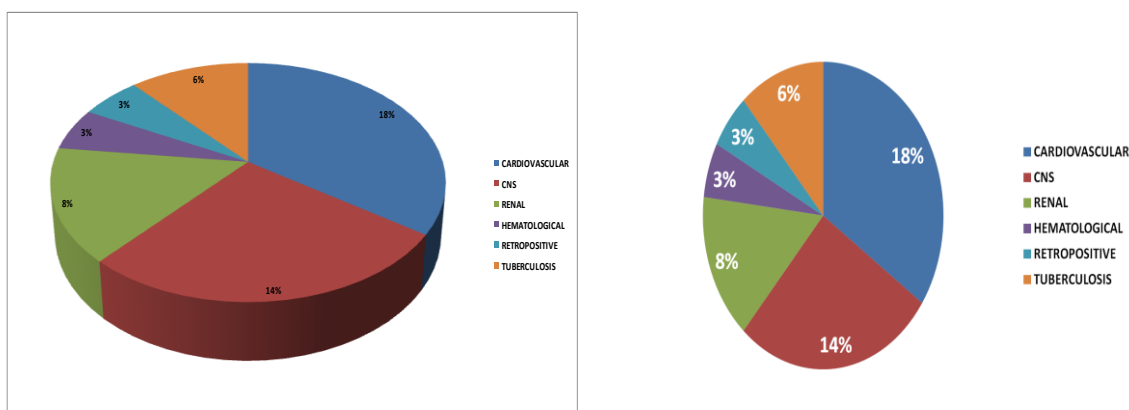
Figure 1.



Distribution of prevalence of Coronary Artery Disease (%)

Fig 1 represents two sets of diagrams for showing the prevalence of Coronary Artery Disease (CAD) in the rural and urban set up. The diagram on right, is the right way of presenting the data, whereas the one on left is wrong, since it exaggerates the data.

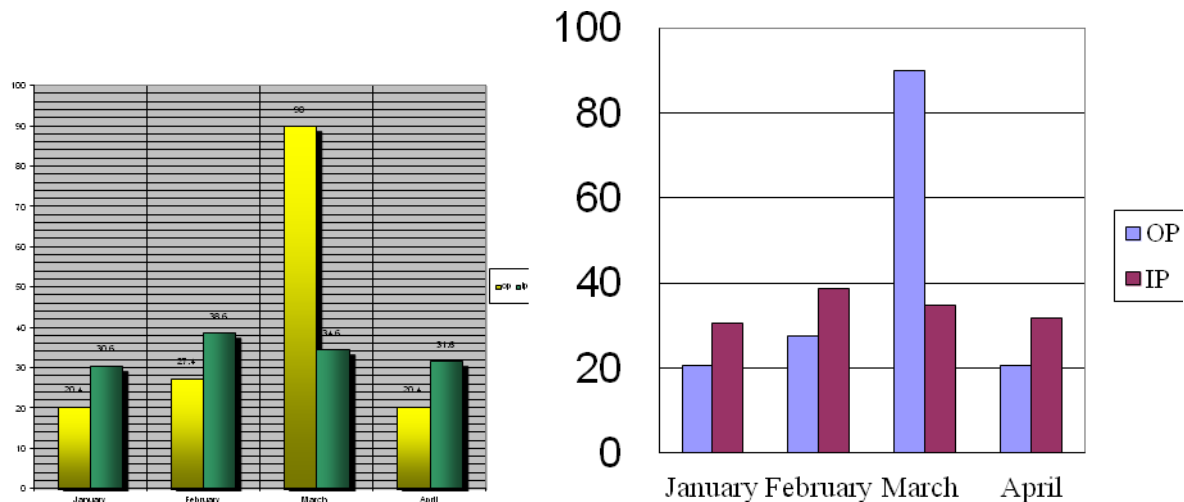
Figure 2 .



Co - morbidities associated with severe acute malnutrition

Any three dimensional or unnecessary shadings in the graphs should be avoided. Fig 2 and 3 shows the right way of representation of data. The diagrams on right side of both figures indicates the correct ones.

Figure 3.



Patients enrolled From January 2013 – April 2013

Always avoid unnecessary gridlines and shading in the graphs since it might distract the reader, however the major gridlines as depicted on the right side of Fig 3 could be used. Using small font or illogical colors should be avoided. The graphs should always be numbered and the title of the graphs or diagrams should always be mentioned. Also remember, that technology is merely a servant to us and it should not rule over humans. So, always use statistically meaningful and acceptable diagrams. Totally, there were 224 diagrams in the 21 dissertations, out of which 71 were three dimensional diagrams and 14 were statistically not acceptable diagrams (diagrams with background shaded) which may acceptable but not statistically meaningful.

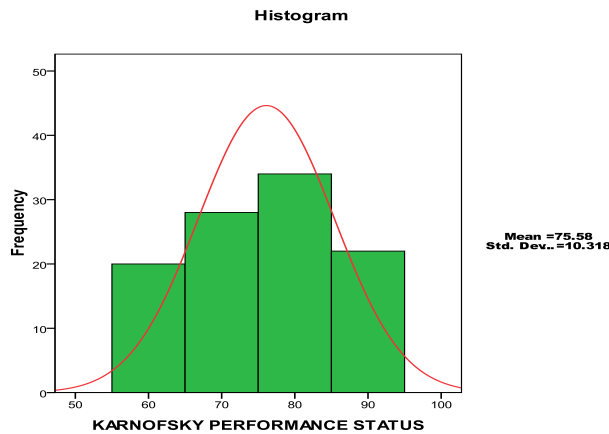
7 Describing the Data

To describe data, summary or descriptive statistics is used. Percentages, Proportions are summary statistics whereas the measures of central tendencies which includes Mean (representative value), Median (middle value), Mode (often repeated value) and the measures of dispersion which are Standard Deviation (variation from mean), Quartiles (dividing the data into four parts), Range (the difference between the highest value and the lowest value) are descriptive statistics. Percentages must be clearly displayed in parentheses with no decimal places, for example to present Summary Statistics, one could write that :- Nearly half 80(49%) of the sample were married and for Descriptive Statistics:- The sample as a whole was relatively young ($M = 19.22$, $SD = 3.45$) which is wrong method of texting whereas the right method is:- The average age of students was 19.22 years ($SD = 3.45$).

8 Evaluating Normality of the Data

A normal distribution is determined by the parameters, mean and standard deviation. There will be different normal distributions for different values of mean and standard deviation. The normal curve is “bell-shaped” and symmetrical in its appearance. W.J Youden expresses his view of the normal distribution as, “The normal curve law of error stands out in the experience of mankind as one of the broadest generalizations of natural philosophy”. Always, it is necessary to evaluate whether the data is normal or not, so that, appropriate tests could be applied and generalized to the whole population. The data must atleast satisfy the condition that it is approximately normal.

Figure 4.



Example for Normal Distribution

The tips for evaluating the normality of the data are as follows :

1. Draw a histogram/frequency curve and check whether the curve is as shown in fig 4.
2. Check whether the mean of the variable is equal to median or mode
3. Standard deviation must be less than half of the mean
4. Kolmogrov Smirnov test for large samples and Shapiro Wilks test for small samples could be used and the results of which are depicted in table 6 as an example. A Researcher would prove his research hypothesis whenever the $P < 0.05$ as significant, but in the case of normality test, one should look for a $P > 0.05$ implying that the data is normally distributed.

TABLE 6. Example for Test for Normality

Tuberculosis Unit (Zone)	Kolmogrov Smirnov test	P value	Shapiro Wilks test	P value
1	0.122	0.163	0.105	0.286
2	0.191	0.268	0.233	0.096
3	0.122	0.217	0.117	0.219

Out of the 21 dissertations, only three had reported about the normality checking and that was in the year 2009, 2011 and 2012. All the others have used parametric as well as non-parametric tests but have missed to report about the normality conditions of the data.

9 Relationship of variables

In medical research, one is often interested to know whether two or more variables are related or not. Association and correlation roughly means the same and is used to indicate when the relationship exists. Often these words seems to be juggled in medical literature but statistically, association is the term used for assessing relationship between categorical variable while correlation is the term used for assessing relationship between continuous variables. The test to be used to know the relationship is depicted in table7.

TABLE 7. Test for finding relationship of variables

Type of variables	Type of group	To be reported	Interpret using
Nominal Vs Nominal	Equal	Contingency table	Chi square test Significant Values
	Not Equal	Contingency table with percentages	
Continuous Vs Continuous	Equal	Correlation	Co-efficient of Determination and Significant Values
Ordinal Vs ordinal Ordinal Vs Continuous	Equal	Spearman rank correlation	Significant values

Usually when Chi square test is used, the variables are cross-tabulated. But when the number of observations obtained for analysis is small, then Chi square may produce misleading results. Chi square is used

when only the expected cells contain values which are greater than or equal to five. If that is not satisfied, one must use Fisher exact test. 10 dissertations have reported chi-square test when the cell values were less than 5, 10 reported properly and one pasted the statistical software output which contained both the above mentioned tests as well as other tests. By default, many tests results will come as a output, but the researcher should know what exactly to report.

10 Testing of Hypothesis

A sample of investigation produces results and with these results, decisions are made about the population. But, such decisions involves an element of uncertainty misleading to wrong decisions. Hypothesis is an assumption which may or may not be true about a population parameter. The goal of testing an hypothesis is to test some hypothesis about the parent population from which the sample is drawn. If the data is normally distributed, parametric tests are to be used, otherwise, non-parametric tests needs to be used.

In the 21 dissertations, there were 620 Chi-square test reported to see the associated factors, out of that 110 chi-square were not interpreted in the results section and 17 just presented the statistical software output. The other tests that were reported are correlation(49), logistic regression(36), Independent t Test(5), Paired t Test(5), in which 2, 2, 1, 2 were not interpreted respectively. 16 non-parametric, 7 advanced statistics and 1 factor analysis have been used and interpreted the results properly

Table 8 is very important for a researcher to understand which type of test has to be applied on their dataset. Different studies uses different tests and the list is exhaustive, however the researcher need not know them all. But, a particular test which is used in the study should be known properly in order to avoid misuse and misinterpretation. The Researcher may use Table 8 as guidelines for the choice of test.

II. Conclusion

A good dissertation is that it is well planned and structured. As altruism, proper use of statistics is a determinant of the best dissertation. Students involved in health research skimp on the time they set aside because of their busy work nature in the wards. Most of the researchers have been taught about statistical significance but in digging the significance they miss out the clinical significance. Students must attend at least one or two workshops on basic biostatistics before they take up their dissertation work. Always one should bear in mind that appropriate statistical methods with respect to the nature of the data and variables should to be used. They need not know all the statistical method however they should be clear with the one they use. The mistakes pointed out in this paper are not a serious issue but when rectified it will enhance the quality of the dissertation such that the researchers work could be a worthwhile contribution to the world.

Table 8 :- Guidelines for the choice of test			
GOAL	NUMERICAL DATA	ORDINAL DATA	NOMINAL DATA
Describe a Group	Mean \pm SD	Median and Interquartile range	Proportion and percentage; mode
Compare 1 group to hypothetical value	One sample t test	Wilcoxon test	Chi square test
To be reported: Degrees of freedom in parentheses; Statistic(rounded to two decimal places) significance level	Texting the results: There was a significant effect for gender, $t(54) = 5.43, p < .001$, with men receiving higher scores than women	Texting the results: The sets of scores differed significantly from hypothetical value with $W(24) = 14.01, Z = 3.22, p = .007$ "	Texting the results: The percentage of participants that were married did not differ by gender, $\chi^2(1, N = 90) = 0.89, p = .35$
Compare 2 unpaired groups	Independent t test	Mann-Whitney test	Fisher's test (Expected cell count <5)
To be reported: Sample means and S.Ds of each groups; Degrees of freedom in parentheses Statistic(rounded to two decimal places) significance level	Texting the results: An independent samples t test was performed comparing the mean consistency scores of college students and individuals diagnosed with schizophrenia. As predicted, the schizophrenics ($M = .76, SD = .20, N = 10$) were more illogical than the college students ($M = .17, SD = .13, N=9$), $t(17) = 7.53, p < .01$, two-tailed	Texting the results: Test Statistic(denoted by U) Level of significance, Drugs users (Median=33.50) were significantly more depressed than alcohol users(Median=7.50), $U=4.00, p<0.001, r=-0.78$ Draw Box-Whiskers Plot.	Texting the results: The percentage of the ownership of device A significantly differed by gender (p -value = 0.038)
Compare 2 paired groups	Paired t test	Wilcoxon signed rank test	Mc Nemar's test
To be reported: Degrees of freedom in parentheses Statistic(rounded to two decimal places) significance level	Texting the results: A significant increase in well-being occurred ($t(9) = -3.14, p = .012$,) in the Prozac treated group compared to the control group.	Texting the results: There was a significant difference in how the faculty ranked the candidates, $z = -2.33, p < .05$. The results indicate that Dr. Ram was the preferred candidate and received significantly more favourable rankings than Dr. Ramesh	Texting the results: Kappa Statistic should reported There is no difference in the opinions of the two cardiology experts about the patients who were suspected to have ischemic heart disease.
Quantify association between 2 variables	Pearson correlation	Spearman correlation	Chi square
To be reported: Degrees of freedom in parentheses Statistic(rounded to two decimal places) significance level	Texting the results The two variables were correlated, $r(55) = .49, p < .01$. There was a significant correlation between amount of education and annual income, $r = .65, n = 30, p < .01$, two tails. Higher annual income was associated with more years of education.		Texting the results The percentage of participants that were married did not differ by gender, $\chi^2(1, N = 90) = 0.89, p = .35$
Predict value from another measured variable	Simple linear regression	Ordinal Regression	Simple logistic regression
Predict value from several measured or binomial variables	Multiple Linear regression	Ordinal Regression	Multiple logistic regression
To be reported	R-squared value, Beta coefficients, Confidence Interval Significant level, ANOVA Scatter Plot for Independent Vs Dependent Variable with Best line	Pseudo R squared value Odds Ratio Confidence Interval	R-squared value, Exp(B) and its confidence interval Hosmer and Lemeshow test P Value (LR or Wald test)

Acknowledgement

We profusely thank Dr.K.Sathish kumar and Dr.G.Srinivasan ,The Tamilnadu Dr.MGR medical university without whom this article would not have bloomed unto this stage of a dream article becoming a reality.

References

Journal papers:

- [1] Shunquan Wu,¹Zhichao Jin,¹Xin Wei,²Qingbin Gao,¹Jian Lu,¹Xiuqiang Ma,¹Cheng Wu,¹Qian He,¹Meijing Wu,¹Rui Wang,¹Jinfang Xu,¹and Jia He¹ 'Misuse of Statistical Methods in 10 Leading Chinese Medical Journals in 1998 and 2008, *Scientific World Journal*. 2011; 11: 2106–2114.Published online 2011 November 2.
- [2] Krithikadatta.J, Valarmathi.S, Research Methodology in Dentistry :Part-II The relevance of statistics in research, *J.Conserv Dent* 2012;15:206-13.
- [3] Gardenier JS, Resnik DB. The misuse of statistics: concepts,tools, and a research agenda. *Account Res*. 2002;9:65–74.

Books:

- [4] David E.Matthews , Vernon.T.Farewell , *Using and understanding medical statistics* 3rd revised edition, S.Karger AG,Switzerland,1996.