Reading Notes on α and β Errors

Masaru Okada

November 10, 2025

1 The Concept of Statistical Significance and Its Practicality

In statistical discussion, the expression '**statistically significant**' carries a meaning far more critical than a mere numerical difference in magnitude. Consider, for example, a major corporation where a new initiative contributes to an annual sales increase of only one yen; from a practical standpoint, this information is almost meaningless. Yet, even if this extremely minuscule and 'imperceptible difference' is present, statistical science deems the difference '**significant**' if objective evidence suggests it's unlikely to have arisen from the **random variability of data** (random noise).

1.1 Statistical Significance versus Practical Importance

Typically, some difference will emerge in representative values like the mean or proportion between two groups—say, a group trialing a new marketing method and a control group using the conventional one. Given data variability, this is natural. However, when the difference is **substantial** enough, for instance, exceeding two standard deviations ($\pm 2SD$) of the data, it's considered reasonable to assume a systematic underlying factor, or a 'meaningful difference,' rather than simple chance. This criterion serves as a benchmark for judging statistical significance. The essence of this concept, though, lies not in the **magnitude** of the difference, but in the **low probability that the difference is attributable to chance**. What we seek to ascertain is whether the observed difference is a repeatable, **essential difference**, not just a matter of 'happenstance.'

2 The Importance of Power and Realistic Challenges

In real-world data analysis, it's rare to find mean values between compared groups diverging by as much as two standard deviations. Should such a large difference exist, the distinction would likely be **intuitively obvious** without requiring specialized statistical analysis. Thus, the real challenge in statistics is determining how to find **realistic differences** that are smaller than two standard deviations but still hold practical significance, using the **minimum amount of data**. This capability is referred to in statistics as '**Statistical Power**'.

2.1 Definition of Statistical Power and the Difficulty of Maximization

Statistical power is simply defined as 'the **probability of correctly detecting a fact as a significant difference** through a statistical test, **given that the hypothesis of a true difference is**

correct'. Maximizing this power is key to enhancing the accuracy of research and business decision-making. However, simply maximizing power doesn't solve everything. In an extreme sense, by adopting the strategy of 'irresponsibly asserting every idea conceived, without any supporting data,' one would find the meaningful difference 100% of the time if the hypothesis were indeed correct. While this might appear to maximize power, it is a **harmful approach** that simultaneously generates many false claims. There are individuals who plausibly assert unfounded ideas, and their actions can be viewed as an extreme example of trying to maximize power alone. By constantly predicting some 'crisis' or 'success,' they end up hitting a prediction by chance, much like a 'broken clock being right twice a day.'

3 α and β Errors: Types of Mistake

The approach of maximizing power alone is harmful because it ignores the risk of the opposite mistake —that is, the risk of '**mistakenly accepting a false hypothesis as true**'—while only considering the risk of failing to detect a correct hypothesis, known as the β error. There are two types of error in statistics.

3.1 Type I Error (α Error)

The mistake of **erroneously asserting that a difference exists** when in fact there is none is called the 'α **error**' (Type I error). This is the error of rejecting the null hypothesis when the truth is 'no difference,' and it is often referred to as the '**a**larmist's error.' People who spout baseless claims are arguably rushing to zero out the blockhead's error (β error) by asserting something, without considering the risk of committing this α error.

3.2 Type II Error (β Error)

Conversely, the mistake of **failing to detect a difference** and concluding 'no difference' when a difference **actually exists** (i.e., the alternative hypothesis is true) is called the 'β **error**' (Type II error). This is the error of failing to reject the null hypothesis when the truth is 'a difference exists,' and it is known as the '**b**lockhead's error.' An extreme example of seeking to eliminate this error is found in those who advocate for endless, cautious debate, avoiding all action or hypothesis formation on the grounds that 'it's not strictly knowable, no matter who asserts what.' They can zero out the risk of α error, but they continually overlook even patent truths. However, most real-world decisions are a race against time; just as a doctor who only observes a patient cautiously without starting treatment risks a mounting loss (or life), being a blockhead carries a significant cost.

4 Setting the Significance Level and the Trade-Off

The α error (alarmist's error) and the β error (blockhead's error) are fundamentally in a **trade-off** relationship. Since we are dealing with the uncertain event of data variability, it's impossible to completely eliminate both errors simultaneously. Statistics formalizes the process of making realistic and rational judgments between these two types of error.

4.1 Framework for Optimal Decision-Making

As a first step towards this, statistics requires clearly defining the **acceptable range for the α error**. This tolerance range is called the '**significance level** (α level).' This level is typically set at 5% or 1%. This is the **risk threshold** set in advance by the researcher or practitioner, stating, 'We will limit the probability of mistakenly concluding a difference exists when there is none to a maximum of this level.' After setting this constraint (upper limit) on the α error via the significance level, the goal then becomes minimizing the β error, or in other words, **maximizing statistical power**.

5 Statistical Hypothesis Testing and the Most Powerful Test

Statistical hypothesis testing is the general term for the entire set of **methods used to determine whether a specific hypothesis is likely to be correct**.

5.1 Selecting a Test Method and Statistical Power

Simply **increasing the amount of data (sample size)** used in the analysis increases power and reduces the risk of overlooking the truth. However, in practice, data is often limited. Thus, to avoid carelessly overlooking the truth even with limited data, it is extremely important to **select the optimal test method based on the hypothesis and type of data**. For example, the t-test is an option for comparing differences in means, and the χ^2 test for differences in proportions. Furthermore, within statistics, there exist testing methods that are theoretically proven to have the **highest power** under the constraint of a pre-determined **significance level** (α error upper limit). This is called the '**Most Powerful Test**'.

6 Practical Example of Testing in Marketing

The concept of statistical testing is routinely used, particularly in the **evaluation of marketing initiatives** like A/B testing. For instance, suppose a website's new design results in a marginal 0.01% increase in conversion rate, from 0.10% to 0.11%. Although this 0.01% difference is very small, if it truly represents a meaningful advantage (essential superiority), it could potentially boost the service's long-term sales by a factor of 1.1. Conversely, if this 0.01% difference was merely due to **random fluctuation**, all subsequent design changes and system modifications would be wasted costs, leading the company into a cycle of pursuing 'mere chance.' It is this statistical testing framework that is used to judge whether this **mere** 0.01% **difference should be deemed 'statistically significant'** or rejected as an 'insignificant difference due to chance.' Testing provides an objective method for quantitatively managing the risks (α and β errors) in this decision-making process, enabling the most rational judgment.

7 Universal Theories and Concepts

Statistical significance, Data variability, Standard deviation, Statistical power, α error, β error, Null hypothesis, Alternative hypothesis, Significance level, Statistical hypothesis testing, Most powerful

test, t-test, χ^2 test, Trade-off

7.1 Comprehension Check Quiz

1. In statistical testing, what is the term for the error of mistakenly concluding that a difference exists when the truth is 'no difference'?
2. In statistical testing, what is the term for the error of overlooking a difference and concluding 'no difference' when a difference truly exists?
3. What is the statistical metric defined as the probability of correctly detecting a difference when one truly exists?
4. What is the risk threshold that a researcher sets in advance as the acceptable upper limit for the probability of committing the error of mistakenly concluding a difference exists?
5. In the basic framework of statistical testing, which hypothesis is initially doubted and subjected to verification?
6. What does statistics call the state where a difference found in data analysis is judged not to have occurred due to random data fluctuation?
7. What is the technical statistical term for the error nicknamed the 'alarmist's error'?
8. In statistical testing, what is the term for the test method that achieves the highest statistical power under the constraint of a predetermined risk level?
9. What is one of the representative statistical indicators showing the degree of data variability, which expresses the spread around the mean?
10. In a situation where a doctor only observes a patient cautiously without starting treatment, potentially losing a life that could have been saved, which type of statistical error is at risk of being increased?
11. What is the relationship called when the probability of mistakenly asserting a difference (Question 1) and the probability of overlooking a true difference (Question 2) cannot both be simultaneously reduced to zero?
12. What is a representative test method used in statistical testing, such as when verifying a hypothesis about the difference in population means?
13. What is the systematic method that formalizes the two types of errors in statistical testing (Question 1 and Question 2) to guide rational decision-making?
14. What is a commonly used test method in marketing A/B testing and similar situations for verifying the association or difference between two groups when the data are categorical or frequency-based (proportions)?
15. What statistical metric is an economist, who constantly predicts 'a recession is coming soon,' ultimately pursuing and trying to maximize?

Answer Key

1. α Error, 2. β Error, 3. Statistical Power, 4. Significance Level, 5. Null Hypothesis, 6. Statistically Significant, 7. Type I Error, 8. Most Powerful Test, 9. Standard Deviation, 10. Type II Error, 11. Trade-off, 12. t-test, 13. Statistical Hypothesis Testing, 14. χ^2 test, 15. Statistical Power

8 References

References

[1] What is a p-value anyway ? - Vickers, Andrew J.

[2] Statistics is the most powerful discipline (Practical Edition) - Kei Nishiuci