International Journal of Medicine Research
ISSN: 2455-7404; Impact Factor: RJIF 5.42
www.medicinesjournal.com
Volume 1; Issue 5; November 2016; Page No. 36-44

Application of statistical test in clinical research

[1] Anirban Goswami, [2] Dr. Mohd Wasim Ahmed, [3] Dr. Rajesh, [4] Dr. Najmus Sehar, [5] Dr. Mohd Ishtiyaque Alam

[1] Investigator (Statistics), Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India
[2, 3] Research Officer (U), Scientist L-1, Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India
[4] Research Officer (U), Scientist L-3, Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India
[5] Research Officer Incharge, Scientist L-4, Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India

Abstract
Clinical research 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 different type of data set under assumption of normality or non-normality. The statistical tests applied when normality (and homogeneity of variance) assumptions are satisfied otherwise the equivalent non-parametric statistical test used. Advice will be presented for selecting statistical tests on the basis of very simple cases. It is therefore an advantage for any physician or researcher 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 research, statistical test

1. Introduction
Clinical research 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 continuous or discrete. Understanding that the types/assumptions 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.

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). Qualitative data tend to be categories; people are male or female, Indain or Bangladeshi, they have a disease or are in good health and they are belonging to lower or middle or higher socio-economic status. 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 or non-parametric data; interval and ratio scales are continuous data or parametric 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 [2].

In clinical research, 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, any serious adverse events, biochemical and pathological findings. In addition the nature of these data can be parametric or non-parametric. In this regard parametric test is used on parametric data, while non-parametric data is examined with a non-parametric test. Parametric statistical tests are done when data follow the normal distribution. Parametric test are the most powerful statistical test because they use all of the information in the numbers. Non-parametric statistical test are used when the data don't follow a particular distribution but can be ordered and sometimes are called distribution free test.

2. Statistical test used in Clinical Research
Z-test
A Z-test is a hypothesis test based on the Z-statistic, which follows the standard normal distribution under the null hypothesis. This test is used when the outcome is continuous and the exposure, or predictor, is binary. We can use this test under the assuming for the sample size is greater than 30, observations should be independent from each other, one observation isn't related or doesn't affect another observations, data should be followed normally distributed and data should be randomly selected from a population, where each item has an equal chance of being selected. There are two type of test under Z-test as one sample Z-test and two sample Z-test. The one sample Z-test, which tests the mean of a normally distributed population with known variance. For example in clinical research, if someone said they had found a new drug that cures cancer, some other would want to be sure it was probably true. A hypothesis test will tell him if it's probably true, or probably not true. The two sample Z-test

used to determine whether two population means are different when the variances are known and statistic is assumed to have a normal distribution. For example in clinical research, suppose two flu drugs A and B, Drug A works on 41 people out of a sample of 195, Drug B works on 351 people in a sample of 605 and to test the effect of two drugs equal or not.

Student t-test

Student t-test, in statistics, a method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown. We can use this test under the assuming for the sample size is lesser than 30, observations should be independent from each other, one observation isn't related or doesn't affect another observations, data should be followed normally distributed and data should be randomly selected from a population, where each item has an equal chance of being selected. There are two type of Student t-test under one sample and two sample. One sample student t-test is a statistical procedure used to examine the mean difference between the sample and the known value of the population mean. It is used to determine if a mean response changes under different experimental conditions. In other hand, two-sample t-test is used to compare the means of two independent populations, denoted μ1 and μ2 with standard deviation of the populations should be equal. This test has ubiquitous application in the analysis of controlled clinical research. For example in clinical research, the comparison of mean decreases in diastolic blood pressure between two groups of patients receiving different antihypertensive agents, or estimating pain relief from a new treatment relative to that of a placebo based on subjective assessment of percent improvement in two parallel groups [3, 4].

Student paired 't' test

It is a statistical technique that is applied to paired data of independent observations from one sample only when each individual gives a pair of observation or compare two population means in the case of two samples that are correlated. Paired sample t-test is used in 'before-after' studies, or when the samples are the matched pairs, or when it is a case-control study. We can use this test under assumptions of the number of observations in each data set must be the same, and they must be organized in pairs, in which there is a definite relationship between each pair of data observations, data were taken as random samples follows as Normal distribution and the variance of two samples is equal, Cases must be independent of each other. This statistical test used in clinical research to compare the effect of two drugs, given to the same individuals in the sample at two different occasions, e.g., adrenaline and noradrenalin on puls rate, number of hours for which sleep is induced by two hypnotics and so on [5].

Hotelling's T^2 test

Hotelling's T^2 test is the multivariate generalization of the Student's t-test [6]. Hotelling's T^2 test should be described by multiple response variables. A one-sample Hotelling's T^2 test can be used to test if a set of objects (which should be a sample of a single statistical population) has a mean equal to a hypothetical mean. A two-sample Hotelling's T^2 test may be used to test for significant differences between the mean vectors (multivariate means) of two multivariate data sets. This test can be used under the assumptions (1) the variables of each data set follow a multivariate normal distribution, each variable may be tested for univariate normality, (2) the objects have been independently sampled, (3) in a two-sampled test, the two data sets being tested have (near) equivalent variance-covariance matrices, Bartlett's test may be used to evaluate if this assumption holds, (4) each data set describes one population with one multivariate mean. No subpopulations exist within each data set. Example in clinical research, a certain type of tropical disease is characterized by fever, low blood pressure and body aches. Suppose a researcher team are working on a new drug to treat this type of disease and wanted to determine whether the drug is effective. They took a random sample of 20 people with this type of disease and 18 with a placebo. Based on the data they wanted to determine whether the drug is effective at reducing these three symptoms.

ANOVA

Analysis of variance (ANOVA) is used in statistics that splits the total variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, but the random factors do not. Analysts use the ANOVA test to determine the result independent variables have on the dependent variable amid a regression study. It is an extension of the two-sample t-test and Z-test. In 1918, Ronald Fisher developed a test called the analysis of variance. This test is also called the Fisher analysis of variance, used to the analysis of variance between and within the groups whenever the groups are more than two[7]. When we set the Type one error to be 0.05, and we have several of groups, each time we tested a mean against another there would be a 0.05 probability of having a type one error rate. This would mean that with six T-tests we would have a 0.30 (.05×6) probability of having a type one error rate. This is much higher than the desired 0.05. ANOVA creates a way to test several null hypothesis at the same time at the Type one error 0.05. We can use this test under the assuming, each group sample is drawn from a normally distributed population, all populations have a common variance, all samples are drawn independently of each other, within each sample, the observations are sampled randomly & independently of each other and factor effects are additive in nature. Example in clinical research, ANOVA method might be appropriate for comparing mean responses among a number of parallel-dose groups or among various strata based on patients' background information, such as race, age group, or disease severity [4].

ANCOVA

In clinical research, patients who meet inclusion and exclusion criteria are randomly assigned to each treatment group. Under the assumption of targeted patient population is homogeneous, we can expect that patient characteristics such as age, gender, and weight are comparable between treatment groups. If the patient population is known to be heterogeneous in terms of some demographic variables, then a stratified randomization according to these variables should be applied. At the beginning of the study, clinical data are usually collected at randomization to establish baseline values. After the administration of study drug, clinical data

are often collected at each visit over the entire duration of study. These clinical data are analyzed to assess the efficacy and safety of the treatments. As pointed out earlier, before the analysis of endpoint values. Characteristics between treatments of the patient are usually examined by an analysis of variance (ANOVA) if the variable is continuous. For the analysis of endpoint values, although the technique of analysis of variance (ANOVA) can be directly applied, it is believed the endpoint values are usually linearly related to the baseline values. Therefore an adjusted analysis of variance should be considered to account for the baseline values. This adjusted analysis of variance is called analysis of covariance (ANCOVA) [8]. In addition, ANCOVA provides a method for comparing response means among two or more groups adjusted for a quantitative concomitant variable, or 'covariate', thought to influence the response. The attention here is confined to cases in which the response, y, might be linearly related to the covariate, x. ANCOVA combines regression and ANOVA methods by fitting simple linear regression models within each group and comparing regressions among groups. Assumptions for ANCOVA as each independent variable, the relationship between the response (y) and the covariate (x) is linear, the lines expressing these linear relationships are all parallel (homogeneity of regression slopes), the covariate is independent of the treatment effects (i.e. the covariant and independent variables are independent. ANCOVA might be applied 1) comparing cholesterol levels (y) between a treated group and a reference group adjusted for age (x, in years) 2) comparing scar healing (y) between conventional and laser surgery adjusted for excision size (x, in mm) 3) comparing exercise tolerance (y) in 3 dose levels of a treatment used for angina patients adjusted for smoking habits (x, in cigarettes/day).

Bartlett Test

Bartlett Test can be used to test for homogeneity of variance [9]. In addition, it can be used when the variances across groups are not equal, the usual analysis of variance assumptions are not satisfied and the ANOVA *F* test is not valid and equal sample sizes from several normal populations. For example, to use this tests for checking equality of variances among the treatment groups. The Levene's,Cochran's, and Hartley's statistical tests are also used to test for homogeneity of variance.

Bonferroni Test

It is a multiple comparison test of significance based on individual p-value is derived [10]. It can be used to correct any set of p-values for multiple comparisons, and is not restricted to use as a test to ANOVA. It works like as (1) compute a p-value for each comparison. Do no corrections for multiple comparisons when you do this calculation. (2) Define the familywise significance threshold. Often this value is kept set to the traditional value of 0.05. (3) Divide the value you chose in step 2 by the number of comparisons you are making in this family of comparisons. If you use the traditional 0.05 definition of significance, and are making 20 comparisons, then the new threshold is 0.05/20, or 0.0025. (4) Call each comparison "statistically significant" if the p- value from step 1 is less than or equal to the value computed in step 3. Otherwise, declare that comparison to not be statistically significant.

Holm's Test

The Holm test is a powerful and versatile multiple comparison test. It can be used in clinical research to compare all pairs of means, compare each group mean to a control mean, or compare preselected pairs of means. It is not restricted to being used as a follow up to ANOVA but instead it can be used in any multiple comparisons context [11].

Newman-Keuls Test

Newman-Keuls Test also referred to as the "Student Newman-Keuls Test". It is described variously as a stepwise or multiple-stage test. The range statistic varies for each pairwise comparison as a function of the number of group means in between the two being compared. A different shortest significant range is computed for each pairwise comparison of means. Means are first ordered by rank, and the largest and smallest means are tested. If there is no significant differences, testing stops there and it is concluded that none is significantly different. Then means of the next greatest difference are tested using a different shortest significant range. Testing is continued until no further significant differences are found.

This tests used when the group sample sizes are equal. For example, to test with 5 treatment means X5 > X1, p-value < 0.05. X4 = X1, p-value = ns. Can't test different between X1 and X3, X1 and X2, or X2 and X3. Can test different between X2 and X5 if the different between the means exceeds the difference between the means of X1 and X5. The Student-Newman-Keuls (SNK) test is more powerful than Tukey's method, so it will detect real differences more frequently [12]. However, Newman-Keuls test offers poor protection against a type I error. This is especially the case when treatment means fall into groups which are themselves widely spaced apart. Differences between means within groups will be significant more often than they should be at the specified level of α.

Tukey Multiple Comparison Test

In clinical research, the researcher may still need to understand subgroup differences among the different experimental and control groups. The subgroup differences are called "pairwise" differences. ANOVA does not provide tests of pairwise differences when the researcher needs to test pairwise differences. Tukey's multiple comparison analysis method tests each experimental group against each control group [13]. The Tukey method is preferred if there are equal group sizes among the experimental and control groups. A modified Tukey-Kramer method can be applied for comparisons of unequal-sized groups. We can use this test under assuming the observations being tested are independent within and among the groups, the groups associated with each mean in the test are normally distributed and there is equal within-group variance across the groups associated with each mean in the test (homogeneity of variance). Example in clinical research, consider the data on effect of maternal smoking on child birth weight, in this case only the effect of duration of smoking is statistically significant. To find which duration or durations are making a significant impact, compare mean birth weight for different duration.

Scheffe Test

The Scheffe Test (also called Scheffe's procedure or Scheffe's method) is a multiple comparison test used in Analysis of Variance [14]. In clinical research, researchers used to ANOVA and got a significant F-statistic (i.e. rejected the null hypothesis that the means are the same), then Sheffe's test to find out which pairs of means are significant. The Scheffe test corrects alpha (level of significant) for simple and complex mean comparisons. Complex mean comparisons involve comparing more than one pair of means simultaneously. For example, suppose four different antibiotics were tested for mortality rates among patients with necrotizing fasciitis. All the ANOVA can determine is if there were significant differences among the groups' mortality rates. It cannot identify which drug produced the lowest mortality rates, or if two or three of the drugs were equivalent in effectiveness and one was ineffective. The Scheffee method provides that detailed information about each drug.

Dunnett's test

The Tukey-Kramer method has wide appeal for all pairwise comparisons, Dunnett's test is the preferred method if the goal is to maintain the overall significance level when performing multiple tests to compare a set of treatment means with a control group. Dunnett developed this method of multiple comparisons for obtaining a set of simultaneous confidence intervals for preplanned treatment versus control contrasts $t_i - t_1$ (i= 2,…,v) where level 1 corresponds to the control treatment [15]. In additions, the Dunnett test is quite useful in clinical research when the researcher wishes to test two or more experimental groups against a single control group [16]. It tests each experimental group's mean against the control group mean. The other methods test each study group against the total group mean (i.e., the grand mean). This difference in testing approach makes the Dunnett method much more likely to find a significant difference because the grand mean includes all group means and thus mathematically it is less extreme than individual group means. The more extreme group means will produce larger mean differences than tests comparing one group mean to the grand mean.

Repeated Measurement of ANOVA

In a Clinical Research we record the data on the patients more than two times. In such a situation using the standard ANOVA procedures is not appropriate as it does not consider dependencies between observations within subjects in the analysis. To deal with such types of study data Repeated Measure ANOVA should be used [17]. We can use this method under the assumptions, (1) the dependent variable should be measured at the continuous level (i.e. measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. (2) The independent variable should consist of at least two categorical, "related groups" or "matched pairs". "Related groups" indicates that the same subjects are present in both groups. (3) The distribution of the dependent variable in the two or more related groups should be approximately normally distributed. (4) The variances of the differences between all combinations of related groups must be equal and there should be no significant outliers in the related groups. Example in clinical research, consider the two groups with two different treatment modalities with measured different physical and biochemical parameters (e.g pulse, systolic blood pressure, serum sodium level etc.) in each group at different time intervals (say pre-intervention, after 1 month and after two months) and to test the effect of each treatment modality on these parameters over time and at the same time look for any significant difference existing between the two groups using repeated measurement of ANOVA.

Repeated Measurement of ANCOVA

It is used in randomized clinical research, suppose measurements are often collected on each patient at a baseline visit and several post-randomization time points. In the longitudinal analysis of covariance in which the post baseline values form the response vector and the baseline value is treated as a covariate can be used to evaluate the treatment differences at the post baseline time points. A constrained longitudinal data analysis in which the baseline value is included in the response vector together with the post base line values and a constraint of a common baseline mean across treatment groups is imposed on the model as a result of randomization [18]. If the baseline value is subject to missingness, the constrained longitudinal data analysis is shown to be more efficient for estimating the treatment differences at post baseline time points than the longitudinal analysis of covariance. The efficiency gain increases with the number of subjects missing baseline and the number of subjects missing all post baseline values, and, for the pre-post design, decreases with the absolute correlation between baseline and post baseline values.

Pearson Correlation Test

Pearson Correlation is a statistical procedure applied to calculate association between two continuous or ordinal scale variables. It is used when both variables being studied are normally distributed. This coefficient is affected by extreme values, which may exaggerate or dampen the strength of relationship, and is therefore inappropriate when either or both variables are not normally distributed. Pearson's coefficient test while the significance of the coefficient is expressed by p-value. Pearson's correlation is denoted by a small letter 'r' and its values may range from -1 to +1. The value of the correlation coefficient from 0 to 1 is positive correlation and it designates proportional growth of values in both variables. An example of positive correlation is the duration of diabetes mellitus and the degree of damage of eye capillaries. The correlation coefficient value from 0 to -1 indicates negative correlation, i.e. a rise in the value of one variable that is proportional to a decline in the value of the other; e.g. oxygen concentration in the air drops with the rise in altitude above sea level. Perfect correlations, i.e. the values of the coefficient of correlation $r = \pm 1$ are not characteristically for biological systems and most frequently refer to theoretical models. The zero value of the coefficient of correlation indicates absence of linear correlation, i.e. by knowing the values of one variable, we can conclude nothing on the values of the other.

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 research, 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.

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).

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).

Fisher Exact Test

The Fisher's exact test is used in the approximation of the chi-squared and normal test for a 2 x 2 contingency table, when cells have an expected frequency of five or less [19]. The chi-square test assumes that each cell has an expected frequency of five or more, but the Fisher's exact test has no such assumption and can be used regardless of how small the expected frequency is. For example in clinical research, a study to compare two treatment regimes for controlling bleeding in haemophiliacs undergoing surgery when cell frequency of 2 x 2 contingency table is five or less [20].

G–test of independence

G–test of independence used when researcher has two nominal variables, each with two or more possible values and researcher want to see whether the proportions of one variable are different for different values of the other variable. For example, suppose researcher wanted to know whether it is better to give the diphtheria, tetanus and pertussis (DTaP) vaccine in either the thigh or the arm, so they collected data on severe reactions to this vaccine in children aged 3 to 6 years old. One nominal variable is severe reaction vs. no severe reaction; the other nominal variable is thigh vs. arm [21]. In this case, a higher proportion of severe reactions in children vaccinated in the arm; a *G*–test of independence will tell whether a difference this big is likely to have occurred by chance. Fisher's exact test is more accurate than the *G*–test of independence when the expected numbers are small.

Binomial Test

It is used for testing whether a proportion from a single dichotomous variable is equal to a presumed population value. Binomial test as an alternative to the z -test for population proportions. The assumptions for the test are that a) the data are dichotomous, b) observations should be independent from each other, and c) the total number of observations in category A multiplied by the total number of observations (i.e. A + B) > 10, and that the total number of observations in category B multiplied by the total number of observations > 10 (this way we can use the normal approximation for the binomial test and calculate the z-score). In clinical research, a common use of the binomial test is for estimating a response rate, p, using the number of patients (X) who respond to an investigative treatment out of a total of n studied.

McNemar test

In clinical research, 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").

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 [22, 23]. In clinical research, 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 [24]. 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 [25]. Bhapkar test is a more powerful alternative to the Stuart-Maxwell test. In large sample both will produce the same chi-squared value [24].

Cochran's Q **test**

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 and forty-six subjects receives drugs A, B, and C [26]. The response to each drug is either favorable or unfavorable and to test that differences of favorable response for the three drugs.

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 [27]. 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

Cronbach's α (alpha) Statistic

The Cronbach's alpha is a statistic for investigating the internal consistency of a questionnaire [28, 29]. 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 research 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).

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

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 research. 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 research, 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 research, this test find out the improvement of the drug(s) among the patients follow ups for a particular disease.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov test is a nonparametric statistical test that compares the cumulative distributions of two data sets. It does not assume that data are sampled from Gaussian distributions (or any other defined distributions). This test (K-S test) is used to decide if a sample comes from a population with a completely specified continuous distribution and also assumed that the population distribution is fully specified (i.e. it assumes that you know the mean and Standard deviation (SD) of the overall population perhaps from prior work) [30, 31]. For example in clinical research, to compare the serum Antioxidant levels in 30 patients with pemphigus vulgaris, an auto-immune blistering disorder [32].

Spearman Correlation Test

Spearman correlation to test the association between two ranked variables, or one ranked variable and one measurement variable. It is appropriate when one or both variables are skewed or ordinal [33] and is robust when extreme values are present. It is used instead of linear regression/correlation for two measurement variables if you're worried about non-normality, but this is not usually necessary. Spearman correlation coefficient solely tests for monotonous relationships for at least ordinally scaled parameters. The advantages of the latter are its robustness to outliers and skew distributions. Correlation coefficients measure the strength of association and can have values between –1 and +1. The closer they are to 1, the stronger is the association. A test variable and a statistical test can be constructed from the correlation coefficient. The null hypothesis to be tested is then that there is no linear (or monotonous) correlation.

Cochran Armitage trend test

In clinical research, 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 is used to test for trend among binomial proportions across levels of a single factor or covariate [34, 35]. 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.

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 [26]. 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.

Peto log-rank or Peto's generalized wilcoxon test

This test give more weight to the initial interval of the study where there are the largest number of patient's risk. If the rate of death is similar over time, the Peto log-rank test and log-rank test will produce the similar results. Log-Rank test is more appropriate than the Peto generalized Wilcoxon test when the alternative hypothesis is that the risk of death for an individual in one group is proportional to the risk at that time for a similar individual in the other group. In additions, the validity of this proportional risk assumption can be elucidated by the survivor functions of both groups. If it is clear they do not cross each other than the proportional risk assumption is quite probably true and then Log-rank test should be used. In other case, the Peto log-rank test used instead.

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 research, 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.

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.

Simpson's Paradox

Simpson's paradox, also known as Yule–Simpson effect was first described by Yule [36] and is named after Simpson's [37]. In clinical research, 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 [38].

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 [34]. 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 [39].

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 DM [40].

3. Conclusion

Statistical test are used to analyze the different type of data in different situations and nature of the data set. The statistical test has its limitations, and to overcome that another method is used. Before using the statistical test in clinical research 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 research, to make the decision on the objectives. Researchers / Physicians are helpful to used statistical tests to determine results from experiments, clinical research 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.

4. References

1. Wang D, Bakhai A. Clinical Trials-A Practical Guide to Design, Analysis, and Reporting. Remedica Publishing, USA. 2006.
2. Campbell MJ. Statistics at Square Two (2nd Ed.). Blackwell, USA. 2006.
3. Box JF, Guinness, Gosset, Fisher, and Small Samples. Statistical Science. 1987; 2(1):45-52.
4. Walker GA, Shostak J. Common Statistical Methods for Clinical Research with SAS® Examples (3rd Ed.). SAS Publishing, USA. 2010.
5. Mahajan BK. Methods in biostatistics for medical students and research workers (7th Ed.). Jaypee, India, 2010.
6. Hotelling H. The generalization of Student's ratio. Ann Math Stat. 1931; 2(3):360-378.
7. Scheffé H. The Analysis of Variance (Classics Ed.). John Wiley & Sons, USA, 1999.
8. Chow SC, Liu JP. Design and analysis of clinical trials: Concepts and Methodologies (2nd Ed.). John Wiley & Sons, New Jersey, 2004.
9. Bartlett MS. Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London Series A. 1937; 160:268-282.
10. Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika. 1988; 75(4):800-802.
11. Motulsky H. Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking (2nd Ed.). New York, NY: Oxford University Press. 2010.

12. Herve Abdi, Lynne JW. Newman-Keuls Test and Tukey Test. Neil Salkind (Ed.), Encyclopedia of Research Design. Thousand Oaks, CA: Sage. 2010.
13. Mary L. McHugh. Multiple comparison analysis testing in ANOVA. Biochemia Medica. 2011; 21(3):203-9.
14. Scheffé H. The Analysis of Variance. New York: Wiley. 1959.
15. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association. 1955; 50:1096-1121.
16. Dunnett CW. New tables for multiple comparisons with a control. Biometrics. 1964; 20:482-491.
17. Singh V, Rana RK, Singhal R. Analysis of repeated measurement data in the clinical trials. Jounranl of Ayurveda and Integrative Medicine. 2013; 4(2):77-81.
18. Liang KY, Zeger S. Longitudinal Data Analysis of Continuous and Discrete Responses for Pre-Post Designs. The Indian Journal of Statistics. 2000; 62:134-148.
19. Fisher RA. Statistical methods for research workers. Genesis Publishing Pvt Ltd. 1925.
20. Sarmukaddan SB. Clinical Biostatistics (1st Ed.). New Age International, India. 2014.
21. Jackson LA, Peterson D, Nelson JC, *et al*. (13 co-authors). Vaccination site and risk of local reactions in children one through six years of age. Pediatrics. 2013; 131:283-289.
22. Stuart A. A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification. Biometrika. 1955; 42:412-416.
23. Maxwell AE. Comparing the classification of subjects by two independent judges. British Journal of Psychiatry. 1970; 116:651-655.
24. 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.
25. Keefe TJ. On the relationship between two tests for homogeneity of the marginal distributions in a two-way classification. Biometrics. 1982; 69:683-684.
26. Agresti A. Categorical Data Analysis (2nd Ed.). John Wiley & Sons, New Jersey. 2002.
27. Bland JM, Altman DG. statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; i:307-10.
28. Cronbach LJ. Coefficient alpha and the internal structure of test. Psychometrika. 1951; 16:297-334.
29. Bland JM, Altman DG. Statistics notes: Cronbach's alpha, British Medical Journal. 1997; 314:572.
30. Lilliefors H. On the Kolmogorov Smirnov test for normality with mean and variance unknown. JASA. 1967; 62:399:402.
31. Sprent P, Smeeton NC. Applied Nonparametric Statistical Methods (4th Ed.). Florida: Chapman and Hall/CRC. 2001.
32. Alireza AB, Shima Y, Sara J, Maryam Y, Farid Z, Farid AJ. How to test normality distribution for a variable: a real example and a simulation study. Journal of Paramedical Sciences (JPS). 2013; 4(1):73:77.
33. Altman DG. Practical Statistics for Medical Research. Chapman & Hall/CRC. 1990.
34. Armitage P. Tests for Linear Trends in Proportions and Frequencies, Biometrics. 1955; 11:375-386.
35. Cochran WG. Some Methods for Strengthening the Common Chi-Square Tests. Biometrics. 1954; 10:417-51.
36. Yule G. Notes on the theory of association of attributes of statistics. Biometrika. 1903; 2:121-134.
37. Simpson EH. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society, Series B. 1951; 13:238-241.
38. Julious SA, Mullee MA. Confounding and Simpson's paradox. British Medical Journal. 1994; 309:1480-1481.
39. Duffy JC. Alcohol consumption and all-cause mortality, International Journal of Epidemiology. 1995; 24(1):100-5.
40. Zucker DM. Permutation Tests in Clinical Trials. Wiley Encyclopedia of Clinical Trials, 2007. (http://pluto.mscc.huji.ac.il/~mszucker/DESIGN/perm.pdf).