UNDERSTANDING OF STATISTICS IN TESTING OF PHARMACOLOGICAL HYPOTHESIS

Anoop Kumar, Neelima Sharma and Dinakar Sasmal*

Department of Pharmaceutical Sciences and Technology, Birla Institute of Technology, Mesra, Ranchi -835215, Jharkhand, India. Article type: Review

Running title : Role of Statistics in Pharmacology.

* Corresponding author : Dr. Dinakar Sasmal. Department of Pharmaceutical Sciences and technology, Birla Institute of Technology, Mesra, Ranchi -835215, Jharkhand, India.

Email id: dsasmal@bitmesra.ac.in. Telephone No. +91 651 2275444/2275896.

ABSTRACT

Statistics is an important tool in pharmacological research to conduct hypothesis testing. It is used to summarize experimental data in terms of central tendency (mean, median or mode) and variance (standard deviation, standard error of the mean, confidence interval).The purpose of statistical analysis is to determine whether the observed differences between the treated and untreated animals/humans could have arisen by chance or by the drug? Now days, there has been a huge increase in the use of statistics in pharmacological (preclinical and clinical) research, especially after the availability of user friendly statistical software. The biostatical tools give meaning to raw data generated during the research studies. Results of various statistics tests help to draw a valid conclusion from the observations. However, the basic understanding of methodology (study design, sample size justification and correct use of sampling techniques) and bio statistical tests (parametric and non–parametric tests) is still lacking among pharmacologists. Therefore, it is essential for every pharmacologist to have an understanding of the correct uses of statistics. Thus, in this review, we tried to simplify the understanding of statistical tool in pharmacology field.

Keywords : Null hypothesis, Alternate hypothesis, Type 1 error, Type 2 error, p-value, Power of study.

INTRODUCTION

Pharmacology is the study of drugs and their actions on target sites in biological systems. Pharmacology includes pharmacokinetics (What does the body do with the drug?) and pharmacodynamics (What does the drug do with the body) [1]. The ultimate goal of pharmacology is the prevention and treatment of disease and illness. To achieve this goal, pharmacological experiments must be designed, performed, interpreted and reported correctly [2]. Pharmacology routinely employs statistics for collection, organization, analysis and interpretation of experimental data. The appropriate use of statistics is essential to ensure the best methods are used to collect data in an unbiased fashion and the analysis of the collected data is done in a proper way [3], but

unfortunately, most of times, data is analyzed inappropriately and switching between statistical tests is done by the researchers until they get the expected result. Thus, it is not surprising that there is a loss of faith in the literature. There has been growing concern that much of what is published in preclinical and clinical studies is misleading [4, 5, 6]. If you are confused about the most appropriate statistics for your experiment, you could simply talk to a statistician but it's not possible to take help of stastician in academic research. So, it's important for every pharmacologist to understand the basic concepts of statistics. Thus, the aim of this review is to understand the use of stastisics in the testing of pharmacological hypothesis.

GENERATION OF HYPOTHESIS

Hypotheses are generated on the basis of good research questions. Detailed knowledge about a subject may generate a number of questions. It then becomes necessary to ask whether these questions can be answered through one study or if more than one study is required. All questions, primary and secondary, should be developed at the beginning and planning stages of a study. The primary question should never compromise because it is the primary research question that forms the basis of the hypothesis and study objectives (Fig. 1) [7].

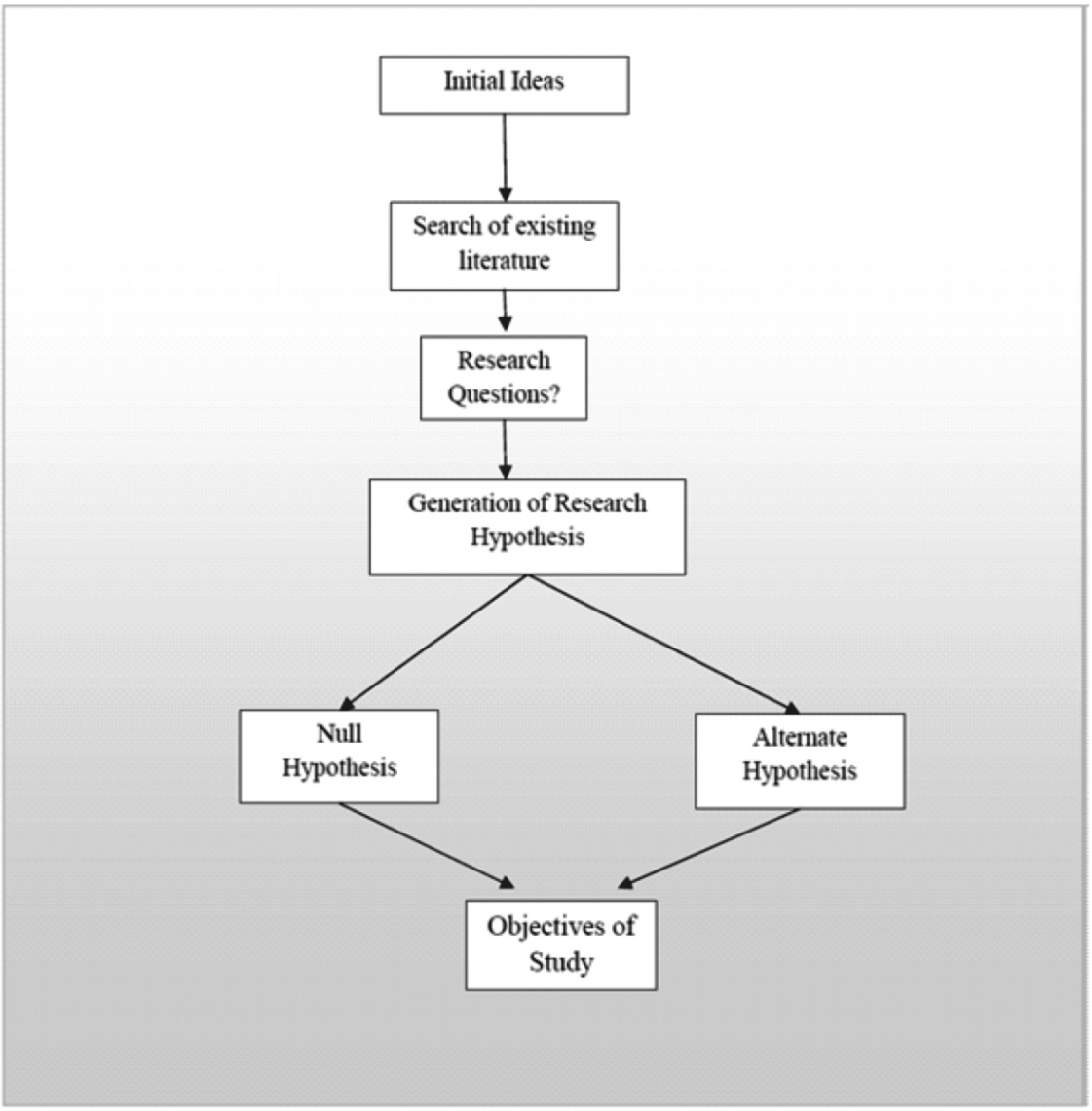

Figure 1: Layout of hypothesis and study objectives in statistics

TESTING OF HYPOTHESIS

The first step of hypothesis testing is to convert the research question into null and alternative hypotheses. The null hypothesis is a claim of no difference in the means of the population, and that any difference in the sample means can be explained just by chance. The alternative hypothesis is that the two means really are different, and it's not just chance. For example: Deltamethrin does not have anticancer properties (Null Hypothesis), Deltamethrin have anticancer effects (Alternate Hypothesis) [8].

STUDY DESIGNS

After conversion of research questions into a null and alternate hypothesis which a lead to the formulation of objectives, the study is designed (Fig. 2). Studies can generally be classified in one of two ways: observational or experimental. In an observational study, the researchers simply "observe" a group of subjects without actually "doing" anything to the subjects. Case/control, cohort and cross–sectional studies are the examples of observational studies [9]. Experimental studies may be thought of studies where a research makes an intervention (such as giving a particular drug treatment to a patient). Preclinical study and randomised clinical trial are the good example of the experimental study. Experimental studies are further designed into parallel groups, crossover studies, sequential and factorial design according to the objectives of the study [10].

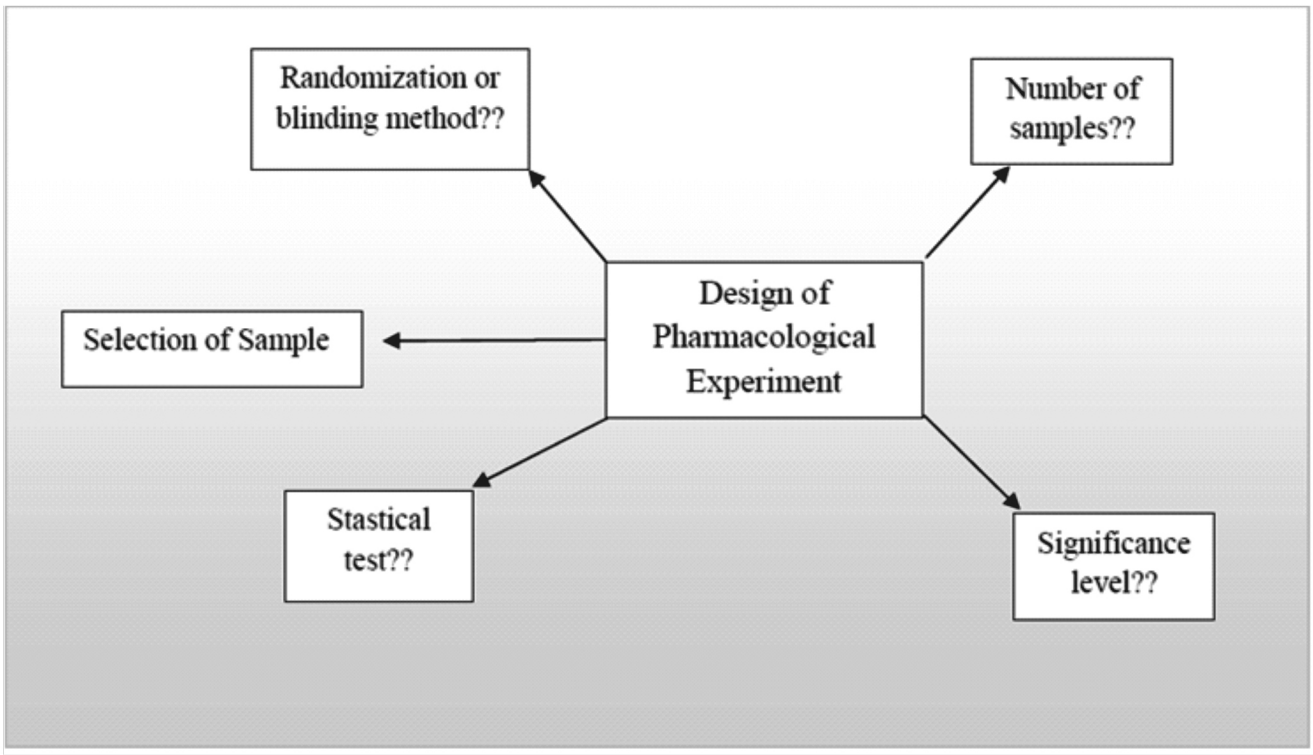

Figure 2: Systematic design of Pharmacological Experiments

CONCEPT OF SAMPLING

Sampling is at the core of scientific investigation. In reality, only a portion of the true population can be sampled in the almost all experimental design and finally, generalized the results in the whole population [11]. For example, the response of every Swiss albino mice on the planet cannot be measured. Sample size is also the most important issue in the design of experiments. Generally, more the sample size, more accurate the results, but it is unethical to use large numbers of animals. So, the number of animals used in an experiment can be reduced by employing a better experimental design. Sample bias occurs when a sample is taken intentionally from the specific part of the population. For example, if researcher, on the basis of experience, chooses only those animals (preclinical) or volunteer (clinical), who he/she know they will respond better to the test drug leads to bias in the result. Sample bias can be reduced by the use of various randomization techniques [12].

DESCRIPTIVE STATISTICS

After completion of experiments, data are collected and methods are needed to summarize the whole data. A measure of central tendency and dispersion is commonly used for this purpose. A measure of central tendency indicates the approximate centre of the sample distribution, and a measure of dispersion indicates the degree to which individual members of the sample set depart from the central value.

Measures of central tendency

The measure of central tendency provides a characteristic of the sample population and is useful when provided with additional statistics such as a measure of dispersion [13].

Mean

The mean is one of the most commonly used statistics in scientific research. It is simply the sum of all observations divided by the number of observations. However, the mean is not always the best choice for characterizing a distribution with a single number as it is highly sensitive to outliers [14].

Median

The median is defined as the middle term in a data set arranged in rank order. Its use is less as compare to mean, but it can be quite useful, especially in describing a data set containing outliers. The median is determined by rank ordering all the members in a data set and choosing the middle term for which 50% of the values lay above or below. This method provides a single number for data sets containing an odd number of values, but in case of an even number of values, a pair of numbers must be averaged to get the median for a data set. For non-parametric datasets, the median is used as a measure of central tendency [15].

Mode

The mode is simply the most commonly occurring value in the sampled distribution but often not useful in pharmacological research.

Measures of dispersion

The central tendency value only provides a central value of the data. It is unable to describe the distribution beyond the point of central tendency. For example, if a given concentration of DLM produces 50% decrease in cell viability, then it is expected that the next time if this concentration of DLM is applied, the loss in cell viability is likely to be close to 50%. The key question however is how likely? How surprising would an observation of 70% or 30% be? It is therefore essential that a measure of central tendency be coupled with a measure of dispersion, which describes the spread of the sample values about the point of central tendency in order to provide an interpretable summary.

Range, variance, and standard deviation

The basic goal of calculating a measure of dispersion is to generate a description of how far the data are spread out about the mean. The simplest and most intuitive measure is the range, which is simply the difference between the largest and smallest value but it is sensitive to outliers. The variance provides a reasonable estimate of the

spread in a distribution. The standard deviation (SD) is the most common measure of variability, measuring the spread of the data set and the relationship of the mean to the rest of the data. SD tells us dispersion of individual observations about the mean. In other words, it characterizes typical distance of an observation from distribution center or middle value. If observations are more disperse, then there will be more variability. Thus, a low SD signifies less variability while high SD indicates more spread out of data [16].

Standard error of the mean

The standard error of the mean (SEM) is a very commonly used as a measure of dispersion in the biological sciences. Despite its common use, it is generally misunderstood and used more by convention that for logical reasons. In fact the SEM is not a measure of dispersion and does not tell anything regarding the scatter of data about the mean. So, question arises, why SEM is commonly used? The fact that the SEM will always be smaller than the standard deviation may, in part, be responsible for its use in a graphical representation of data, providing an illusion that the measurements are more precise than they actually are but this is clearly not a justifiable reason to choose the SEM over the standard deviation.Actually, the SEM is a measure of how accurately the population mean has been estimated based on the central limit theorem (CLT) [17]. In conclusion, descriptive statistics simply tells, what is there in our data. Thus, to reach a valid conclusion, inferential statistics should be used. Hypothesis testing (using p- values) and point estimation (using confidence intervals) are two concepts of inferential statistics that help in making inference about population from samples.

INFERENCE

In most of pharmacological experiments, an inappropriate statistical method of inference is used due to lack of understanding of the underlying assumptions of the methods. The correct use of statistical methods in the hypothesis testing gives us a tool to decide between the null and alternative hypotheses. The null hypothesis is a default statement that the difference between the two groups has no effect. The alternative hypothesis is the opposite statement. At a very basic level, hypothesis testing involves the statement of the null and alternative hypothesis, selection of a distribution (e.g. t or F distribution), determination of the rejection and non-rejection regions of the chosen distribution, calculation of a test statistic, and decision on whether or not to reject the null hypothesis. The primary goal is to minimize Type I (α) representing false positives and Type II (β) errors representing false negatives [18, 19].

Type I (α) error

We could make a Type I error (reject when null is true) if we said there was a real difference when it was just by chance. If the P value is small enough, we will reject the null hypothesis and conclude there is a difference, but actually there is no difference (Table 1).

Table 1 : Outcomes of Hypothesis testing

Decision	Outcome if null hypothesis true	Outcome if null hypothesis false
Do not reject null hypothesis	Correct decision	Type II error
Reject null hypothesis	Type I error	Correct decision

Type 2 (β) errors or Power of study

Type II error arises when we accept the null hypothesis when actually it is not true or in other words failure to reject the null hypothesis when you should (Table 1). Usually we talk about power, which is 1-P (Type II error). So if the chance of Type II error is 10%, the power is 90%, which is considered very well. A study with low power may fail to reject even if the null hypothesis is false. We want to

design studies with good power. Once a power analysis is completed, a choice is made to either proceed with the experiment as designed using the appropriate sample size identified by the analysis, or to revisit the design in an attempt to either increase the reliability of measurement or increase the effect size. Therefore, the analysis of power plays a central role in experimental design.

HOW TO SELECT APPROPRIATE STATISTICAL TEST?

Selection of appropriate statistical test is very important for analysis of research data. Use of wrong or inappropriate statistical test is a common phenomenon observed in articles published in pharmacological journals. Wrong statistical tests can be seen in many conditions like the use of paired tests for unpaired data or use of parametric statistical tests for the data which does not follow the normal distribution etc. Because of the availability of different types of user friendly statistical software, performing the statistical tests become easy, but selection of appropriate statistical test is still a problem. Selection of appropriate statistical tests depends on the following three things: 1) what kind of data we are dealing with? 2) Whether our data follow the normal distribution or not? 3) What is the aim of the study? If we deal with continuous data, parametric test is applied. Secondly, if your data are following the normal distribution, parametric statistical test should be used and nonparametric tests should only be used when normal distribution is not followed. Lastly, what we want to compare? Whether we want to compare the drug with placebo? Or we want to compare the effect of intervention with standard. That comparison between the groups can be done by applying additional tests like Tukey (Compare all pairs of columns),Dunnett (Compare all columns with control) etc.[20, 21, 22].

PARAMETRIC AND NON-PARAMETRIC TESTS

Parametric statistical methods assume that the sample distribution is normally distributed. Parametric statistics include the most commonly employed tests such as t-test and ANOVA. Non parametric tests are used when data does not follow the normal distribution. Chi Square test, Mann–Whitney U test, Wilcoxon signed-rank tests are most commonly used non parametric tests (Fig. 3) [23, 24].

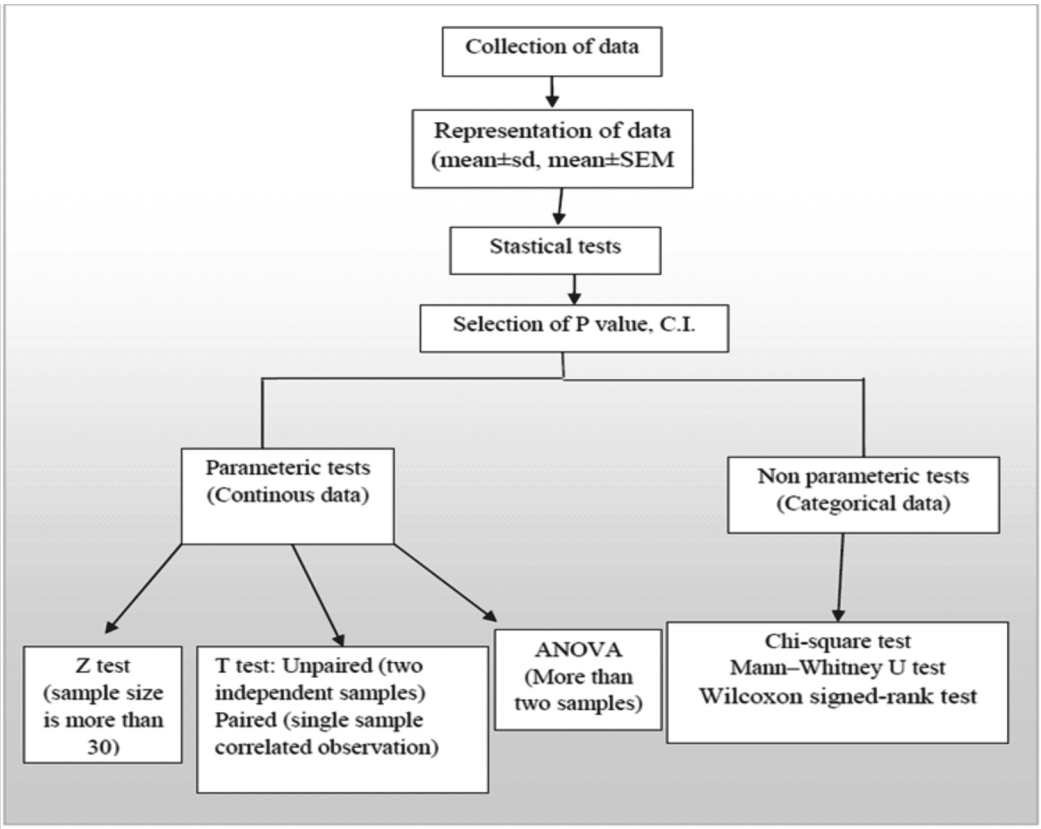

Figure 3: Parametric and non-parametric tests in Biostatistics

Parametric tests: assumptions and interpretation

Parametric tests are applied for quantitative (or continuous) data. Z test : Z test is applied when samples are large in size (more than 30). Student's t-test : The Student t-test is probably the most widely used parametric test. It is the first choice for comparing two independent groups. The t-test has several underlying assumptions that are critical to understand. It assumes that the means were drawn from normally distributed populations and are truly independent (i.e. no repeated measures).The paired version of Student's t-test is the best parametric choice for two groups, repeated measures. The unpaired t-test is used to determine whether the mean of a sample is different from a known average [25]. The t-test uses the mean, standard deviation, and number of samples to calculate the test statistic. The critical value for the Student's t-test is 1.96 for an alpha of 0.05, obtained from a t-test table. The t-test can be run as either a 1 tailed or 2 tailed tests with the two tailed being more conservative. When choosing between 1 and 2 tailed tests, one must decide is there is reason to believe that the treatment could only move the mean in one direction. If so, a one tailed test is appropriate. The p-value is interpreted against avalue (typically but not necessarily 0.05) and can be interpreted as the probability of rejecting the null hypothesis when it is in fact true (i.e. making a type I error).

Analysis of variance (ANOVA)

The analysis of variance (ANOVA) provides a statistical test of whether 3 or more means differ significantly. This is done by dividing the observed variance into different components (e.g. within groups, between groups).The test statistic for ANOVA is called the F-ratio. The F ratio and the degree of freedom (df) from each of the variability estimates are then used to obtain a p-value. The null hypothesis can be stated as the treatment does not produce any change in the mean values, and in general, an F ratio close to 1 indicates that the null hypothesis should be retained. The p-value is the probability that the observed differences would be expected to occur by chance alone. The ANOVA only indicates that there is a difference between the means, but does not indicate which means are different. Assuming the ANOVA does not allow the null hypothesis to be retained; subsequent post hoc test like Tukey (Compare all pairs of columns), Dunnett (Compare all columns with control) can be performed to determine which means are significantly different [26].

NON PARAMETRIC TESTS: ASSUMPTIONS AND INTERPRETATION

If the data do not meet the criteria for a parametric test (normally distributed, equal variance, and continuous), it must be analysed with a nonparametric test.

Chi Square Test

The chi-squared test is usually used to compare multiple groups. The most important assumption for this test, input and output variable should be binary.

Mann–Whitney U test

The Mann–Whitney U test (also called the Wilcoxon rank-sum test) is the most common nonparametric test used for the comparison of two independent groups. The Mann Whitney U test does not assume that the samples were drawn from normal distributions, but it does assume that the distributions have the same basic shape [22].

Wilcoxon signed-rank test

The Wilcoxon signed-rank test provides a non-parametric alternative to the paired t-test for the evaluation of matched samples. This test does not assume that difference values are normally distributed; however, it does assume that the distribution of differences is symmetrical about the median [27].

Kruskal–Wallis test

The Kruskal–Wallis test is a non-parametric test that can be used to compare 3 or more groups. It has the same assumptions as the Mann–Whitney U test [28].

Confidence interval and p-value

The aim of tests of significance is to calculate the "probability" that an observed outcome has merely happened by chance. This probability is known as the "p-value".If the p- value is small ($p < 0.05$), then the null hypothesis can be rejected and we can assert that findings are'statistically significant'. When $p < 0.05$, the degree of difference or association being tested would occur by chance only five times out of a hundred.When $p < 0.01$, the difference or association being observed would occur by chance only once in a hundred[29]. Hence, if our goal is to assess whether experimental results are likely to have occurred simply through chance or they are real, then the p-value calculation can be helpful but it cannot tell us how large or small the effect is. If we want an estimate of the actual effect, we need the confidence interval. Confidence interval (CI) is defined as 'a range of values for a variable of interest constructed so that this range has a specified probability of including the true value of the variable. The specified probability is called the confidence level, and the end points of the confidence interval are called the 'confidence limits'. By convention, the confidence level is usually set at 95%. The 95% CI is defined as "a range of values for a variable of interest constructed so that this range has a 95% probability of including the true value of the variable". In simple words, it means that we can be 95% sure that truths somewhere between 95% confidence interval. Because we are only 95% confident, there is a 5% probability that we might be wrong, i.e. 5% probability that the true value might lie either below or above the two confidence limits. Thus, the 95% CI corresponds to hypothesis testing with $p < 0.05$ [30, 31]. Thus, it can be concluded, that confidence interval is more informative than pvalue. Thus, that's why, the practice of reporting the p-value with confidence interval is highly recommended by the high impact journals.

STATICAL SIGNIFICANCE AND CLINICAL SIGNIFICANCE

Statistical significance does not necessarily mean a clinically important observation. It is the size of the effect that determines the importance and not the statistical significance. In the end, patients and physicians want to know the magnitude of the benefit or lack thereof, not the statistical significance of individual studies.

HYPOTHESIS TESTING OUTCOMES

Hypothesis testing gives us a tool to decide between the null and alternative hypotheses (Table 1) Hypothesis testing offers us two choices: 1. Conclude that the difference between the two groups is so large that it is unlikely to be due to chance alone. Reject the null hypothesis and conclude that the groups really do differ. 2. Conclude that the difference between the two groups could be explained just by chance. Accept the null hypothesis.

CONCLUSION

Statistics can be very helpful in formulating experimental design and drawings appropriate inferences from the collected data. If we employ better design and analysis, we will reduce the risk of making misleading claims and provide greater confidence that our proof of concept studies may translate into population. Therefore, it is essential for pharmacologists to have an understanding of the statistics.

ACKNOWLEDGMENTS

First Author, Anoop Kumar gratefully acknowledges the Department of Science and Technology (DST), New Delhi, India for providing financial assistance in the form of a DST-INSPIRE fellowship.

Declaration of Interest statement The authors declare no conflict of intersect.

REFERENCES

1. R. Spector, The Scientific Basis of Clinical Pharmacology: Principles and Examples. Boston, Little, Brown, 1986.
2. M.N. Ghosh, Fundamentals of experimental pharmacology, Indian. J. Pharmacol. 39 (2007) 216.
3. E.A. Murphy, A Companion to Medical Statistics.Baltimore, Johns Hopkins University P ress, 1985.
4. F. Prinz, T. Schlange, K. Asadullah, Believe it or

not: how much can we rely onpublished data on potential drug targets?, Nature Reviews Drug Discovery, 10 (2011) 712.

5. J.P. Ioannidis, Why most published research findings are false, PLoS Medicine, 2 (2005) e124.
6. A. Tatsioni, N.G. Bonitsis, J.P. Ioannidis, Persistence of contradicted claims in the literature, JAMA, 298 (2007) 2517–2526.
7. F. Patricia, B.A. Petrisor, F. Farrokhyar, M. Bhandari, Research questions, hypotheses and objectives, J. can. chir, 53 (2010) 278-281. 8. A. Kumar, D. Sasmal, N. Sharma, A. Bhaskar, S. Chandra, K. Mukhopadhyay, M. Kumar, Deltamethrin, a pyrethroid insecticide, could be a promising candidate as an anticancer agent, Medical Hypotheses, (2015) doi:10.1016/j.mehy.2015.04.018.
9. K.J. Rothman, S. Greenland, Modern Epidemiology, 2nd ed., Philadelphia, Lippincott Williams & Wilkins. 1998.
10. C. Cornu, B. Kassai, R. Fisch, C. Chiron, C. Alberti, R. Guerrini, P. Nony, Experimental designs for small randomised clinical trials: an algorithm for choice, Orphanet J Rare Dis, 8 (2013) 48.
11. J. Faugier, M. Sargeant, Sampling hard to reach populations. J. Adv. Nursing, 26 (1997) 790-797.
12. S.R. Anderson, A. Auquier, W.W. Hauck, D. Oakes, W. Vandaele, H.I. Weisberg, H.I. Statistical methods for comparative studies: techniques for bias reduction, John Wiley & Sons. 2009.
13. R.R. Wilcox, H.J. Keselman, Modern robust data analysis methods: measures of central tendency, Psychological methods, 8(2003) 254.
14. M. Maïda, J. Najim, S. Péché, Large deviations for weighted empirical mean with outliers. Stochastic Processes and their Applications, 117 (2007) 1373-1403.
15. J.D. Gibbons, J. D., &Chakraborti, S. Nonparametric statistical inference (pp. 977-979). Springer Berlin Heidelberg. 2011.
16. Jaykaran. "'Mean □} SEM' or 'Mean (SD)'?, Indian Journal of Pharmacol, 42.5 (2010) 329.
17. M.P. Barde, P.J. Barde. What to use to express the variability of data: Standard deviation or standard error of mean? Perspectives in Clinical Research, 3 (2012) 113–116.
18. A. Banerjee, U.B. Chitnis, S.L. Jadhav, J.S. Bhawalkar, S. Chaudhury, Hypothesis testing type I and type II errors, Industrial psychiatry journal, 18(2009) 127-131.
19. W.W. Daniel. In: Biostatistics. Hypothesis testing, 7th ed., New York: John Wiley and Sons, Inc, 2002, pp. 204–294.
20. J.F. Reed,P. Salen, P. Bagher, Methodological and statistical techniques: what do residents really need to know about statistics?, J Med Syst, 27(2003) 233–238.
21. J. Baptist du Prel, R. Bernd,H. Gerhard, B. Maria, Choosing Statistical Tests, Dtsch Arztebl Int, 107 (2010) 343–348.
22. T. Neideen, B. Karen, Understanding Statistical Tests, 64 (2007) 93-96.
23. Ton J. Cleophas, H.Z. Aeilko,Non-Parametric Tests. Statistical Analysis of Clinical Data on a Pocket Calculator (2011) 9-13.
24. K. Adrian, C.O. Daniel, A. Phil, Parametric and non-parametric tests, Pharmaceutical Medicine, Oxford University Press, 2013, DOI: 10.1093/med/9780199609147.001.0001.
25. E.H. Livingston, Who was student and why do we care so much about his t-test?, Journal of Surgical Research, 118 (2004) 58–65.
26. L. St, S. Wold, Analysis of variance (ANOVA), Chemometrics and intelligent laboratory systems, 6(1989) 259-272.
27. D.G. Altman. Statistics in medical journals: some recent trends, Stat. Med. 19 (2000) 3275–3289.
28. D.S. Kerby, The simple difference formula: An approach to teaching nonparametric correlation, Innovative Teaching, 3 (2014) doi:10.2466/11.IT.3.1.
29. A. Attia. Why should researchers report the confidence interval in modern research? Middle East Fertil Soc J. 10 (2005) 78–81.
30. A.K. Akobeng, Confidence intervals and p- values in clinical decision making. Acta Paedia tr, 97(2008) 1004–1007. 31. S.K. Gupta, The relevance of confidence interval and p-value in inferential statistics. Indian J Pharmacol, 44 (2012) 143–144.

FIGURE LEGENDS

Figure 1. Generation of Hypothesis.

Figure 2: Design of Pharmacological study.

Figure 3: Selection of appropriate Stastical test.