COMMENTARY

General Concepts of Group Comparisons for Clinical Practitioners

Chandraketu Singh[1] and M Srivastava[2]

ABSTRACT

A researcher always looks for favourable response to the intervention he/she applies. This favourable response if often understood as statistical significance. Obtaining p-value of significance is one of the backbones of standard statistical methodology. People who read scientific articles must be familiar with the interpretation of p-values while assessing statistical findings. There are separate statistical methods of significance for both continuous and discrete responses. When variables follow the assumptions of normality, homogeneity and independence then parametric class of tests are applied otherwise nonparametric methods are used. If the outcome of response is binary (Yes/No) type, separate method of testing of proportions is applied. Hypothesis testing using a p-value is a binary decision (accept or reject hypothesis). Some of these simple methods have been discussed for medical researchers and practitioners.

Keywords: Group comparison, hypothesis testing, test of significance

INTRODUCTION

Research as it was defined by Leedy and Ormond[1] is a systematic process of collecting and analyzing information to increase our understanding of the phenomenon under study. Misuse of statistics in published articles is often reported.[2] Kim *et al.*[3] reported various types of shortcomings in the reported results and analysis of data in dental journal. Interested readers may also read the following articles.[4,5] Though several difficulties are reported on the use of statistical methods, but our focus in this article is limited to group comparisons only. These comparisons may arise from simple designs of two groups.[6]

The phrase "test of significance" was introduced by R. A. Fisher. The statistical significance refers to whether any difference observed between groups being studied are "real" or whether they are simply due to chance.[7] The question of statistical significance is most important and is the commonly used statistics in pharmaceutical and clinical studies. It is one of the backbones of standard statistical methodology. This helps to make familiar statement, which often opens many regulatory doors.

There are many sources of errors to be controlled, for example, sampling error, researcher bias, problems with reliability and validity, simple mistakes, etc. The random errors cannot be controlled. Tests for statistical significance are used to address the question: what is the probability that we think is a difference between two groups is really just a chance occurrence? One can never be 100% sure that a difference exists between two groups. These groups may be patients participating in a clinical trial, different doses of the same drug, before or after of an intervention or treatment etc. In fact, tests for statistical significance tell us what the probability is that the difference we have found (we think) is due only to random chance.

The knowledge of probability theory and the normal curve, helps us to estimate the probability of being wrong, if we assume that our finding 'a difference' is true. If the probability of being wrong is small, then we say that our observation of the difference is a statistically significant. So a statistical significance tells, with what probability one would be make an error in his decision if we have found that a difference exists.

The mean or median are most widely used measures of the 'location'. The standard deviation (SD) is a measure of the 'spread or scatterness' of the data. SD is also known as scale

Chandraketu Singh is a graduate from Lucknow Christian College, Lucknow. He did his M.Sc. in Applied Statistics from Babasaheb Bhimrao Ambedkar University (Central University), Lucknow in the year 2013. Currently, he is engaged as Academic Associate (Post Graduate Programmes), Indian Institute of Management Lucknow, Prabandh Nagar, Off-Sitapur Road, Lucknow – 226013. INDIA.

[1]Academic Associate, IIM Lucknow, [2]Principal Technical Officer, Lucknow (UP)-226031.

Address for Correspondence:

Dr. M. Srivastava, 25/14, Indira Nagar,
Lucknow-226016, India

Contact: 0522-27060010

E-mail: mukeshlko@yahoo.com

Date of Submission	:	26-08-2014
Reviews Completed	:	10-09-2014
Date of Acceptance	:	19-09-2014

parameter. Thus, Mean±SD is the simple summary in which continuous data is usually expressed. Once mean and SD of a variable are known and it follows a normal distribution[8] then everything about that variable can be understood.

Statistical testing starts off by assuming something impossible: that the two groups of people (say; systolic blood pressure, SBP) were exactly alike at the start of study. This means that on an average the starting SBP in each group was same. Mathematical procedures are then used to examine differences in outcomes (reduction of SBP) between the groups after treatment. The goal is to determine how likely it is that the observed difference (change) might have occurred by chance alone.

Plenty of introductory level biostatistics text are available.[9,10] But with the increased computational power due to recent statistical packages more analysis can be easily performed on the same set of data. Since computer is an obedient servant it will display results to whatever been asked, it should be used after understanding concepts of statistics. Students and faculty have to use statistics for practical, projects and paper publications. Often users get confused in interpreting results of their study with the output they have obtained. What's important at this stage is to know the right statistical test to be used to interpret the outcome with clarity.

Most frequently performed activity by researchers is the group comparison.[11] The activity of groups comparisons can be done in several ways. It depends upon the study design and the type of variable.[12] With regard to the type of variable, which are of four types; interval, ratio, ordinal and nominal type variables,[1,9] different statistical test should be preferred.

One may require to compare the mean of study sample with the mean of population. The study may be done to see the therapeutic effect of drug. This can be performed in three ways. If the design of study is such that observations after treatment are taken on the same individuals on whom it was taken before the treatment, then paired t-test will be performed. Zhi ei studied- Paired t-test was used to assess the changes in LAC and BMD values within each group.[13] If individuals in two groups are different then student's t-test is undertaken.[9] If the sample mean has to be compared some fixed value, mean of the population then one sample test is used.[9] It is expected that readers will use some statistical software for analysis hence use of mathematical expressions has been kept at minimum.

Hypothesis testing

Why is hypothesis testing so important? Hypothesis testing is a process of testing the significance regarding a population parameter on the basis of sample observations. It provides an objective framework for making decisions using probabilistic methods, rather than relying on subjective impressions. The hypothesis being tested is called null hypothesis. In order to test a null hypothesis, a statistical test of difference is required. There are many tests that all seem to answer the same type of question but each is appropriate when certain types of data are being considered. Any hypothesis complementary to the null hypothesis is called an alternative hypothesis.

Suppose a scientist is testing the effect of a drug on the response by injecting 50 rats with a unit of the drug, and recorded the area under the curve (AUC) of serum glucose profile recorded every 30 minutes for 120 minutes. The scientist knows that the mean response (AUC) for rats not injected with drug was 17085 mg.M/dl. The mean of 50 rats response was 16479 mg.M/dl with sample SD of 1414 mg.M/dl. Does the drug have an effect on response time? We can set two hypothesis for this. The first one is Null hypothesis (H_0), i.e the drug has no effect in lowering the serum glucose level. The H_0 is always in the null form. This means that the AUC is going to be around 17085 mg. M/dl even if drug is taken i.e. there is no effect of the drug. The second hypothesis, alternate hypothesis (H_1), could be that the drug at least do something or the drug has an effect. In other words, the mean is not equal to 17085 mg. M/dl when the drug was given. The question is, how do we know that whether we accept the alternate hypothesis or null hypothesis? To begin with let us assume that the null hypothesis is true. If this is the situation, what is the probability that we will get these results in the sample? If the probability of getting such results is really very small then we will reject the null hypothesis and we would say that the alternate hypothesis is true. If the null hypothesis is true then the sample mean would be population mean i.e. the drug treated group would have an equal AUC as with untreated group.

Categorically speaking, almost every time a statistical test is carried out it is testing the probability that the null hypothesis is correct. If the probability is small ($p<0.05$) then the hypothesis is deemed to be untrue and it is rejected in favour of the alternative. Suppose we want to see the effect of a new drug on tooth pain then the null hypothesis would be "the new drug would not reduce the pain". It means that according to null hypothesis, the drug will not be able to control the tooth pain. In case the drug is effective and is able to reduce the pain then this phenomenon is against the null hypothesis and the null hypothesis is liable to be rejected. The null hypothesis is never said to be true or false. It is always either accepted or rejected. Once the null hypothesis is rejected then the alternative hypothesis will be accepted and the decision would be that 'the drug is effective in controlling the tooth pain'. The alternative hypothesis includes values not specified in the null hypothesis.

Type I and II errors

A group of men who have died from heart disease within the past year were identified, and the cholesterol levels of their offspring were measured. Suppose the "average" cholesterol level in the familial aggregation of cardiovascular risk factors study in children was 190 mg/dL. Two hypotheses are considered: (1) The average cholesterol level of these children was 190 mg/dl; (2) The average cholesterol level of these children was >190 mg/dL.

What are the type I and type II errors as per this example? The type I error is the probability of deciding that offspring of men who died from heart disease have an average cholesterol level higher than 190 mg/dL when in fact their average cholesterol level was 190 mg/dL. The type II error is the probability of deciding that the offspring have normal cholesterol levels when in fact their cholesterol levels were above average (i.e. >190 mg/dL).

The concept of type I and type II errors can also be understood in the following way. Suppose we have drawn two samples from a healthy population and measured their systolic blood pressure (SBP). Ideally there should be no difference (the null hypothesis) in the mean SBP of two groups. Upon testing if it was found that the means are statistically equal. The null hypothesis will be accepted because it was considered to be equal. One more sample from the same population was taken. If it was found that the SBP of two groups was significantly different. Then the inference was against the reality and the null hypothesis will be rejected (because the samples drawn from same population should be equal). This type of reversing inference in rejecting the null hypothesis is Type I error (α). Ideally this kind of extreme individuals (group) should not be frequently present in the homogeneous population. Conventionally one such rejection in 20 samples (i.e. 5%) has been considered significant. So, in a Type I error the null hypothesis is really true but the statistical test has led us to believe that it is false. Statistically, Type I error is the probability of rejection of null hypothesis when it is true.

Now, suppose the SBP of a hypertensive and healthy group is studied and the statistical test fails to pick up the difference between two groups (may be due to large SD). We accept the null hypothesis. Our decision for accepting the H_0 is against the norm of study i.e. i.e. the SBP of hypertensive and healthy groups are bound to be different. Thus Type II error (β) is the probability of accepting the null hypothesis when it is false. The beta (β) error is the probability of accepting H_0 (no treatment difference) when, in fact, some specified difference included in H_1 is the true difference.

There are two ways in which a wrong inference can be made from the test. This give rise to the two types of errors. The endeavour should be to keep both type of errors low. Unfortunately, we work on a sample which is considered to be the best representative of the population under study. Due to fixed sample size, both types of errors cannot be controlled simultaneously. The statistical test only gives an indication of how likely it is that the null hypothesis is true. This inference is based upon the information available in the sample. A good sample will provide us a reliable result. Using a lower critical p-value will increase the chance of making a type II error. Choosing a higher critical p-value will increase the chance of making a type I error. It is only a convention to use $p < 0.05$ as type I error. The type I error is potentially dangerous and could be seen as a false positive.

Level of significance

We keep the probability of committing of type I error at certain minimum level, called level of significance (α). It is also known as the region of rejection. The level of significance is pre-specified while applying a test. The level of significance is defined as the probability that the statistical test results in a decision to reject H_0 (a significant difference) when, in fact, the treatments do not differ (H_0 is true). By definition, the level of significance represents the chance of making a mistake when deciding to reject the null hypothesis. If the statistical test results in rejection of the null hypothesis, we say that the difference is significant at the α level. If α is chosen to be 0.05, the difference is significant at the 5% level. A significant difference is often expressed, equivalently, as $p< 0.05$. If two means are significantly different.

P-value: It is important to understand here is that p-value is always for testing the null hypothesis. The term "p" used to describe the probability of observing such a large difference purely by chance in two groups is known as the p-value. The p-value for any hypothesis test is the α-level at which we would be indifferent between accepting or rejecting H_0 given the sample data at hand. That is, the p-value is the α-level at which the given value of the test statistic (such as t) is on the borderline between the acceptance and rejection regions.

We always look for a low p-value to reject the hypothesis and accept the more interesting alternative hypothesis. In biology it is usual to take a value of 0.05 or 5% as the critical level for the rejection of a null hypothesis. If it is unlikely enough that the difference in outcomes occurred by chance alone, the difference is pronounced statistically significant. Rejecting the null hypothesis simply means there is less than one in 20 chance of it being true (and we reject H_0). The smaller the p-value the more confident we are in the conclusions drawn from it. A p-value of 0.001 indicates that if the null hypothesis is true, the chance of seeing data as extreme or more extreme than that being tested is one in 1000. This is much more convincing than a marginal p=0.049. If the results yield a p-value of 0.05, here is what the scientists are saying: "Assuming the two groups of people being compared were exactly the same from the start, there's a very good chance, say 95 per cent, that the 11 mmHg difference in SBP would not be observed if the drug had no benefit whatsoever." From this finding, scientists would infer that the SBP reducing drug is indeed effective.

One Tailed or Two Tailed Test

When the interest is on just to see an effect whether the level is increased or decreased then it is a two tailed test. But suppose we are of the view that the drug has a lowering effect. In this situation the null hypothesis remains the same but the alternative hypothesis changes i.e. H_1: Drug lowers the serum glucose level. Our interest lies in to know what is the probability of lowering the sample mean or more extreme value (a value lower than the sample mean)? If a null hypothesis is true, from the example above the probability of getting a result more extreme than 16479 mg. M/dl will be (0.0027/2=0.00135). The calculated p-value (0.00135) is smaller than 0.05, therefore, it is very less likely to happen so the null hypothesis will be rejected. Thus a one-sided test allows for the possibility of a difference in only one direction.

Comparison with known Mean (μ_0): Here one tries to test the sample mean with some prespecified mean (value) because data often come from a single population. A comparison of the sample mean to some hypothetical or standard or known or fixed value is desired. Suppose we are interested in checking whether the population mean μ is equal to some prespecified value, say μ_0. This problem can be formulated as a two-sided hypothesis test. It means that μ can be smaller or greater than μ_0 The null hypothesis, or the hypothesis under test, is $H_0 : \mu = \mu_0$, whereas the alternative hypothesis is $H_1 : \mu \neq \mu_0$. The one-sample t-test is the most commonly used statistic when one wants to compare the mean to a constant value. If n<30 then the statistic will follow normalised t-distribution. If the sample size is more than 30 then

$$Z = \frac{\bar{x} - \mu}{SE}$$

follows approximately normal distribution. One can also use a one-sample Z-test if the population standard deviation is known and do not need to estimate it based on the sample data. From the example above, we are trying to get a p-value at this extreme of sample mean (16479 mg. M/dl) of the treated group.

With the given SD of 1414 mg. M/dl of the sample, the standard error (SE) =202. Hence, the standard deviation of our sampling distribution is SD/√n = 202. We approximate the population standard deviation with the sample SD. Our interest might be to know 'what is the probability of getting the mean 16479 mg. M/dl in our sample'. How many SD away the drug treated group mean is away from the population (untreated group) mean. i.e.

$$Z = \frac{\bar{x} - \mu}{202}$$

Z=(17085-16479)/202 = 3.0 .

Thus the drug treated group mean is 3 SD below (away) from the population mean.

What is the probability of getting this extreme or below by chance is decided from standard normal distribution table. Depending upon the objective, the result can either be on the lower side or the upper side of the extreme. Empirically it is known that 99.73% area under normal distribution is covered under Mean±3SD. Only 0.27% area remains on both sides of extreme. The probability that the difference is within 3 SD is 0.9973 (99.73%). If the null hypothesis is true then it is only 27 out of 10000 chances that we would have this extreme result or more. So under the circumstances we must reject the null hypothesis in favour of alternative hypothesis. The one-sample Z-test is more powerful than the one-sample t-test.[14]

Paired t-test: When a single individual is tested twice (e.g. before and after intervention), then the paired comparison is done.[17] Here the data should be continuous and, at least approximately, normally distributed. The variances of the two sets should be homogeneous. The homogeneity of variance can be tested by the Bartlett test or Levene's test.[9,10,21] The null hypothesis is that the there is no difference between the two samples i.e. the differences are not significantly different from zero (H_0: μ_1-μ_2=0; or μ_1=μ_2). If the interest is to test a difference other than zero, then a single tailed test should be preferred (i.e. we are interested in the hypothesis that 'before' is smaller than 'after', with the null hypothesis that 'before' is not greater than 'after').

Parametric Method (t-test)

Two groups are obvious in the study; for example males and females, healthy and diseased, control and drug treated etc. in research studies. The comparison of means of two groups is often required to ascertain the real difference or the effect of intervention. t-test is used to compare means of two different set of values. t-test is often called *Student's t-test* in the name of its founder "Student". It is generally performed on a small set of data. This comparison is admissible when the observations are coming from a normally distributed population, observations are independent and the variances of the groups are equal.[9,15] The degrees of freedom against which the critical t-value (t_{crit}) at desired level of significance (say, 0.05) can be had from n_1+n_2-2. Adebayo[16] studied, the values for fracture toughness and compared them using independent samples t-tests at α= 0.05.

A normally distributed variable can be completely characterised by its mean and SD. The two means in can be compared only if their SDs are equal. The t-test uses means and standard deviations of two samples to make a comparison. Before actually performing the independent group t-test, a statistical pre-test is often performed to verify the hypothesis that the variances are equal. Depending upon the interest in studying the drug/intervention effect, the t-statistic for one or two tailed is referred. Reject H_0 in favour of H_1 (i.e. decide that H_0 is false, based on the data) if $t_{obs} > t_{crit}$ for the given degrees of freedom; else, do not reject H_0.

If the variances of the two groups are equal, then the formula for t-test is given as

$$t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

S is the pooled standard deviation of two samples which is calculated as follows;

$$S = \sqrt{\frac{(n_1-1)S_1{}^2 + (n_2-1)S_2{}^2}{n_1+n_2-2}}$$

Where $\overline{x_1}\ and\ \overline{x_2}$ are means of two groups, with SD, S_1 and S_2 and sample size n_1 and n_2 respectively?

Unequal variance (Welch test) If the variances of the two groups are not equal then apply Welch's t-test.[9,15]

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})}}$$

The Welch test uses the Satterwaite-Welch adjustment for the degrees of freedom:

$$v \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(\frac{S_1^2}{n_1})^2}{(n_1-1)} + \frac{(\frac{S_2^2}{n_2})^2}{(n_2-1)}}$$

Standard t-distibution table can be used for critical value. Menicucci *et al.*[19] studied the difference in BIC rates was evaluated using t-test.

Test of proportions

Instead of comparing mean or median if the responses are in yes-no, responded-not responded; cured-not cured etc. (binomial response) then the groups can be compared by the test of proportions. On comparing the performance of two analgesics in tooth extraction if p_1 and p_2 were the response of two groups respectively, then comparison of groups is done Z test using following formula

$$Z = \frac{p_1 - p_2}{\sqrt{P.Q(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$P = \frac{r_1 + r_2}{n_1 + n_2}$$

$$Q = 1 - P$$

Where r_1 and r_2 are number of responses from the testing of n_1 and n_2 samples (cases) such that $p_j = r_j/n_j$; j=1,2.

Suppose a public health researcher wants to know how the complaint of caries differs in the percentage of students who are vegetarians (20/80) and non vegetarians (27/70). With P=0.2933 and Q=0.7067, Z=1.78. Since Z is less than 1.96 so the difference between the two groups is not significant. Here we can say that there is insufficient evidence to conclude that the food habit is a cause of dental caries.

Arrow[18] applied proportional t-test to study the efficacy of articaine 4% with 1:100,000 adrenaline (test) and lignocaine 2% with 1:80 000 adrenaline (control), delivered either through an inferior alveolar nerve block (IANB) or BI for routine restorative procedures in mandibular posterior teeth among children.

Non-parametric method - *Mann-Whitney test*

If the observations of an experiment recorded are not normally distributed then a different kind of comparison test might need to be employed – a nonparametric test.[22] In the case of independent groups, the nonparametric test usually performed is the Mann-Whitney U-test. For paired data that are not normally distributed, the Wilcoxon signed-rank test is usually performed.

The Mann-Whitney test, also called the Wilcoxon rank sum test. Use it when the data do not meet the requirements for a parametric test i.e. the Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed. The Mann-Whitney U-test is used to test whether two independent samples of observations are drawn from the same (or identical) distributions. This test is a non-parametric alternative to the t-test for independent samples. Unlike the t-test which compares mean values between two groups, Mann-Whitney U-Test compares their median. An advantage with this test is that the two samples under consideration may have different number of observations. In order to interpret the results properly one should ascertain that the shapes of distribution of the two independent groups are nearly similar.

The logic behind the Mann-Whitney test is to rank the data for each condition, and then see how different the two rank totals are. If there is a systematic difference between the two conditions, then most of the high ranks will belong to one condition and most of the low ranks will belong to the other one. The Mann-Whitney test statistic "U" reflects the difference between the two rank totals.

$$U_1 = n_1 * n_2 + \frac{n_1 * (n_1 - 1)}{2} - R_1$$

$$U_2 = n_1 * n_2 + \frac{n_2 * (n_2 - 1)}{2} - R_2$$

The smaller value of U_1 and U_2 is the one used when consulting significance tables which can be easily calculated using

$$U_1 + U_2 = n_1 * n_2$$

If this smaller value of U is higher (larger) than the characteristic value (the one obtained from statistical Table) then we say that the distributions are not significantly different ($p>0.05$). i.e. the two medians are not significantly different. If the smaller value of U is lower than the characteristic value then the distributions are not same (but different) and the two medians are significantly significant ($p<0.05$).

DISCUSSION

A good clinical research has been defined as the integration of the best evidence with clinical expertise. Statistical methods have a central role in the production and analysis of scientific data and drawing proper inferences.[23] Manoeuvring best evidence involves systematically collecting and analysing scientific evidence to answer a specific clinical question.[24]

Statistical significance means that there is a good chance that we are right in finding that a difference exists between two groups. But it is not the same as practical significance. We can have a statistically significant finding, but the implications of that finding may not be relevant while answering the query for which the study was conducted. For example, if while screening for antidiabetic activity the test compound showed significantly high glucose level than the control group. Statistical significance exists, but this result is of no relevance to a drug researcher who is looking for compound to lower the glucose levels. The researcher must always examine both the statistical and the practical significance of any research finding. This aspect of statistics is not only important to researchers in terms of applications to data analysis and interpretation, but is critical to an understanding of the statistical process.[8]

Generation of high quality, reliable and statistically sound data from experiments and clinical trials is necessary for reproducible results.[20] A careful planning accompanied with the knowledge of statistical tools helps in execution of experiments and decision making. The common choice of Type I error is 0.05 where as choice of type II error varies according to researcher. The performance of statistical test largely depends upon the sample size. The choice of optimal sample for study is necessary because of limitations of resources time and finance. Ioannidis *et al.*[25] studied the optimal choice of type I and type II errors which has been shown varying according to the available sample size and the plausible effect sizes.

Statistical tests are commonly seriously misinterpreted by non-statisticians, but the misinterpretations are very natural. One clarity required here is that what a statistical test determines is the probability that the null hypothesis is true (called the p-value). If the probability is low then the null hypothesis is rejected and the alternate hypothesis accepted.

Unpaired t-test has been the most frequently used statistical test for establishing significant difference between two groups by Medical researchers. Mao *et al.*[26] examined the effects of vitamin D deficiency on the impaired bone repair in streptozotocin (STZ)-induced diabetes by unpaired t-test using female C57BL6 mice fed by normal and vitamin D-deficient diet. de Andrade *et al.*[27] compared mean physical performance and cardiorespiratory responses in the six-minute walk test (6MWT) in asthmatic children with reference values for healthy children in the same age group using Student's t-test. The statistical conclusions drawn from applying t-test give direct information about the causative factor also. In the study by de Andrade *et al.*[27] asthmatic children's performance in the 6MWT evaluated through distance walked was significantly lower than the predicted values for healthy children of the same age, and was directly influenced by sedentary life style. Using t-test Al-Hana[28] found that nano-composite recorded the highest micro-shear bond strength close to that required to resist the polymerization contraction stress in their study of shearing load with tensile mode of force applied via materials. Kamatagi *et al.*[29] applied unpaired t-test to demonstrate significantly lower mean microleakage for immediate apical to coronal leakage for immediate (ICA) and delayed (DCA) techniques and not significant difference between apical to coronal (IAC) and Delayed apical to coronal (DAC) groups. After the comparison of two groups, the testing of the means of more than two groups simultaneously is done by *post hoc* tests.

Depending on the method chosen to adjust for baseline measures, varying inferential results has been shown by Carlsson *et al.*[30] They investigated the Type I error and statistical power of tests comparing treatment outcomes based on parametric and nonparametic methods. Martínez-Murcia *et al.*[31] applied Mann–Whitney U-Test for the selection of voxels in terms of their significance to be used in Factor Analysis to carry out the feature reduction step in early diagnosis with the help of Alzheimer's Disease Neuroimaging Initiative (ADNI).

REFERENCES

1. Leedy PD, Ormrod JE . Practical Research Planning and Design. 9th Ed. Pearson. 2010.
2. Hannigan A, Lynch CD. Statistical methodology in oral and dental research: Pitfalls and recommendations. J Dentistry 2013; 41: 385-92.
3. Kim JS, Kim D-K, Hong SJ. Assessment of errors and misused statistics in dental research. Int Dental J 2011; 61: 163–7.
4. Vahanikkila H, Nieminen P, Miettunen J, Larmas M. Use of statistical methods in dental research: comparison of four dental journals during a 10-year period. Acta Odontologica Scandinavica 2011; 67: 206–11.
5. Fernandes-Taylor S, Hyun JK, Reeder RN, Harris AHS. Common statistical and research design problems in manuscripts submitted to high impact medical journals. BMC Research Notes 2011; 4: 304.
6. Greenhalgh T. How to read a paper: Statistics for the non-statistician. II: "Significant" relations and their pitfalls. BMJ 1997; 315: 422
7. Sterne JAC, Smith GD. Sifting the evidence - what's wrong with significance tests?. BMJ 2001; 322: 226–31
8. Richard G. Brereton. The normal distribution. Journal of Chemometrics 2014. [Last access 25-06-2014. doi: 10.1002/cem.2655].
9. Zar, JH. *Biostatistical Analysis. 5th Edition.* Pearson Prentice-Hall, Upper Saddle River, NJ. 2010.
10. Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. John Wiley & Sons, 2008.
11. Neelakantan P, Grotra PD, Sharma S. Retreatability of 2 mineral trioxide aggregate-based root canal sealers: a cone-beam computed tomography analysis. J Endod 2013; 39: 893-96.
12. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy I: Medical. Stats in Med 1989; 8: 441-54.
13. Zhi QH, Lo ECM, Kwok ACY. An in vitro study of silver and fluoride ions on remineralization of demineralized enamel and dentine Aust Den J 2013; 58: 50–56.
14. DeCoster J. Testing Group Differences using T-tests, ANOVA, and Nonparametric Measures. Retrieved (month, day, and year you downloaded the notes, without the parentheses) (2006). [Available online: http://www.stathelp.com/notes.html]
15. Bolton S, Bon C. Pharmaceutical Statistics: Practical and Clinical Applications, Fifth Edition (Drugs and the Pharmaceutical Sciences). CRC Press; 4th edition October 17, 2003
16. Adebayo OA, Burrow MF, Tyas MJ. Relationship between composite fracture toughness and bond strengths to enamel and dentine. Aust Dent J 2012; 57: 319–24.
17. Pahwa N, Kumar A, Gupta S. Short term clinical e□ectiveness of a 0.07% cetylpyridinium chloride mouth rinse in patients undergoing fixed orthodontic appliance treatment. The Saudi Den J 2011; 23: 135–41
18. Arrow P. A comparison of articaine 4% and lignocaine 2% in block and infiltration analgesia in children. Aust Dent J 2012; 57: 325–33.
19. Menicucci G, Mussano F, Schierano G, Rizzati A, Aimetti M, Gassino G, *et al.* Healing properties of implants inserted concomitantly with anorganic bovine bone. A histomorphometric human study. Aust Dent J 2013; 58: 57–66.
20. Krishnankutty B, Bellary S, Naveen BRK, Moodahadu LS. Data management in clinical research: An overview. Indian J Pharmacol. 2012; 44: 168–72.
21. Garson, GD. Testing Statistical Assumptions. North Carolina State University. 2012. [Available online: http://www.statisticalassociates.com/assumptions.pdf]
22. Corder GW, Foreman DI. Nonparametric Statistics for Non Statisticians: A step-by-step approach. Wiley. 2009
23. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996; 313: 36.
24. Allan G. A critique of using grounded theory as a research method. Electronic Journal of Business Research Methods 2003; 2: 37-46.
25. Ioannidis JP, Hozo I, Djulbegovic B. Optimal type I and type II error pairs when the available sample size is fixed. J Clin Epid. 2013; 66: 903–10.
26. Mao L, Tamura Y, Kawao N, Okada K, Yano M, Okumoto K, Kaji H. Influence of diabetic state and vitamin D deficiency on bone repair in female mice. Bone 2014; 61: 102–8
27. de Andrade LB, Silva DARG, Salgado TLB, Figueroa JN, Silva NL, Britto MCA. Comparison of six-minute walk test in children with moderate/severe asthma with reference values for healthy children. J Pediatr (Rio J) 2014; 90:250–57
28. Abo Al-Hana DA, El-Messairy AA, Shohayb FH, Alhadainy HA. Micro-shear bond strength of different composites and glass-ionomers used to reinforce root dentin. Tanta Dent J 2013; 10: 58e66
29. Kamatagi L, Saler S, Rastogi A, Chhabra N. Permeability of remaining endodontic obturation: Comparison of immediate versus delayed post space preparation. An in vitro study. Asian J Oral Health Allied Sci 2013; 3: 8-13.
30. Carlsson MO, Zou KH, Yu CR, Liu K, Sun FW. A comparison of nonparametric and parametric methods to adjust for baseline measures. Contemporary Clinical Trials 2014; 37: 225-33.
31. Martínez-Murcia FJ, Górriz JM, Ramírez J, Puntonet CG, Salas-González D. Computer aided diagnosis tool for Alzheimer's Disease based on Mann–Whitney–Wilcoxon U-Test. Expert Systems with Applications, 2012; 39: 9676-85