International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.960

Errors & power in biostatistics

Dr Dinesh Kumar Bagga
Professor & Head
Dept of Orthodontics
School of Dental Sciences
Sharda University

Received: 10 July Revised: 18 July Accepted: 26 July

Abstract

In hypothesis testing, the p value is supposed to reject the false null hypothesis but it must fail to reject the true null hypothesis. The two types of errors namely type I and type II errors are possible due to wrong decision by incorrectly rejecting or failing to reject the null hypothesis. These errors need to be addressed before running the experiment. Power analysis is an important step to calculate the minimum sample size for a given effect size at the desired level of significance (α error) assuring adequate statistical power (1-β) by controlling β error to detect the statistical significance. Power is the probability of not making type II error thereby detecting the statistically significant difference when it exists.

Keywords:-Type I error; Type II error; Statistical power; Statistical significance; hypothesis.

Introduction

Hypothesis formulation is the initial step in scientific problem solving which offers an explanation to the observed phenomenon based on some rationale. This observed phenomenon needs to be measured on a scale so that a testable hypothesis can be formulated. In contrast to the scientific theory for the explanation of certain phenomenon, which is tested and backed by evidence; the hypothesis is testable and predictable statement and is falsifiable also. It comprises two words "Hypo" meaning "tentative (subject to verification)" and "Thesis" meaning "statement about the solution of the problem". Hypothesis consists of 3 components i.e. variables, population and relationship of the variables[1].

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.960

Hypothesis testing

Hypothesis testing is performed to make statistical decision about the population based on the data collected by conducting a well designed experiment on a sample drawn from a specific population. This sample data is analyzed using various statistical tests and interpreted to report the conclusion about the population.

Null hypothesis represents the presumed default declaring there is no difference between the groups. The counterpart of this is the alternative hypothesis mentioning there is a difference. The goal of NHST (null hypothesis significance testing) is to reject the null hypothesis accurately in favour of the alternative hypothesis based on the *p* value at a specific level of significance. The *p* value is supposed to reject the false null hypothesis and fail to reject the true null hypothesis.

The level of significance is usually set at 0.05. The research outcome is termed statistically significant if $p \leq 0.05$ and statistically insignificant if $p > 0.05$[2,3].

Statistical errors

If the statistical test results do not match the true condition, the errors have occurred. Although the *p* value is supposed to reject the false null hypothesis and fail to reject the true null hypothesis but, the *p* value may reject the true null hypothesis and fail to reject the false null hypothesis. The error of rejecting the true null hypothesis is the Type I error or α error also called as false alarm or the false positive error. In other words, actually there was no difference but the study concluded statistically significant result. The error of failing to reject the false null hypothesis is the Type II error or β error also known as false negative error. In other words, actually there was a difference but the study concluded statistically insignificant result. The two types of errors are possible due to wrong decision by incorrectly rejecting or failing to reject the null hypothesis[4,5].

Both errors are never made at the same time. Only one type of error occurs at a time. A type-II error can only occur when a type-I error did not occur, and vice versa. Before conducting the experiment, these errors need to be addressed.

These two errors need to be taken into consideration during planning the experiment. 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). The *p* values are used to make statistical decision by rejecting or not rejecting the null hypothesis using an arbitrary cut off value known as a level of significance (α level). The α level is acceptable upper limit of Type I error (a "false positive" finding). Researchers routinely choose an alpha level of 0.05 for testing their hypotheses. 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. The smaller the value of alpha, it is less likely to reject a true null hypothesis resulting in alpha error or type I error.

Type II or β error is failure to reject null hypothesis when it is actually false. In other words, it is failure to detect the difference between groups when one exists (concluding there is no difference).

Statistical power

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[4,5]. 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 statistical power level generally accepted is 80% or 0.80 referring to the type II or beta error being set at 20% or 0.2)

The main purpose of power analysis is to find out the smallest sample size needed to detect the given effect size at the desired level of significance. Power analysis is conducted before the data collection.

Important components of power calculations are the sample size, the effect size, the α error and statistical power (1 – β). The fourth value can be calculated when rest three are known. 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 obtained using previous studies or a pilot study.

If the sample is too small, the investigator might commit a Type II error due to insufficient power. Larger samples may lead to a type I error labelling 'insignificant' difference as 'significant'[5].

Bigger effects are easier to detect as compared to the smaller effects. Larger the effect size, the smaller is the sample size required. Significance tests are highly dependent on sample size.

When the sample size is small; the strong and important effects can be categorized as statistically insignificant due to low statistical power of the significance test. When the sample size is large; even small effects can have lower *p* values due to the increased statistical power of detecting even a minute difference. In other words, a statistically significant research outcome may be clinically insignificant whereas a statistically insignificant research outcome may be clinically significant. 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[6].

In order to control type I error, if we choose a stringent alpha level of 0.01, then rejecting the null hypothesis becomes very difficult. So, the probability of rejecting true null hypothesis is reduced thereby reducing the probability of type I error; but at the same time, the probability of rejecting false null hypothesis is also reduced thereby increasing the probability of a Type II error.

On the other hand, if we opt for a lenient alpha level of 0.10, then rejecting the null hypothesis becomes easier. So, the probability of rejecting false null hypothesis is increased thereby reducing

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.960

the probability of type II error; but at the same time, the probability of rejecting true null hypothesis is also increased thereby increasing the probability of a Type I error.

This is amply clear that as we try to minimize Type I error, there is an increase in Type II error and when we try to minimize Type II error, there is an increase in Type I error[7,8]. So, we need to make a balance between these two. Scientists have found an alpha level of 0.05 to be the most acceptable value for the routine use[9].

However, there are some situations where we need to determine which is the lesser evil out of these two errors and we choose one type of error over the other. When we prefer type I error (as type II error is not acceptable) then we opt for a higher alpha level (e.g. 0.10). In other situation, where we prefer type II error (as type I error is more dangerous) then we opt for a lower alpha level (e.g. 0.01)[4].

While screening for a specific disease, Type I error is more acceptable than type II error. A false positive test for a disease (Type I error) will wrongly term some of the healthy individuals to be diseased ones whereas false negative test for a disease (Type II error) will wrongly term some of the diseased individuals as healthy individuals. By making an error, it is better to let the healthy individuals undergo further confirmatory tests rather than letting the diseased individuals remain untreated and putting them on risk of developing complications.

While conducting highly invasive surgical treatment such as organ removal in cancer, Type II error is more acceptable than type I error. One has to be very sure about the decision of surgical removal of an organ. 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). By making an error, it is better to let the patient needing surgical removal of an organ wait rather than the patient that can be managed without surgical removal to undergo surgical removal of the organ.

Misuse of statistical power

Power analysis must always be performed before running the experiment not after the research outcome. It is meant to calculate the appropriate sample size before the selection of the sample for the experiment. It should never be used retrospectively to calculate the observed power or post-hoc power so as to find out the statistical power in the already conducted experiment[10]. The reason is that observed (or post-hoc) power reflects *p*-values. Once the experiment ends with insignificant research outcome against expectation of the researcher, there is a tendency for critical appraisal of the experiment. With that intent, the researcher wishes to check the observed power based on the sample-size taken and effect-size of the data at a fixed α level. Since the research result is insignificant due to higher *p* value, the observed power is bound to be less. Steidl et al. (1997) observed that estimates of true power using retrospective power calculations will never exceed 0.53 in value[11]. This

results in interpretive bias as the researcher doesn't accept the unexpected result showing statistically insignificant difference due to the inadequate statistical power.

Conclusion

Statistical power analysis is an important concept in hypothesis testing which needs to be performed before running the experiment. This estimates appropriate sample size based on the effect size after determining the level of significance (α) and statistical power (1-β).

It should never be performed retrospectively (in post hoc manner) to calculate the observed power using the post-experiment data.

Once the study is over, the effect size can be calculated for further use in the power analysis of future studies.

References

1. Daniel WW. In: Biostatistics. 7th ed. New York: John Wiley and Sons, Inc; 2002. Hypothesis testing; pp. 204-94.
2. Jamart J. Statistical tests in medical research. Acta oncologica 1992;31(1)723-7.
3. 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.
4. 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.
5. Breur T. Statistical Power Analysis and the contemporary "crisis" in social sciences. J Market Anal 2016; 4(2-3):61-5.
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. Salsburg DS. The religion of statistics as practiced in medical journals. Am Stat 1985;39:220-3.
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.
11. Steidl RJ, Hayes JP, Schauber E. Statistical power analysis in wildlife research. J wildl manage 1997; 61(2):270-9.

There is no conflict of interest.