

## ***Hypothesis Testing: The Two Kinds of Error and the Trade-off Between Significance and Power***

**Sayantana R. Majumdar<sup>1</sup>, Ritabrata K. Ganguly<sup>2</sup>**

### **ABSTRACT**

*A statistical hypothesis test decides, from data, whether to reject a claim, such as that a treatment has no effect, but because the data are random the decision can be wrong in two distinct ways: the test may reject the claim when it is in fact true, or fail to reject it when it is in fact false; these are the two kinds of error, and they cannot both be made small at once for a fixed amount of data. This study examines hypothesis testing, the two kinds of error, and the trade-off between significance and power. The chances of the two errors were examined as the threshold for rejecting the claim was varied, the distributions of the test result under the claim and under a real effect overlapping, so that the trade-off between the two errors could be seen. The two errors traded off against each other: setting a strict threshold, so that the claim is rejected only on strong evidence, made it unlikely to reject a true claim, a small chance of the first error, the significance level, but more likely to miss a real effect, a large chance of the second error and so low power; setting a lenient threshold did the reverse, catching real effects more often but rejecting true claims more often too. The two errors could not both be made small at once for a fixed sample, because the distributions of the test result under the claim and under an effect overlapped, and moving the threshold to reduce one error enlarged the other. The only way to reduce both together was to gather more data, which separated the two distributions and so shrank the overlap. The threshold was therefore chosen to fix the chance of the first error at a small, conventional level, accepting the resulting power, with the sample made large enough that the power was adequate. The study shows that the two errors of a test trade off through the rejection threshold and are reduced together only by more data, and it discusses the implications for the design and interpretation of statistical tests.*

**KEYWORDS:** *Hypothesis testing, Type I error, Type II error, Significance level, Statistical power, Null Hypothesis, Sample size, Statistical inference, Trade-off*

## INTRODUCTION

Much of empirical science and engineering turns on deciding, from data, whether some claim is true: whether a new treatment works, whether a process change made a difference, whether an observed effect is real or just chance. A statistical hypothesis test formalises this decision. It begins with a claim to be tested, usually a claim of no effect, the null hypothesis, and asks whether the data are too unlikely under that claim to be believed; if so, the claim is rejected in favour of the alternative, that there is an effect (1, 6).

Because the data are random, however, the decision can be wrong, and in two distinct ways. The test may reject the claim when it is in fact true, declaring an effect that is not there, a false alarm; or it may fail to reject the claim when it is in fact false, missing an effect that is really there. These are the two kinds of error, and they are different in nature: one is a false positive, the other a false negative, and a test that guards against one does not thereby guard against the other (2, 3).

Crucially, the two errors are in tension: for a fixed amount of data, making one less likely makes the other more likely, so they cannot both be made small at once. This paper examines hypothesis testing, the two kinds of error, and the trade-off between significance and power, examining the chances of the two errors as the threshold for rejecting the claim is varied. The aim is to show how the two errors trade off and how more data reduces both, and to discuss the implications for the design and interpretation of statistical tests.

## LITERATURE REVIEW

Hypothesis testing grew from the significance testing introduced to judge whether results were due to chance and the framework that formalised testing as a decision with two kinds of error, developed in foundational work and set out in the texts on statistical inference (1, 2, 3, 4). The two errors, the rejection of a true null and the failure to reject a false one, and the related notions of significance level and power, are central to this framework (2, 3).

## **ERRORS, SIGNIFICANCE, AND POWER**

The literature describes the two errors and their measures:

- A Type I error rejects a true null; its probability is the significance level (2, 3).
- A Type II error fails to reject a false null; one minus its probability is the power (2, 8).
- For a fixed sample, reducing one error increases the other (2, 3).
- Both errors are reduced together only by a larger sample (6, 8).

A recurring theme is a trade-off between the two errors, escaped only by more data.

The literature establishes that the significance level, the chance of rejecting a true null, is fixed by the experimenter at a small conventional value, and that the power, the chance of correctly rejecting a false null, then follows from the threshold, the size of the effect, and the sample (2, 3, 8). It shows that for a fixed sample the two errors trade off, since the distributions of the test statistic under the null and under an effect overlap, so moving the rejection threshold trades one error for the other, and that only enlarging the sample, which separates these distributions, reduces both, themes treated throughout the texts on inference and power analysis (4, 5, 6, 7, 9, 10).

## **RESEARCH GAP**

Although the two errors and their trade-off are foundational, a concise demonstration that shows the distributions under the null and under an effect overlapping, and how the threshold trades one error for the other, remains useful as a clear illustration. There is therefore value in a study that displays the two distributions and relates the trade-off to test design.

## **OBJECTIVES**

The objectives of this study are:

- To examine the chances of the two errors as the rejection threshold is varied.
- To show how a strict threshold reduces the first error but increases the second, and vice versa.
- To show that the two errors cannot both be made small at once for a fixed sample.
- To relate reducing both errors to enlarging the sample, and to test design.

## METHODOLOGY

### 1. The Two Errors and the Threshold

The chances of the two errors were examined as the threshold for rejecting the claim was varied, the distribution of the test result under the claim and under a real effect overlapping, so that the chance of the first error (rejecting a true claim) and of the second (missing a real effect) could be seen (2, 3). The elements are summarised in Table 1.

*Table 1: Elements of a hypothesis test*

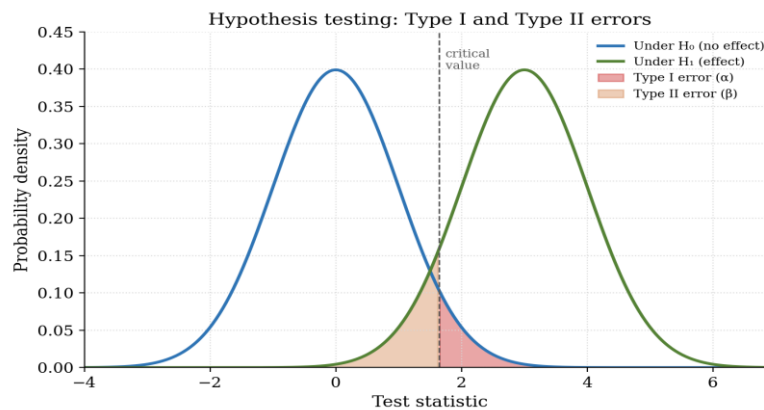
| Element            | Description                                |
|--------------------|--------------------------------------------|
| Null hypothesis    | The claim tested, usually of no effect     |
| Significance level | Chance of rejecting a true null (Type I)   |
| Type II error      | Failing to reject a false null             |
| Power              | Chance of correctly rejecting a false null |

### 2. Varying the Threshold

The two error chances were examined as the rejection threshold was moved, a strict threshold reducing the first error but enlarging the second, and the reverse, with the overlap of the two distributions making both small at once impossible for a fixed sample (2, 3). The reduction of both by a larger sample, separating the distributions, was considered (8).

## RESULTS AND FINDINGS

The two errors traded off against each other, as shown in Figure 1, which displays the distributions of the test result under the claim and under a real effect, with the threshold and the two error areas. Setting a strict threshold, so that the claim is rejected only on strong evidence, made it unlikely to reject a true claim, a small chance of the first error, the significance level, but more likely to miss a real effect, a large chance of the second error and so low power (2, 3).



**Figure 1: Hypothesis testing: Type I and Type II errors**

Setting a lenient threshold did the reverse, catching real effects more often but rejecting true claims more often too. The two errors could not both be made small at once for a fixed sample, because the distributions under the claim and under an effect overlapped, and moving the threshold to reduce one enlarged the other; only gathering more data, which separated the distributions, shrank both together (8). The findings are summarised in Table 2.

**Table 2: Errors with the rejection threshold**

| Threshold / action | First error (Type I)          | Second error (Type II) |
|--------------------|-------------------------------|------------------------|
| Strict             | Small                         | Large (low power)      |
| Lenient            | Large                         | Small (high power)     |
| Fixed sample       | Reducing one raises the other | —                      |
| More data          | Both can fall                 | Both can fall          |

## DISCUSSION

The results show that the two errors of a test trade off through the rejection threshold and are reduced together only by more data, and the trade-off arises from overlap. The test decides by comparing a test statistic to a threshold: above it, the claim is rejected. Under the claim the statistic has one distribution, under a real effect another, shifted from it; but the two overlap, so some values are plausible under both. The threshold cuts through this overlap. Set it high and few true-claim cases exceed it, so the first error is rare, but many real-effect cases also fall

below it, so the second error is common; set it low and the reverse holds. Moving the threshold thus shifts probability from one error to the other, and cannot reduce both at once, because the overlap is fixed for a given sample (2, 3).

This is why the two errors are treated asymmetrically in practice, which shapes how tests are set up. The first error, a false positive, is usually regarded as the more serious, since it amounts to claiming a discovery that is not real, so it is controlled directly: the threshold is set to fix the chance of the first error at a small conventional level, the significance level, before the data are seen. The power, the chance of catching a real effect, is then whatever that threshold and the data allow; it is not controlled directly but follows. So a test is built to limit false positives first, and its ability to detect true effects is a consequence to be checked, not the primary lever (1, 2, 6).

Escaping the trade-off requires more data, which is the key practical lesson and the basis of experimental design. Because the trade-off comes from the overlap of the two distributions, and that overlap shrinks as the sample grows and the distributions narrow and separate, a larger sample reduces both errors at once: it allows a strict threshold, keeping false positives rare, while still detecting real effects, giving high power. So the way to make a test both cautious and sensitive is to gather enough data; how much is the question power analysis answers, finding the sample size that gives adequate power at the chosen significance level for an effect of interest (6, 8). Too small a sample yields a test that, however correct its significance level, is too weak to detect real effects, a common pitfall; this is why power and sample size are considered before an experiment, not after. The practical conclusion is that a test's threshold fixes the false-positive rate and the sample must be large enough for adequate power, with the two errors reduced together only by more data, the main caveats being that the effect size and assumptions enter the power and that the figure illustrates the trade-off rather than a specific test.

## CONCLUSION

Hypothesis testing, the two kinds of error, and the trade-off between significance and power were examined by examining the chances of the two errors as the rejection threshold was varied. The errors traded off: a strict threshold made the first error (rejecting a true claim) rare

but the second (missing a real effect) common, and a lenient threshold the reverse, and both could not be small at once for a fixed sample because the distributions under the claim and under an effect overlapped; only more data reduced both. The study shows that the two errors trade off through the rejection threshold and are reduced together only by more data, so that a test fixes the false-positive rate by its threshold and must have a sample large enough for adequate power, with the caveats that the effect size and assumptions enter the power and that the figure illustrates the trade-off rather than a specific test.

## LIMITATIONS

The following limitations apply to this study:

- The power depends on the size of the real effect, which is usually unknown in advance.
- The figure illustrates the trade-off rather than the distributions of a specific test.
- The analysis assumes the test's distributional assumptions hold.
- Significance and power were treated for a single test; multiple testing raises further issues.

## FUTURE SCOPE

Future work could extend the present study as follows:

- Power analysis to determine sample sizes for adequate power.
- The effect of the assumed effect size on power and required sample.
- Robustness of tests when distributional assumptions are violated.
- Control of error rates under multiple simultaneous tests.
- Interpretation of p-values and significance in relation to the two errors.

## REFERENCES

1. Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.
2. Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, 231(694–706), 289–337.
3. Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Springer.
4. Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.

5. Rice, J. A. (2007). *Mathematical statistics and data analysis* (3rd ed.). Thomson Brooks/Cole.
6. Montgomery, D. C., & Runger, G. C. (2018). *Applied statistics and probability for engineers* (7th ed.). John Wiley & Sons.
7. Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications* (7th ed.). Thomson Brooks/Cole.
8. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
9. Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. Springer.
10. Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2016). *Probability and statistics for engineers and scientists* (9th ed.). Pearson.

**\*Author for Correspondence**

**Sayantan R. Majumdar**

**E-mail:** sayantan\_majumdar7@yahoo.com

<sup>1</sup>Associate Professor, Department of Statistics, Rarh Institute of Technology, Bardhaman, West Bengal

<sup>2</sup>Assistant Professor, Department of Mathematics and Statistics, Rarh Institute of Technology, Bardhaman, West Bengal

Received Date: May 21, 2026

Accepted Date: June 04, 2026

Published Date: June 11, 2026

**Citation:** Sayantan R. Majumdar, Ritabrata K. Ganguly. Hypothesis Testing: The Two Kinds of Error and the Trade-off Between Significance and Power. Journal of Engineering Mathematics & Statistics. 2026; 10(1): 9-16p.