Importance of Hypothesis Testing, Type I, and Type II Errors – A Study of Statistical Power

Bibhu Dash
School of Computer and Info. Sciences
University of the Cumberlands
Williamsburg, KY USA
bdash6007@ucumberlands.edu

Azad Ali
School of Computer and Info. Sciences
University of the Cumberlands
Williamsburg, KY USA
azad.ali@ucumberlands.edu

***Abstract*—The statistical interpretation of data is crucial to research techniques and scientific studies. To infer something statistically, one must estimate and test a hypothesis. This is known as statistical inference and the 'Testing of Hypothesis' discussion is crucial when doing a statistical observation. The null and alternative hypotheses used in hypothesis testing result in Type I and Type II errors depending on whether they are accepted or rejected. We focus more on Type II mistakes when discussing statistical power since they are less likely to occur as statistical power increases. This study offers fresh perspectives and a thorough examination of each of these statistical variables, as well as how statistical power affects daily statistical inference and decision-making.**

***Keywords*—Hypothesis Testing, Type I error, Type II error, Statistical Power, Effect Size, Inference, Significance level**

I. Introduction

There is no such thing as a pure observation that does not depend on theory, which is why Karl Popper makes the crucial point that empirical scientists put the wagon before the horse when they assert that science moves from observation to theory (Proper, 1976). In essence, theoretical research leads to observatory experiments, which are subsequently applied as inductive evidence in statistical studies and research methodologies (Banerjee et al., 2009).

The creation of a hypothesis, which can subsequently be critically verified through observations and experiments, comes first in the scientific method rather than observation. Additionally, Popper makes the crucial point that the objective of a scientist's work is to refute an original premise rather than to confirm it. The validity of a general law cannot be rationally established by numerous observations, but it is theoretically conceivable to disprove the validity of such a law through a single observation. A strong research topic is the foundation of a strong hypothesis. For any scientific inferences, it should be clear, concise, and expressed upfront (Browner et al., 2001).

II. Hypothesis Testing and Types

In research and data analysis, hypothesis testing is a fundamental statistical technique that is used to draw conclusions about a population from a sample of data (El-gohary, 2019). It is assessing conflicting hypotheses to see if there is substantial evidence to support one of them. Typically, these competing hypotheses include a null hypothesis (H_0) and an alternative hypothesis (H_1). The level of significance and test power define the inverse relationship between the risks of these two errors. Therefore, before assessing their risks, you should decide which inaccuracy will have the most detrimental effects on your research. Researchers must be aware that no hypothesis test can be done with absolute certainty.

Making a type I error has probability *p*. When the null hypothesis is rejected, a value of 0.05 means the researcher is willing to accept a 5% possibility that the data are incorrect. The ramifications of committing an error that could cause a greater loss than another must be carefully considered by the researcher. The process to formulate hypothesis testing process in research:

I. *Null Hypothesis(H_0)*: This is a claim that there is no population-wide impact, distinction, or association. It stands for the current situation or a presumption.

II. *Alternate Hypothesis(H_1)*: This claim, which frequently makes a population-level effect, difference, or association, is the one you wish to test.

III. *Collect Data Sample*: Collect a sample of data relevant to the research questions and methodologies.

IV. *Significance Level(α)*: The cutoff point for statistical significance is represented by the significance level (α), which is typically set at 0.05 (5%). It shows the likelihood of committing a Type I error, which is to reject the null hypothesis when it is true. Based on the required level of confidence and the effects of making Type I errors, researchers select.

V. *Perform Statistical Test*: Depending on the research issue, the type of data (continuous, categorical, etc.), and the assumptions, choose the best statistical test. Determine a test statistic, which expresses the degree to which the null hypothesis is supported by the evidence.

VI. *Determine p-value to alpha*: If the null hypothesis is true, the p-value indicates the likelihood of obtaining a test statistic that is equally extreme to or more extreme than the one that was observed. You reject the null hypothesis (H_0) if the p-value is less than or equal to the significance level (Allua & Thompson, 2009). If the p-value exceeds, the null hypothesis is not rejected. It just indicates you don't have enough evidence to reject the null hypothesis, not that it is always correct.

VII. *Draw Conclusion and Interpretation*: If the null hypothesis is rejected, you get to the conclusion that there is strong evidence supporting the alternative hypothesis (H_1). If the alternative hypothesis is not supported by sufficient evidence, the null hypothesis is assumed to be true. Consider the results' theoretical and practical ramifications. Now is the moment to interpret the findings in light of the study question.

III. TYPE I (‘α’) AND TYPE II (‘β’) ERROR

A sample may not be a true representative of the population just by chance. As a result, the sample results do not reflect reality in the population, and the random mistake leads to an incorrect inference. A type I (false-positive) mistake happens when an investigator rejects a null hypothesis that is actually true in the population; a type II (false-negative) error arises when an investigator fails to reject a null hypothesis that is actually untrue in the population (see Table I) (Sedgwick, 2014). Although type I and type II mistakes can never be completely avoided, the investigator can lower their risk by increasing sample size (the larger the sample, the less likely it is to diverge significantly from the population).

Table I. Truth in population Vs. Results in Study Sample

Truth in the population	Association + nt	No associatic
Reject null hypothesis	Correct	Type I error
Fail to reject null hypothesis	Type II error	Correct

Many times, we receive false statistical inferences. The reasons may be many. Bias (observer, instrument, recall, faulty data sample, etc.) can also generate false-positive and false-negative outcomes. (Biased errors, on the other hand, are not referred to as type I and type II errors.) Such errors are problematic since they might be difficult to detect and are rarely quantifiable. Type II error is more critical for any scientific observation. The less the Type II error, the more accurate the research findings are, and that is called the true representation of Statistical Power.

IV. EFFECT SIZE AND STATISTICAL POWER

A. *Effect Size*

For statistical discoveries, the relationship between the predictor variable and the result variable is critical. The accurate measurement of the correct sample size is the magnitude of the association in the population. If it is large (say, 90%), the association in the sample will be easy to notice. In contrast, if the association is small (say, 2%), it will be difficult to detect in the sample. Unfortunately, the investigator frequently does not know the exact magnitude of the association - one of the study's goals is to estimate it. Instead, the investigator must select the size of the association that he wishes to find in the sample, known as effect size.

In other cases (healthcare or drug testing), a 1% difference is significant. As a result, the size of the effect is always somewhat discretionary, and feasibility issues are frequently crucial (Neumann & Kutterer, 2009). When the number of available individuals is limited, the investigator may have to work backward to evaluate whether the effect size that his study will be able to detect with that number of subjects is reasonable.

B. *Statistical Power*

The Statistical Power was formed because of the Type II mistake; the likelihood of rejecting the null hypothesis when it is untrue. Therefore, this null hypothesis should be rejected to avoid a Type II error. As a result, one must maintain a high Statistical Power, because the higher our Statistical Power, the fewer Type II mistakes can be predicted or occur.

Statistical Power analysis can be performed on either pre-collection or post-collection data. Statistical power is typically figured out by:

- The required power levels.
- The desired level of significance in the test.
- The population's strength of association or impact size.
- The sensitivity of the data.
- The sample sizes.

The power level in Statistical Power specifies the likelihood of not making a Type II error. Typically, the researcher selects 0.80 as the power level. In other words, the researcher has an 80% chance of not committing a Type II error. The level of significance in Statistical Power is the smallest probable possibility that a sample is likely to be related to the population. Assume the degree of significance is 5%. This indicates that the sample selected from the population should have at least 5% of the characteristics of the population from which it was drawn.

The effect size or strength of association in Statistical Power is essentially the strength of the relationship between the two variables. As a result, the larger the impact size, the bigger the Statistical Power. As a result, the test has a better likelihood of being legitimate. As a result, a larger effect size indicates better Statistical Power (Biau et al., 2010).

The number of true positives out of the total of true positives and false negatives is referred to as sensitivity in Statistical Power. Sensitivity, in layman's terms, perceives the actually right info. This means that a high sensitivity will produce solid data and, as a result, a high Statistical Power, which means data with fewer Type II mistakes. As a result, data sensitivity is a critical aspect of Statistical Power. It is the sample size that maintains the Statistical Power number high. This means that the greater our sample size, the higher our Statistical Power (Banerjee et al., 2009).

C. *p Value*

The *p-value* is defined as the likelihood of obtaining a result equal to or more extreme than what was actually seen on the premise of no impact or difference (null hypothesis). The *P* stands for probability, and it indicates the likelihood that any observed difference between groups is attributable to chance. Because *P* is a probability, it can have any value between 0 and 1. Values near 0 indicate that the observed difference is unlikely to be due to chance, whereas values close to 1 imply that there is no difference between the groups other than via chance. As a result, it is usual in medical publications to see adjectives such as "highly significant" or "very significant" after mentioning the *P* value, depending on how near the *P* value is to the significance level. As a result, it is typical to see adjectives like "highly significant" or "very significant" after stating the *P* number, depending on how near to zero the value is.

V. WHY THESE CONCEPTS ARE IMPORTANT?

Hypothesis testing, and type I and type II errors are key ideas in research techniques, especially in statistics. They contribute significantly to the scientific process by assisting researchers in making informed decisions about the validity of their hypotheses and the conclusions they derive from their

findings. Here's why these ideas are important:

i. Testing Hypothesis:

Formulating Clear Questions: The creation of a null hypothesis (H_0) and an alternative hypothesis (H_1) is the first step in hypothesis testing. These hypotheses describe research issues and aid in clarifying the study's objectives.

Objectivity and Precision: It offers a methodical and objective manner to evaluate the data supporting or refuting a certain hypothesis, which is essential for upholding objectivity and accuracy in research.

ii. Scientific Rigor:

Quality Control: In the scientific method, testing hypotheses serves as a quality control technique. It ensures that study conclusions are supported by statistical analysis rather than just intuition or anecdotal evidence (Louangrath, 2013).

Replicability: The dependability and fidelity of scientific discoveries are enhanced by well-conducted hypothesis testing, which enables the replication of research by other scientists.

iii. Decision Making:

Informed Decision-Making: A framework for deciding whether to accept or reject a null hypothesis is provided by hypothesis testing. In a number of disciplines, including engineering, economics, and medicine, this choice may have real-world repercussions (Biau et al., 2010).

Resource Allocation: Resources like time and finance are frequently in short supply for researchers. By concentrating efforts on hypotheses that have the best chance of being correct, hypothesis testing aids in the optimal allocation of these resources.

iv. Statistical Validity:

Validity and Reliability: As these are founded on meticulous statistical analysis, it guarantees that research findings are statistically valid and accurate.

Control Over Variability: The process of testing hypotheses takes data variability into consideration and aids in separating random fluctuations from significant impacts (Sedgwick, 2010).

Because hypothesis testing encourages the rigor, objectivity, and dependability of scientific studies, it is crucial to research technique, along with taking type I and type II errors into account. It enables researchers to generate defensible inferences, manage errors, and make wise judgments based on empirical data.

VI. Conclusion and Recommendations

The cornerstone of empirical research and the newly popular approach of evidence-based scientific inquiry is hypothesis testing. However, both hypothesis testing and empirical research have their limitations. The empirical method of study cannot totally remove uncertainty. It can, at most, measure uncertainty. Type I error, which involves incorrectly rejecting a null hypothesis, and type II error, which involves falsely accepting a null hypothesis, are both examples of this uncertainty. For computations of sample sizes, it is crucial to know the permissible magnitudes of type I and type II errors. Another crucial thing to keep in mind is that statistical analysis and hypothesis testing cannot 'prove' or 'disprove' anything.

Technology advancements always require a positive and supportive research environment and a good researcher attitude for developing successful data insights and decision-making. The null hypothesis can only be disproved or rejected, and the alternative hypothesis is thereby automatically accepted. The null hypothesis is accepted if we are unable to reject it. *P* value, sample size, and research questionnaires are all necessary for a successful scientific investigation. In order to prevent misinterpreting the P value, especially if they are utilizing statistical software for their data analysis, researchers are generally advised to consult a statistician at the outset of their study.

References

[1] Popper, K. R. (1976). Unended Quest: An intellectual autobiography Fontana.

[2] Browner, W. S., Newman, T. B., Cummings, S. R., & Hulley, S. B. (1988). Getting ready to estimate sample size: hypotheses and underlying principles. Designing clinical research, 2, 51-63.

[3] Banerjee, A., Chitnis, U. B., Jadhav, S. L., Bhawalkar, J. S., & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. Industrial psychiatry journal, 18(2), 127.

[4] El-gohary, T. M. (2019). Hypothesis testing, type I and type II errors: Expert discussion with didactic clinical scenarios. International Journal of Health and Rehabilitation Sciences (IJHRS), 8(3), 132.

[5] Allua, S., & Thompson, C. B. (2009). Hypothesis testing. Air medical journal, 28(3), 108-153.

[6] Sedgwick, P. (2014). Pitfalls of statistical hypothesis testing: type I and type II errors. Bmj, 349.

[7] Neumann, I., & Kutterer, H. (2009). The probability of type I and type II errors in imprecise hypothesis testing with an application to geodetic deformation analysis. International Journal of Reliability and Safety, 3(1-3), 286-306.

[8] Louangrath, P. (2013). Alpha and Beta Tests for type I and type II inferential errors determination in hypothesis testing. Available at SSRN 2332756.

[9] Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P value and the theory of hypothesis testing: an explanation for new researchers. Clinical Orthopaedics and Related Research®, 468, 885-892.

[10] Sedgwick, P. (2010). Errors when statistical hypothesis testing. BMJ, 340.

Authors:

Bibhu Dash is a Ph.D. scholar and researcher at the School of Computer and Information Sciences, University of the Cumberlands, KY. His research interests are inferential statistics, AI, and DeepNLP.

Dr. Azad Ali is a professor at the School of Computer and Information Sciences, University of the Cumberlands, KY. He has 32 years of experience as an educator and his research publications (on Statistics, AI, ERM) are available in all major high-impact journals.

Article Submitted: October 2019; Approved: December 2019.