ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2016; 2(10): 71-75
www.allresearchjournal.com
Received: 12-08-2016
Accepted: 13-09-2016

Anirban Goswami
Investigator (Statistics), Regional Research Institute of Unani Medicine, Patna, Bihar, India

Dr. Rajesh
Research Officer (U), Scientist L-1, Regional Research Institute of Unani Medicine, Patna, Bihar, India

Dr. Mohd Washim Ahmed
Research Officer (U), Scientist L-1, Regional Research Institute of Unani Medicine, Patna, Bihar, India

Correspondence
Anirban Goswami
Investigator (Statistics), Regional Research Institute of Unani Medicine, Patna, Bihar, India

Utilization of statistical test in clinical trials for categorical data

Anirban Goswami, Dr. Rajesh and Dr. Mohd Washim Ahmed

Abstract
A clinical trial is increasingly based on the empirical studies and the results of these are usually presented and analyzed with statistical methods. Therefore discuss frequently used statistical tests for categorical data. Advice will be presented for selecting statistical tests on the basis of very simple cases. It is therefore an advantage for any physician he/she is familiar with the frequently used statistical tests, as this is the only way he or she can evaluate the statistical methods in scientific publications and thus correctly interpret their findings.

Keywords: Clinical trials, categorical data, statistical test

Introduction
Clinical trials are conducted to collect and recorded data on each subject, such as the patient's demographic characteristics, disease related risk factors, medical history, biochemical markers, pathological history, medical therapies, and outcome or endpoint data at different time points. This data may be categorical or continuous. Understanding that the types of data are more important as they determine which method of data analysis is to be use and how to report the results [1]. For the assessment of the safety, efficacy, and / or the mechanism of action of an investigational medicinal product, or new drug or device that is in development.
In clinical trials, patient's and investigator's responses to treatments can be documented according to the occurrence of some meaningful and well-defined event such as death, infection, or cure of a certain disease and any serious adverse events. In addition the intensity of these events can be graded according to some predefined categories. Therefore categorical data can be useful surrogate endpoints for some unobserved latent or hypothetical continuous variables in clinical trials. Sometimes, to provide an easy analysis and/or a better presentation of the results, continuous data are transformed into the categorical data with respect to some predefined criteria. As a result, many efficacy and safety endpoints in clinical trials are in the form of categorical data on either a nominal or ordinal scale [2]. The different types of statistical methods are used to analyze this data in clinical trials.
Categorical data arises whenever a variable is measured on a scale that simply classifies respondents into a limited number of mutually exclusive (incompatible) groups. Data can be divided into two main types: quantitative and qualitative. Quantitative data can be either continuous variables that one can measure (such as height, weight, or blood pressure) or discrete variables (such as numbers of patients attained in OPD per day or numbers of attacks of asthma per child per month). Count data can be discrete and quantitative. Qualitative data tend to be categories; people are male or female, Indian or Bangladeshi, they have a disease or are in good health and they are belonging to lower or middle or higher socio-economic status. These are examples of categorical data. There are four types of scales that appear in social sciences: nominal, ordinal, interval, and ratio scales. They are categorized into two groups: categorical and continuous scale data. Nominal and ordinal scales are categorical data; interval and ratio scales are continuous data. When categorical data has unordered scales it is called nominal scales. Blood group, gender are example of the nominal scale. Categorical data that has ordered scales are called ordinal scale. Severities of illness, amount of pain are example of ordinal scale. There should be distinction between them because the data analysis method is different depending on the scale of measurement [3].

Statistical test used to analyze the categorical data

Chi-square test (of independency): The chi-square test of independency is used to the association between two independence categorical variables. The idea behind this test is to compare the observed frequencies with the frequencies that would be expected if the null hypothesis of no association/statistically independence were true. By assuming the variables are independent, we can also predict an expected frequency for each cell in the contingency table. If the value of the test statistic for the chi-squared test of association is too large, it indicates a poor agreement between the observed and expected frequencies and the null hypothesis of independence/no association is rejected. For example in clinical trials, it will be used to test the association between adverse event and the treatment used. The assumptions of chi-square test as independent random sampling, no more than 20% of the cells have an expected frequency less than five, and no empty cells. If the chi-square test shows significant result, then we may be interested to see the degree or strength of association among variables, but it fails to explain another situation where more than or equal to 20% of the cells have an expected frequency less than five. In this case, the usual chi-square test is not valid. Then the Fisher Exact test will be used to test the association among variables. This method also fails to give the strength of association among variables.

Chi-square test (of Homogeneity): The chi-square test of homogeneity is applied to a single categorical variable from two different populations. It is used to determine whether frequency counts are distributed identically across different populations. We can use this test under the assuming for each population, the sampling method is simple random sampling and sample data are displayed in a contingency table (Populations x Category levels), the expected frequency count for each cell of the table is at least 5. For example, in multicenter clinical trials it will be used to test differences among the centres for response of the particular drug(s).

Cochran Armitage trend test: In clinical trials, it is often of interest to investigate the relationship between the increasing dosage and the effect of the drug under study. Usually the dose levels tested are ordinal, and the effect of the drug is measured in binary. In this case, Cochran-Armitage trend test [4, 5] is used to test for trend among binomial proportions across levels of a single factor or covariate. This test is appropriate for a two-way table where one variable has two levels and the other variable is ordinal. The two-level variable represents the response, and the other variable represents an explanatory variable with ordered levels.

McNemar test: In clinical trials, It's used when researcher interested to the test of improvement in response rate after a particular treatment or finding a change in proportion for the paired data (e.g., studies in which patients serve as their own control, or in studies with before and after design). The three main assumptions for this test are variable must be nominal with two categories (i.e. dichotomous variables) and one independent variable with two connected groups, two groups of the dependent variable must be mutually exclusive and sample must be a random sample and no expected frequencies should be less than five. Data should be placed into a 2×2 contingency table, with the cell frequencies equalling the number of pairs. For example, a researcher is testing a new medication and records if the drug worked ("yes") or did not ("no").

Cochran's Q tests: This test is used to determine if there are differences on a dichotomous dependent variable between three or more related groups. In addition, when a binary response is measured several times or under different conditions, Cochran's tests that the marginal probability of a positive response is unchanged across the times or conditions. The Cochran Q test is an extension to the McNemar test for related samples that provides a method for testing the differences between three or more matched sets of frequencies or proportions. We can use this test under the assuming for one dependent variable with two, mutually exclusive groups (i.e., the variable is dichotomous), dichotomous variables include perceived safety (two groups: "safe" and "unsafe"), one independent variable with three or more related groups and the cases (e.g., participants) are a random sample from the population of interest. For example, the data set drugs contain data for a study of three drugs to treat a chronic disease [2] and forty-six subjects receives drugs A, B, and C. The response to each drug is either favorable or unfavorable and to test that differences of favorable response for the three drugs.

Generalized McNemar/Stuart-Maxwell Test: The generalization of McNemar's test extend 2x2 square tables to KxK tables is often referred to as the generalized McNemar or Stuart-Maxwell test [6, 7]. In clinical trials, this testing is used to analyze matched-pair pre–post data (treatment) with multiple discrete levels (e.g. severity of pain) of the exposure (outcome) variable.

Bhapkar's test: This test is the marginal homogeneity by exploiting the asymptotic normality of marginal proportion [8]. The idea of constructing test statistic is similar to the one of generalized McNemar's test statistic, and the main difference lies in the calculation of elements in variance-covariance matrix. Although the Bhapkar and Stuart-Maxwell tests are asymptotically equivalent [9]. Bhapkar test is a more powerful alternative to the Stuart-Maxwell test. In large sample both will produce the same chi-squared value [8].

Cohen's kappa statistic: Cohen's kappa statistic is a measure of agreement between categorical variables. For example, kappa can be used to compare the ability of different raters to classify subjects into one of several groups. Kappa also can be used to assess the agreement between alternative methods of categorical assessment when new techniques are under study. In clinical aspect, comparison of a new measurement technique with an established one is often needed to check whether they agree sufficiently for the new to replace the old. Correlation is often misleading [10]. Cohen's Kappa used and the level of agreement between raters were assessed in terms of a simple categorical diagnosis (i.e., the presence or absence of a disorder).

The kappa coefficient (κ) is used to assess inter-rater agreement. One of the most important features of the kappa statistic is that it is a measure of agreement, which naturally controls for chance. Kappa is always less than or equal to 1.

A value of 1 implies perfect agreement and values less than 1 imply less than perfect agreement. In rare situations, Kappa can be negative. This is a sign that the two observers agreed less than would be expected just by chance. Possible interpretation of kappa coefficient (κ) as follows:

- Poor agreement = Less than 0.20
- Fair agreement = 0.20 to 0.40
- Moderate agreement = 0.40 to 0.60
- Good agreement = 0.60 to 0.80
- Very good agreement = 0.80 to 1.00

Wilcoxon signed-rank test: The Wilcoxon signed rank test is a non-parametric or distribution free test for the case of two related samples or repeated measurements on a single sample. It can be used (a) in place of a one-sample t-test (b) in place of a paired t-test or (c) for ordered categorical data where a numerical scale is inappropriate but where it is possible to rank the observations when the population can't be assumed to be normally distributed. For example, the hours of relief provided by two analgesic drugs in patients suffering from arthritis and to test that one drug provides longer relief than the other.

Mann–Whitney U test (Wilcoxon Rank Sum Test): The Mann–Whitney U test is a non-parametric or distribution free test to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed. The Mann-Whitney (or Wilcoxon-Mann-Whitney) test is sometimes used for comparing the efficacy of two treatments in clinical trials. It is often presented as an alternative to a *t* test when the data are not normally distributed. Whereas a't' test is a test of population means, the Mann-Whitney test is commonly regarded as a test of population.

Kruskal-Wallis H test: The Kruskal-Wallis H test is a rank-based nonparametric test that can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable. Sometimes this test described as an ANOVA with the data replaced by their ranks. It is an extension of the Mann-Whitney U test to three or more groups. For example in clinical trials, it will be used to test assess differences in albumin levels in adults different diets with different amounts of protein.

Friedman Post Hoc test: It is a non-parametric test (distribution-free) used to compare observations repeated on the same subjects. This test is an alternative to the repeated measures ANOVA, when the assumption of normality or equality of variance is not met. Friedman's Test and found a significant P- value, that means that some of the groups in data have different distribution from one another, but it is don't know which. There for, it is needed to find out which pairs of groups are significantly different then each other. But when we have N groups, checking all of their pairs will be to perform [n over 2] comparisons, thus the need to correct for multiple comparisons arises. In that situation we will used the Friedman Post Hoc test. In clinical trials, this test find out the improvement of the drug(s) among the patients follow ups for a particular disease.

Cronbach's α (alpha) Statistic: The Cronbach's alpha is a statistic for investigating the internal consistency of a questionnaire [11, 12]. Generally, many quantities of interest in medicine, such as anxiety or degree of handicap, are impossible to measure explicitly. In such cases, we ask a series of questions and combine the answers into a single numerical value. For example, Quality of Life (QoL) scale used in clinical trials should have demonstrated reliability and validity, and be responsive to change in health status, reliability is assessed through examination of the internal consistency at a single administration of the instrument using Cronbach's α (alpha).

Odds Ratio (OR): The Odds ratio is the ratio of the odds of disease in the exposed to the odds of disease in the non-exposed. It is used to measure of association the risk of a particular outcome (or disease) if a certain factor (or exposure) is present. In addition, odds ratio is a relative measure of risk, telling us how much more likely it is that someone who is exposed to the factor under study will develop the outcome as compared to someone who is not exposed.

For a 2x2 contingency table:

- OR=1 suggests there is an equal chance of getting the disease among exposed group compared to unexposed group.
- OR>1 suggests there is a more chance or likelihood of getting the disease exposed group compared to unexposed group.
- OR<1 suggests there is a less chance or likelihood of getting the disease among exposed group compared to unexposed group. Odds ratio can be used in both retrospective and prospective studies.

The Odds Ratio useful to analyse associations between groups from case-control and prevalent (or cross-sectional) data, rare diseases (or diseases with long latency periods) the OR can be an approximate measure to the RR (relative risk) and to estimate the strength of an association between exposures and outcomes.

Relative Risk (RR): The risk of the disease is probability of an individual becoming newly disease given that the individual has the particular attribute. The Relative Risk is a ratio of the risk of disease for those with the risk factor to the risk of disease for those without the risk factor. In clinical trials, it is used to compare the risk of developing a disease in people not receiving the treatment (or receiving a placebo) versus people who are receiving the treatment. Alternatively, it is used to compare the risk of developing a side effect in people receiving a drug as compared to the people who are not receiving the treatment.

For a 2x2 contingency table:

- RR=1 implies that the two groups (exposed and unexposed group) have same risk.
- RR>1 implies that higher risk of getting disease among exposed group compared to unexposed group.
- RR<1 implies that lower risk of getting disease among exposed group compared to unexposed group.

Simpson's Paradox: Simpson's paradox, also known as Yule–Simpson effect was first described by Yule [13] and is named after Simpson's [14]. In clinical trials, Simpson's Paradox arises when the association between an exposure and an outcome is investigated but the exposure and outcome are strongly associated with a third variable. This is a real-life example from a medical study comparing the success rates of two treatments for kidney stones [15].

Tests for Linear Trend: In clinical study researcher may interested to dose-response effect, that is situation in which an increased value of the risk factor means a greater likelihood of disease. It is used to test for a dose-response trend whenever the different level of the risk factor (i.e., the risk factor is ordinal or at least treated as such). Armitage described the details of the theory [4]. For example, it is used to trend test of prevalence cough would be greater for greater amount of smoking.

Tests for Nonlinearity: Sometimes the relationship between the risk factor and disease is nonlinear. For example, it could be that low and high doses of the risk factor are harmful compared with average doses. In this case a U-shaped relationship has been found by the several authors who have investigated the relationship between alcohol consumption and death from any cause and to test the nonlinear relationship [16].

Sensitivity, specificity, Predictive Value Positive Test (PPT) and Predictive Value Negative Test (NNT)

- **Sensitivity:** Sensitivity of a test is the ability to identify correctly those who have the disease and it is the proportion of patients with disease in whom the test is positive.
- **Specificity:** Specificity of a test is the ability to identify correctly those who do not have the disease and it is the proportion of patients without disease in whom the test is negative.
- **Predictive Value Positive Test (PPT):** Predictive value of a positive test is the likelihood of an individual with a positive test has the disease.
- **Predictive Value Negative Test (NNT):** Predictive value of a negative test is the likelihood of an individual with a negative test Predictive value of a positive test is the likelihood of an individual with a positive test does not have the disease

Mantel Haenszel (MH) test: Mantel Haenszel (MH) statistic used to analysis of two dichotomous variables while adjusting for a third variable to determine whether there is a relationship between the two variables controlling for levels of the third variable. For example, compare the frequency of smoking vs. non-smoking in teenage boys vs. girls in several different cities for 2x2 replicated tables.

Cochran Mantel Haenszel (CMH) test: Mantel Haenszel is a non-model based test used to identify confounders and to control for confounding in the statistical analysis. It is used to test the conditional independence in 2x2xK tables. The Cochran-Mantel-Haenszel test is often used in the comparison of response rates between two treatment groups in a multi-center study using the study centres as strata [2]. The CMH can be generalized to IxJxK tables.

Log-rank test: The Log-rank test is a nonparametric test to comparing distributions of time until the occurrence of an event of interest among independent groups. The event is often death due to disease, but event might be any binomial outcome, such as cure, response, relapse, or failure. Examples where use of the log-rank test might be appropriate include comparing survival times in cancer patients who are given a new treatment with patients who receive standard chemotherapy, or comparing times-to-cure among several doses of a topical antifungal preparation where the patient is treated for 10 weeks or until cured, whichever comes first.

Permutation test: Permutation test is used to perform a nonparametric test to find out the difference between treatment groups in the assessment of new medical interventions. In addition, it is used to study efficacy in a randomized clinical trial which compares, in a heterogeneous patient population, two or more treatments, each of which may be most effective in some patients, when the primary analysis does not adjust for covariates. The general discussion and application of permutation test describe by Zucker, D.M [17].

Conclusion

Different type of statistical test is used to analyze the categorical data in different situations and nature of the data. The statistical test has its limitations, and to overcome that another method is used. Before using the statistical test we need to check the assumptions and type of the study. Most of these statistical tests play a very important role to getting appropriate and desired result in clinical trials, to make the decision on the objectives. Researchers / Physicians are helpful to used statistical tests to determine results from experiments, clinical trials of medicine and symptoms of diseases. The use of statistical test in medicine provides generalizations for the public to better understand their risks for certain diseases, links between certain behaviors of diseases, effectiveness of drug(s) and to significant finding of experimental objectives.

References

1. Wang D, Bakhai A. Clinical Trials-A Practical Guide to Design, Analysis, and Reporting, Remedica Publishing, USA. 2006.
2. Agresti A. Categorical Data Analysis (2nd Ed.), John Wiley & Sons, New Jersey. 2002.
3. Campbell MJ. Statistics at Square Two (2nd Ed.). Blackwell, USA, 2006.
4. Armitage P. Tests for Linear Trends in Proportions and Frequencies, Biometrics. 1955; 11:375-386.
5. Cochran WG. Some Methods for Strengthening the Common Chi-Square Tests, Biometrics. 1954; 10:417-51.
6. Stuart A. A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification, Biometrika. 1955; 42:412-416.
7. Maxwell AE. Comparing the classification of subjects by two independent judges, British Journal of Psychiatry. 1970; 116:651-655.
8. Bhapkar VP. A note on the equivalence of two test criteria for hypotheses in categorical data, Journal of the American Statistical Association. 1966; 61:228-235.
9. Keefe TJ. On the relationship between two tests for homogeneity of the marginal distributions in a two-way classification, Biometrics. 1982; 69:683-684.
10. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between two methods of clinical measurement, lancet. 1986; i:307-10.
11. Cronbach LJ. Coefficient alpha and the internal structure of test, Psychometrika. 1951; 16:297-334.
12. Bland JM, Altman DG. Statistics notes Cronbach's alpha, British Medical Journal. 1997; 314:572.

13. Yule G. Notes on the theory of association of attributes of statistics, Biometrika. 1903; 2: 121-134.
14. Simpson EH. The Interpretation of Interaction in Contingency Tables, Journal of the Royal Statistical Society, Series B. 1951; 13:238-241.
15. Julious SA, Mullee MA. Confounding and Simpson's paradox, British Medical Journal. 1994; 309:1480-1481.
16. Duffy JC. Alcohol consumption and all-cause mortality, International Journal of Epidemiology. 1995; 24(1):100-5.
17. Zucker DM. Permutation Tests in Clinical Trials, Wiley Encyclopedia of Clinical Trials. 2007. (http://pluto.mscc.huji.ac.il/~mszucker/DESIGN/perm.pdf).