ORIGINAL RESEARCH PAPER

Volume-9 | Issue-1 | January-2020 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

ERRORS IN HYPOTHESIS TESTING: AN OVERVIEW

Dental Science

Dr. Dinesh Kumar Bagga Professor & Head Dept Of Orthodontics School Of Dental Sciences Sharda University Knowledge Park III Greater Noida (UP) India

Dr. Poonam Agrawal* Professor Dept Of Orthodontics School Of Dental Sciences Sharda University Knowledge Park III Greater Noida (UP) India *Corresponding Author

Dr. Madhurima Nanda Reader Dept Of Orthodontics School Of Dental Sciences Sharda University Knowledge Park III Greater Noida (UP) India

Dr. Sakshi Tiwari Lecturer Dept Of Orthodontics School Of Dental Sciences Sharda University Knowledge Park III Greater Noida (UP) India

Dr. Aartika Singh Lecturer Dept Of Orthodontics School Of Dental Sciences Sharda University Knowledge Park III Greater Noida (UP) India

Dr. Prashant Kumar Shahi Lecturer Dept Of Orthodontics School Of Dental Sciences Sharda University Knowledge Park III Greater Noida (UP) India

ABSTRACT

In hypothesis testing, the p value is in routine use as a tool to make statistical decisions. It gathers evidence to reject null hypothesis. Although it is supposed to reject the null hypothesis when it is false and fail to reject the null hypothesis when it is true but there is a potential to err by incorrectly rejecting the true null hypothesis and wrongly not rejecting the null hypothesis even when it is false. These are named as type I and type II errors respectively. The type I error (α error) is chosen arbitrarily by the researcher before the start of the experiment which serves as an arbitrary cutoff to bifurcate the entire quantitative data into two qualitative groups as 'significant' and 'insignificant'. This is known as level of significance (α level). Type II error (β error) is also predetermined so that the statistical test should have enough statistical power ((1-β)) to detect the statistically significant difference. In order to achieve adequate statistical power, the minimum sample size required for the study is determined. This approach is potentially flawed for the precision crisis due to choosing of arbitrary cutoff as level of significance and due to dependence of statistical power for detecting the difference on sample size. Moreover, p value does not tell about the magnitude of the difference at all. Therefore, one must be aware of these errors and their role in making statistical decisions.

KEYWORDS

Type I Error; Type II Error; Hypothesis; p Value

INTRODUCTION

Hypothesis testing aims at testing the null hypothesis by analyzing the data obtained following the experiment. The test result is supposed to reject the false null hypothesis while not rejecting the true null hypothesis; but there remains the probability of two types of error due to contradiction between the true condition and the obtained test result. These errors result due to wrong decision by incorrectly rejecting or failing to reject the null hypothesis[1,2,3,4]. When one makes an error by rejecting the true null hypothesis, this is known as Type I error (α error). When one makes an error by failing to reject the false null hypothesis, this is known as Type II error (β error). The study may contain only one out of these two errors. Both the errors are never made at the same time. Before conducting the experiment, the probabilities of these errors need to be controlled within limits. Usually type I error or alpha error is set at 5% (0.05) & type II error or beta error is set at 20% (0.2) for the statistical inference[2,3].

TYPE I ERROR

Type I error is the false alarm (false positive) error occurred due to concluding the difference when it didn't exist i.e. by wrongly rejecting the null hypothesis when it was true. This probability of type I error is predetermined to an acceptable upper limit of Type I error. This is termed as level of significance (α level). This serves as cutoff value for making statistical decision by *p* value based on it being greater than or less than α level.

Researchers routinely choose an alpha level of 0.05 for testing their hypotheses. The $p \leq 0.05$ rejects null hypothesis terming the difference as 'statistically significant' indicating the difference to be real whereas the $p > 0.05$ fails to reject null hypothesis calling the difference as 'statistically insignificant' indicating the difference to be not actual but due to the sampling variability.

Type I error occurs as a result of random chance. Multiple comparative tests for comparison of multiple experimental groups with the control groups inflate the chance of type I error which is called as cumulative type I error or alpha inflation or familywise error rate. This leads to a greater probability of false positive as compared to predetermined level of significance. These multiple comparisons result in at least one false conclusion showing one comparison to be falsely 'statistically significant'. As number of tests conducted increase, the likelihood of one or more comparisons to be falsely 'statistically significant' just due to chance (Type I error) also increase. Appropriate statistical test must be used to maintain the predetermined alpha level e.g. ANOVA followed by Bonferroni test rather than using multiple student's t test[5].

TYPE II ERROR

Type II error is the false negative error as actually there was a difference but the study concluded statistically insignificant result. The error has occurred as a result of failing to reject the null hypothesis given that the null hypothesis is actually false. In other words, it is failure to detect the difference between groups when one exists (concluding there is no difference). So, the probability of rejecting a false null hypothesis/ detecting the difference when it exists, (by not making a Type II or β error) is denoted as (1-β) and known as the statistical power of the statistical significance test.

When power (1-β) increases, the probability of making a Type II or β error (concluding there is no effect when, in fact, there is one) decreases.

The power analysis is conducted before the data collection to find out the smallest sample size needed to detect the given effect size at the desired level of significance with appropriate statistical power[6].

Statistical significance is generally set at 0.05. The desired statistical power level is between 0.80 and 0.90 but 0.80 is generally accepted. The sample size can be calculated after determining the appropriate effect size using previous studies or a pilot study and/or one's clinical observations. In some specific experimental designs (but not always), given any three of these four components i.e. the statistical power, the sample size, the effect size and the statistical significance level, we can determine the fourth.

If the sample is too small, the investigator might commit a Type II error due to insufficient power and rendering the strong and important effects to be statistically insignificant. Larger samples offer greater test sensitivity than small samples.

Larger the effect size, the smaller is the sample size required. Therefore, a small size is needed for the larger effects.

When the sample size is too large; even trivial effects of little clinical value can have impressive *p* values showing a magnified picture of tiny effect. Significance tests are highly dependent on sample size[7]. In other words, a statistically significant research outcome may be of a little clinical significance whereas statistically insignificant research outcome may be of clinical significance. An increase in sample size leading to an increase in statistical power may enable the statistical significance test to label the 'insignificant' difference as the 'significant' difference. Therefore we need to know appropriate sample size for which power analysis is done. Lack of precision and having no information about the magnitude of the effect (effect size) draw criticism for NHST (null hypothesis significance testing) from the researchers.

The smaller the value of alpha, it is less likely to reject a true null hypothesis resulting in alpha error or type I error. However, there are some situations where a lower alpha level (e.g. 0.01) or a higher alpha level (e.g. 0.10) may be desirable.

In an attempt to minimize error, if we choose a stringent alpha level of 0.01, then rejecting the null hypothesis becomes very difficult. So, the probability of rejecting null hypothesis is reduced when it is true, thereby reducing the probability of type I error; but the probability of rejecting null hypothesis is also reduced even when it is false thereby increasing the probability of a Type II error.

On the other hand, if we choose a lenient alpha level of 0.10, then rejecting the null hypothesis becomes easier. So, the probability of rejecting null hypothesis is increased when it is false, thereby reducing the probability of type II error; but the probability of rejecting null hypothesis is also increased even when it is true thereby increasing the probability of a Type I error.

The more we try to minimize a Type I error, the more likelihood of a Type II error creeping in, would be and vice versa[6,7].

As there is a delicate balance between type I error and type II error, it is a judgment call for setting a somewhat arbitrary alpha level by making a choice between the lesser of two evils i.e. a type I error or a type II error.

The choice is to be made between type I error (there is an effect when there isn't) or type II error (there is no effect when there is) while preferring one over the other. One error is more acceptable to the researcher than the other one.

Scientists have found an alpha level of 0.05 to be a good balance between these two issues.

In medical investigation, there are some instances in which a larger value of alpha (an α-level of 0.10 or even greater than 0.10) is set to reject a null hypothesis (where it is more acceptable to have a Type I error) whereas there are other instances in which a more stringent value of alpha at 0.01 is set to reject a null hypothesis (where it is more acceptable to have a Type II error)[8,9].

In medical screening, type I error is more acceptable than type II error. A false positive test for a disease (Type I error) will lead to inclusion of some non-diseased individuals into a group of diseased individuals. They will be recommended further confirmatory investigations where non-diseased individuals will be identified. A false negative test for a disease (Type II error) will lead to inclusion of some diseased individuals into a group of non-diseased individuals. They will be not be recommended further confirmatory investigations thereby letting the diseased individuals remain untreated and putting them on risk of developing complications. Given the choice we would rather accept a false positive (Type I error) than a false negative (Type II error) in this condition.

In surgical treatment of organ removal in cancer, type II error is more acceptable than type I error. There must be a high degree of evidence (α-level of 0.01) to reject the null hypothesis making rejection of null hypothesis difficult thereby reducing type I error (false positive). More likelihood of type II error (false negative) categorizing some of the cancer patients needing organ removal to be as the ones not needing organ removal will defer the organ removal in some of the patients. False positive error of recommending organ removal in cancer free individuals is not acceptable. Given the choice we would rather accept a false negative (Type II error) i.e. deferring the organ removal in cancer patients than a false positive (Type I error) i.e. organ removal in cancer free individuals[10].

CONCLUSION

Hypothesis testing requires thorough understanding of the errors and limitations of statistical inference. This helps in hypothesis formulation, predetermination of errors and sample size along with methods of sample selection, data collection and data analysis for the statistical model before running an experiment followed by a careful interpretation of the statistical result to decide regarding further experimentations and inclusion into the evidence based practice.

REFERENCES

1. Jamart J. Statistical tests in medical research. Acta oncologica 1992;31(1)723-7.
2. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127-31.
3. Kim HY. Statistical notes for clinical researchers: Type I and type II errors in statistical decision. Restor Dent Endod 2015;40:249-52.
4. Bajwa SJ. Basics, common errors and essentials of statistical tools and techniques in anesthesiology research. J Anaesthesiol Clin Pharmacol 2015;31:547-53.
5. Lee S, Lee DK. What is the proper way to apply the multiple comparison test? Korean J Anesthesiol 2018;71(5): 353-60.
6. Nayak BK. Understanding the relevance of sample size calculation. Indian J Ophthalmol. 2010;58(6):469-70.
7. Biau DJ, Kerneis S, Porcher R. Statistics in brief: The importance of sample size in the planning and interpretation of medical research. Clin Orthop Relat Res. 2008;466(9):2282-8.
8. Palesch YY. Some common misperceptions about p-values. Stroke 2014;45(12):e244-6.
9. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals and power: a guide to misinterpretations. Eur J Epidemiol 2016;31(4):337-50.
10. Bababekov YJ, Stapleton SM, Mueller JL, Fong ZV, Chang DC. A proposal to mitigate the consequences of type 2 error in surgical science. Ann Surg 2018;267(4): 621-2.