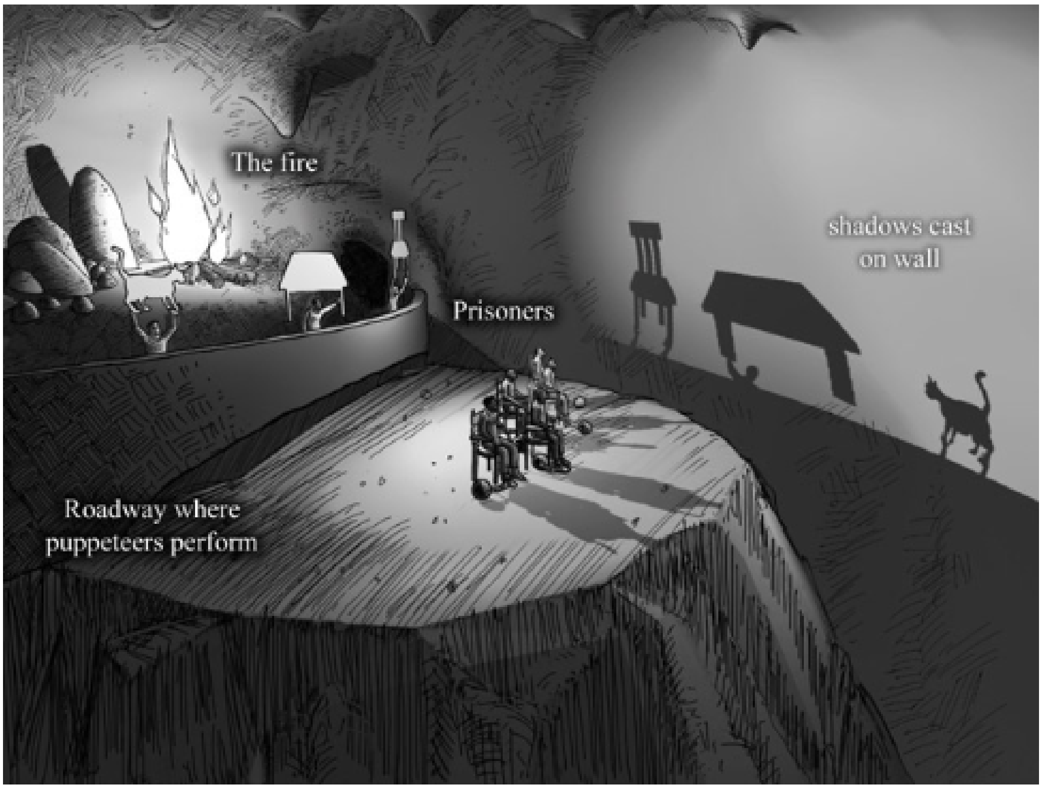
Lecture Notes: Hypothesis Testing

Making informed decisions is a ubiquitous aspect of life, confronting us with questions such as

- "Does a new drug improve health outcomes?"
- "Is a suspect guilty of a crime?"
- "Do environmental factors contribute to disease?"
- "Has a chess grandmaster player engaged in cheating?"

To address these varied queries, a robust statistical framework is essential. Hypothesis testing is the cornerstone of this framework, offering a systematic approach to drawing conclusions from data. This lecture will explore the principles and applications of hypothesis testing, demonstrating its critical role in deciphering data to make informed decisions.



In this exploration, we will see how hypothesis testing acts as a logical tool, akin to a statistical tautology, enabling us to validate or invalidate hypotheses based on empirical evidence. The allegory of Plato's Cave serves as a metaphor for this process, where just like deciphering shadows to uncover a hidden reality, we use data to reveal underlying truths.

Hypothesis testing

Statistical hypothesis testing is a key technique of frequentist statistical inference. In order to understand what hypothesis testing means, we describe an analogy. A statistical test procedure is comparable to a criminal trial; a defendant is considered innocent as long as his guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough incriminating evidence the defendant is convicted.

Modus tollens

The *modus tollens* is the Latin name for a valid argument form and rule of inference. It means “the way that denies by denying”. It is the inference that

1. if $P \Rightarrow Q$, i.e. if P implies Q,
2. and the second premise, Q, is false,
3. then it can be logically concluded that P must be false.

For example, we know that if it rains, that then the streets are wet. When we look outside, we see that the streets are dry, so therefore we can conclude that it doesn’t rain. The argumentation used in hypothesis testing is something like a *modus tollens.* The only thing is that we are not completely certain about the truth or falsehood of $P \Rightarrow Q$ anymore. We explain this in the following sections.

Null and alternative hypotheses

At the start of the procedure, there are two hypotheses:

H_0: “the defendant is not guilty”, and
H_1: “the defendant is guilty”.

The first one is called the *null hypothesis*, and is *for the time being* accepted. The second one is called the *alternative (hypothesis).* It is the hypothesis the prosecutor tries to prove.

Two kinds of errors

The hypothesis of innocence is only rejected when the evidence is really strong, because one doesn’t want to convict an innocent defendant. The error of convicting an innocent person is called an *error of the first kind.* The testing procedure should be set up in such a way that the occurrence of this error is rare. The reverse error, i.e. acquitting a person who committed the crime, called the *error of the second kind* is considered less serious. As a consequence of this asymmetric behaviour, the frequency of occurrence of the second type error is often rather large.

	H_0 is true	H_1 is true
	truly not guilty	truly guilty
Accept Null Hypothesis	Right decision	Wrong decision
Acquittal		Type II Error
Reject Null Hypothesis	Wrong decision	
Conviction	Type I Error	Right decision

- **Type I Error (False Positive)**: Occurs when the null hypothesis is true, but we incorrectly reject it. Denoted by α (alpha), also known as the significance level.
- **Type II Error (False Negative)**: Occurs when the null hypothesis is false, but we fail to reject it. Denoted by β (beta). The power of a test is defined as $1 - \beta$ and represents the probability of correctly rejecting a false null hypothesis.

Test statistic

A criminal trial can be regarded as either or both of two decision processes: guilty vs not guilty or evidence vs a threshold ("beyond a reasonable doubt"). What we wish is to judge the defendant, but really we are judging the performance of the prosecution (which bears the burden of proof). Similarly in hypothesis testing, these decisions are always made using some (hopefully) clever summary of the data. This summary is called a *test-statistic*. There are many summaries of the data possible: one might show nothing (e.g. “the suspect is a woman”), whereas another (e.g. “the suspect has blood-stained clothes”) might give some idea that the that the suspect is guilty.

P-value

We now attempt to reformulate the *modus tollens*, where

- P = “null hypothesis is true”
- Q = “not observing this value of the test statistic”

If $P \Rightarrow Q$ was certain, then if the null hypothesis is true, one would *never* observe this value of the test statistic. BUT we do observe exactly this value of the test-statistic, and therefore the null hypothesis must be false.

However, often $P \Rightarrow Q$ is not certain at all. If the suspect is not guilty, it is not certain that she has no blood on her clothes. She could have tried to help the victim immediately afterwards. We want to replace the certain statement,

$$P \Rightarrow Q$$

by an uncertain statement: “the probability that Q happens if P is true”, $P(Q|P)$. If this value is almost 1, then the implication is almost certain and then we can tolerate the *modus tollens* argument. However, if this is far from 1, then the implication is too uncertain to be able to use the *modus tollens*.

In actual fact, one does not use $P(Q|P)$, but $P(\text{not } Q|P) = 1 - P(Q|P)$. This value is called the *p-value*, i.e.,

> **p-value**: probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed, assuming that the null hypothesis is true.

P-values close to zero, therefore, mean that the implication $P \Rightarrow Q$ is almost certain and that we can use the *modus tollens* to reject the null hypothesis.

Significance level

One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom. The question is what is enough? Should the probability of the implication be 99% (i.e. the p-value = 0.01) or 99.9% (i.e. the p-value=0.001) before we accept the validity of the *modus tollens*? In most fields of science and society, one uses 95% as the cut-off – i.e. a cut-off of 0.05 for the p-value). This is the result of an arbitrary choice by the inventor of hypothesis tests, R.A. Fisher, in the 1930s. The cut-off is called the *significance level*.

In statistics, a result is called *statistically significant* if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The phrase “test of significance” was coined by Ronald Fisher: “Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first.” A result that was found to be statistically significant is also called a positive result; conversely, a result that is not unlikely (i.e. quite likely) under the null hypothesis is called a negative result or a null result.

The probabilities of the two types of errors are determined by the distributions of the test statistic under the null and alternative hypotheses.

If Z is our test statistic:

- The probability of a Type I error, $P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha$.
- The probability of a Type II error, $P(\text{Fail to Reject } H_0 | H_0 \text{ is false}) = \beta$.

Given a significance level α, the critical value Z_α is the value for which:

$$P(Z > Z_\alpha | H_0 \text{ is true}) = \alpha$$

And for the power of the test:

$$P(Z > Z_\alpha | H_1 \text{ is true}) = 1 - \beta$$

In hypothesis testing, we aim to minimize both Type I and Type II errors. However, in practice, there is a trade-off:

$$\text{As } \alpha \text{ decreases, } \beta \text{ increases (and vice versa).} \tag{1}$$

Thus, the goal is often to determine an acceptable level of α (common choices are 0.05, 0.01) and design the test to minimize β for that α. Another approach is to maximize the power of the test, which is $1 - \beta$, given a fixed α.

The choice of α should reflect the relative seriousness of the two types of errors in a particular testing context. For example, in medical testing, a Type I error (falsely diagnosing a healthy person as sick) might be deemed more serious than a Type II error (falsely diagnosing a sick person as healthy), leading to a choice of a smaller α.

Example 1. *We are investigating whether two groups, A and B, have a significant difference in their averages. Our study involves multiple tests to determine how often we fail to detect this difference, depending on the alpha level we choose. Alpha is a threshold that helps us decide if our test findings are significant. The key to this determination is the p-value, calculated as follows:*

$$p\text{-}value = \frac{\textit{Number of times difference in averages} \;\geq\; \textit{observed difference}}{\textit{Total number of tests}}$$

If the p-value is greater than alpha, it suggests that we may have missed a real difference between the groups. This process is repeated across various alpha levels to understand the correlation between alpha and the likelihood of overlooking an actual difference.

Example: a clairvoyant card game

Let's consider a more realistic example. A person (the subject) is tested for clairvoyance. He is shown the reverse of a randomly chosen playing card 25 times and asked which of the four suits it belongs to. The number of hits, or correct answers, is called X.

As we try to find evidence of his clairvoyance, for the time being the null hypothesis is that the person is not clairvoyant. The alternative is the reverse: the person is (more or less) clairvoyant. If the null hypothesis is valid, the only thing the test person can do is guess. For every card, the probability (relative frequency) of any single suit appearing is $1/4$. If the alternative is valid, the test subject will predict the suit correctly with probability different from $1/4$. We will call the probability of guessing correctly p. The two hypotheses, then, are:

H_0: $p = 1/4$ (just guessing)
H_1: $p \neq 1/4$ (true clairvoyant).

We call this a "two-sided test": this means that *a priori* it is logically possible that the clairvoyant has guessing probabilities both greater and less than $1/4$. This is typically the case in many practical situations. Therefore, in practice most tests are two-sided. The main consequence

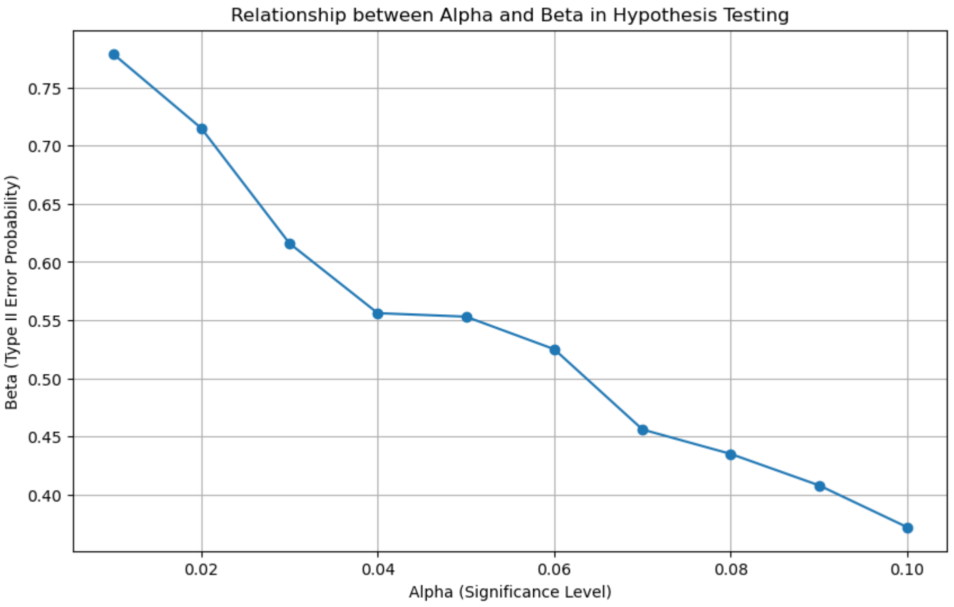


Figure 1: Plot illustrating the relationship between α and β in the simulation study.

of two-sided tests is the calculation of the p-value (the exceedance probabilities have to be multiplied by 2; see below).

When the test subject correctly predicts all 25 cards, it seems reasonable to consider him clairvoyant, and reject the null hypothesis. With only 5 or 6 hits, on the other hand, there is no cause to doubt the null hypothesis. But what about 12 hits, or 17 hits? What is the critical number, c, of hits, at which point we consider the subject to be clairvoyant? How do we determine the critical value c? In practice, one decides how critical one will be depending on the amount of errors (of the first kind) that one is willing to make. That is, one decides how often one accepts an error of the first kind a false positive, or Type I error. With $c = 25$ the probability of such an error is:

$$\text{Type I error} = 2 \times P(X = 25 | H_0 \text{ is true})^1 = \_\_\_\_\_\_.$$

and hence, very small. The probability of a false positive is the probability of randomly guessing correctly all 25 times times 2 – since the test is two-sided. Being less critical, with c=10, gives:

$$\text{Type I error} = 2 \times P(X \geq 10 | H_0 \text{ is true}) = \sum_{k=10}^{25} \binom{25}{k} 0.25^k 0.75^{25-k} \approx 0.14$$

Thus, $c = 10$ yields a much greater probability of false positive. What do you think? Is this enough “evidence” to accept the alternative hypothesis? Can you explain in words what this probability actually means?[2]

Before the test is actually performed, the desired probability of a Type I error is determined. Typically, values in the range of 1% to 5% are selected. Depending on this desired Type 1 error rate, the critical value c is calculated. For example, if we select an error rate of 1%, c is calculated thus:

$$\text{Type I error} = 2 \times P(X \geq c | H_0 \text{ is true}) \leq 0.01$$

[1] We use the notation $P(A|B)$ to mean “the probability of A if we know that B has happened”

[2] Many people made and continue to make a lot of mistakes in interpreting this probability. This probability does **not** mean the chances of the null hypothesis being false given the data! That’s for next chapter... This value means the probability of the data (or data as extreme as what has been observed) assuming that the null hypothesis is true... Got it? Think hard about it!

From all the numbers c, with this property, we choose the smallest. Why do we do this? Remember that there is also another error we can make, namely saying that the person is not clairvoyant, whereas he actually is. In order to minimize the probability of that Type II error, a false negative, we select $c = 12$ in the above example.

But what if the subject did not guess any cards at all? Having zero correct answers is clearly an oddity too. The probability of guessing incorrectly once is equal to $q = (1 - p) = 3/4$. Using the same approach we can calculate that probability of randomly calling all 25 cards wrong is:

$$\text{p-value} = 2 \times P(X = 0|H_0 \text{ is true}) = 0.75^{25} = 0.0014.$$

This is highly unlikely (a 1 in a 1000 chance). This would indeed be evidence for dismissing H_0 in favour of H_1. This may seem a bit strange, but result would suggest a trait on the subject's part of avoiding calling the correct card. That's also clairvoyance.

Forensics

In forensics, hypothesis testing helps determine the innocence or guilt of a defendant based on evidence like fingerprint or DNA matches. The null hypothesis typically assumes innocence, and the decision to reject or retain this hypothesis carries significant implications.

Epidemiology

In epidemiology, drug companies use hypothesis testing to compare new drugs against current treatments. Regulators require proof that a new drug is superior before approval. The null hypothesis in this scenario posits no difference between the new and current treatments.

Hypothesis Testing: Power Plant's Effect on Cancer Incidence

We aim to determine if a power plant influences the incidence of a specific cancer, measurable through a continuous scale like biomarker levels. Our linear regression model is:

$$Y_i = \beta_0 + \beta x_i + e_i$$

Here, Y_i represents cancer severity or biomarker levels, x_i indicates proximity to the power plant, and e_i is the error term. The coefficient β measures the power plant's effect.

Our hypothesis test is:

- $H_0 : \beta = 0$ (No effect)
- $H_1 : \beta \neq 0$ (Some effect)

The least squares estimator for β is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

A possible test statistic T is defined as:

$$T = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})}$$

If H_0 is true, the p-value is:

$$\text{p-value} = P(|T| \geq |t_{obs}|)$$

Where T is the t-statistic under H_0, and t_{obs} is the observed value.

The Central Limit Theorem ensures that, even if e_i is not normally distributed, $\hat{\beta}$ will approximate a normal distribution for large samples, allowing the use of standard t-tests.

Permutation Test

A permutation test is a non-parametric statistical method used for hypothesis testing, particularly to determine if two or more samples originate from the same distribution. The process is delineated as follows:

1. The null hypothesis H_0 posits no effect or difference between the groups, e.g., equal means.
2. All observations from the different groups are pooled together, creating a comprehensive pool of data.
3. This pooled data is then randomly permuted and redistributed into new groups, each matching the size of the original groups.
4. For each permutation, a test statistic is calculated. This statistic could be any measure of interest, depending on the research question:

$$T = f(Y^{(1)}) - f(Y^{(2)})$$

5. This process is repeated numerous times (e.g., 10,000 times) to generate a distribution of the test statistic under H_0.
6. The actual test statistic calculated from the original, unpermuted data is then compared to this distribution.
7. The p-value is determined by the proportion of permuted test statistics that are as extreme or more extreme than the observed statistic:

$$p\text{-value} = \frac{\text{Number of permutations with } T \geq \text{ observed } T}{\text{Total number of permutations}}$$

A low p-value (typically less than 0.05) indicates that the observed difference is unlikely due to chance, leading to a rejection of H_0. This test is valued for its versatility and robustness, as it does not make assumptions about the distribution of the data.

Example

Consider the scenario of evaluating two teaching methods in terms of their effectiveness. Our objective is to use a hypothesis testing framework to ascertain if the observed differences in student pass rates between two schools are statistically significant.

Hypotheses:

- Null Hypothesis (H_0): There is no significant difference in the pass rates between the two schools.
- Alternative Hypothesis (H_1): There is a significant difference in the pass rates between the two schools, attributable to the teaching methods.

The data includes results from two schools: School A, employing a new teaching method with 773 students (of whom 18 passed), and School B, using a traditional method with 1123 students (of whom 11 passed). The permutation test, chosen for its robustness in non-parametric settings, is used to compare the pass rates.

The permutation test produced an observed difference in pass rates of approximately 0.0135. The p-value for this observed difference is 0.0398. Since this p-value is below the conventional threshold of 0.05, we reject the null hypothesis in favor of the alternative hypothesis. This result

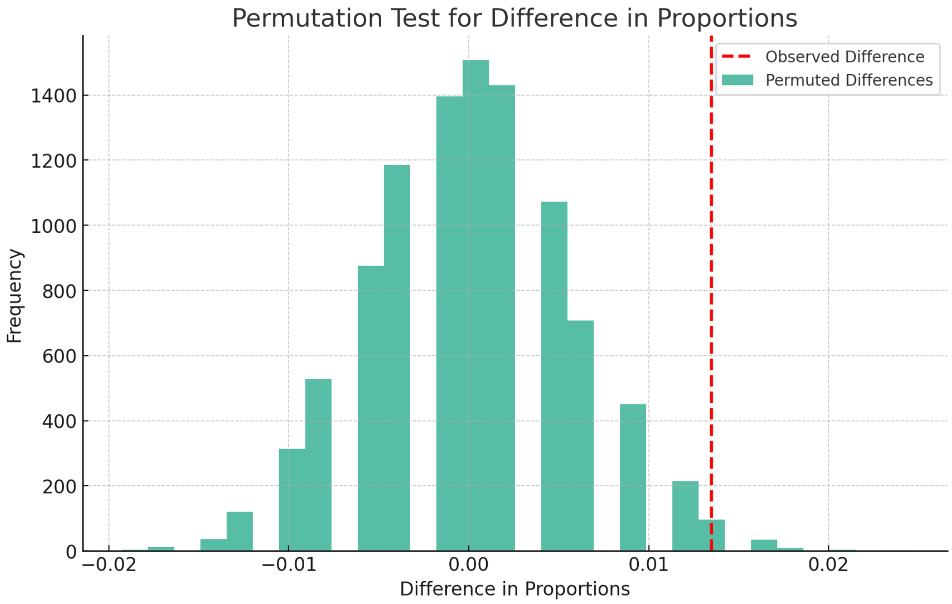


Figure 2: Histogram of permuted differences in pass rates with the observed difference indicated.

suggests that the difference in pass rates between the two schools is statistically significant and may be attributed to the difference in teaching methods.

In conclusion, the statistical analysis supports the hypothesis that the new teaching method implemented in School A is more effective than the traditional method in School B, as evidenced by the significantly different pass rates.

Is Nakamura Cheating?

Following recent allegations of cheating by Vladimir Kramnik against Hikaru Nakamura, this analysis aims to statistically evaluate Nakamura's winning streaks. Kramnik's accusations, made in November 2023, were based on Nakamura's exceptional performance, including a 45.5 out of 46 win rate. Chess.com's investigation found no evidence of cheating, but we aim to independently assess the statistical likelihood of such performance (source).

The hypotheses tested are:

- H_0: The occurrence of long winning streaks is due to chance.
- H_1: The occurrence of long winning streaks is too improbable to be due to chance, indicating potential cheating.

Using the ELO rating system, the probability of Nakamura winning a game is calculated, taking into account his high performance in recent games:

$$P(\text{Win}) = \frac{1}{1 + 10^{(\text{Opponent ELO} - \text{Nakamura ELO})/400}}$$

The simulation parameters are as follows:

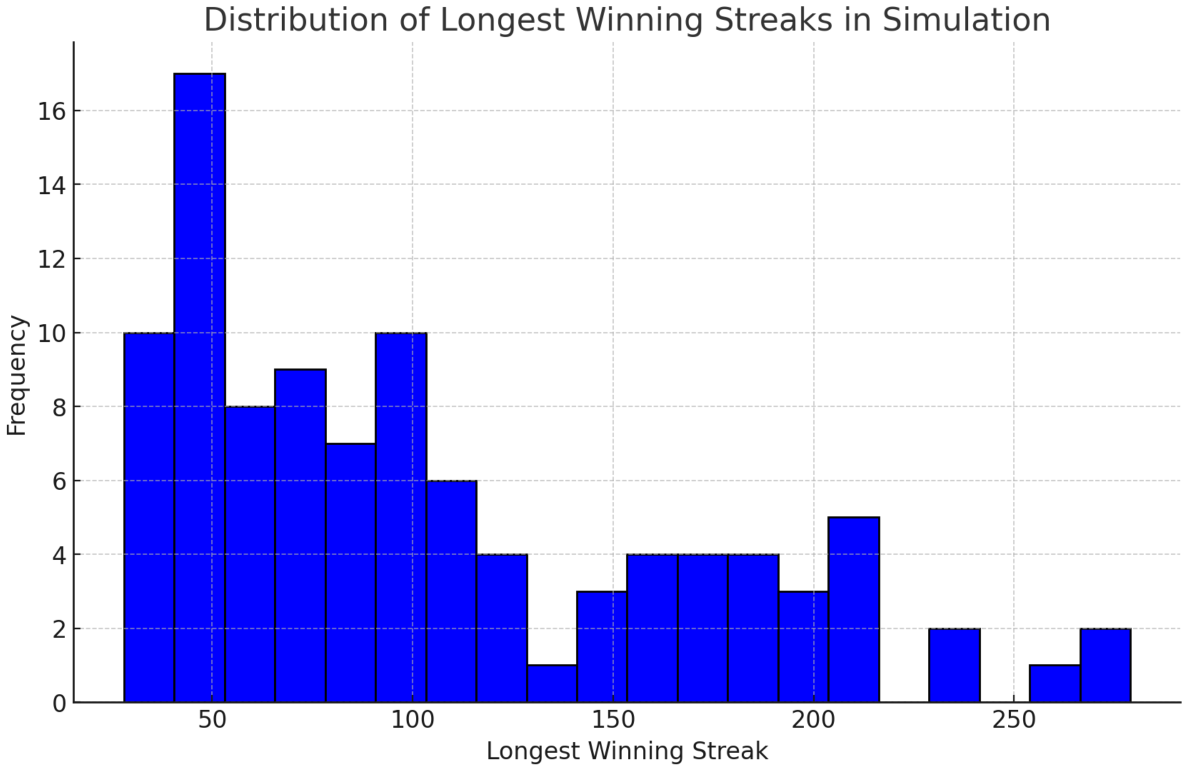
- Number of games per simulation: 25,000
- Number of simulations: 100
- Nakamura's ELO (adjusted for recent performance): 3300
- Opponent's ELO range: 2700 to 3100

- Threshold for significant winning streak: 40 games

The key statistic in this analysis is the longest winning streak. The simulation results, which are analyzed to understand the likelihood of Nakamura's winning streaks occurring by chance, include the mean and median longest winning streaks, as well as the frequency of simulations with winning streaks above the 40-game threshold.

A p-value is calculated to determine the statistical significance of the results, comparing the observed longest winning streak against the distribution from the simulations. The conclusion will be based on whether this p-value is lower than the typical significance level of 0.05.

A histogram is included to visually represent the distribution of the longest winning streaks in the simulations, providing a clear visual interpretation of the data.



Critique of Hypothesis Testing

Hypothesis testing, a staple in statistical analysis, has elicited significant scrutiny over its methodological and interpretational aspects. Key criticisms are outlined as follows:

1. **Persistent Misconceptions**: Common misunderstandings exist regarding the conclusions that hypothesis tests can draw about data and hypotheses. This includes confusion about the implications of rejecting or not rejecting the null hypothesis.

2. **Dependence on Alpha and Beta Levels**: The alpha (α) level, defining the threshold for rejecting the null hypothesis, greatly influences the test results. A lower α reduces Type I error (false positive) probability but raises the risk of Type II error (false negative). High α values increase the test's power—the likelihood of correctly rejecting a false null hypothesis—but at the expense of a greater Type I error risk. Balancing these errors is a critical, yet often overlooked, aspect of hypothesis testing.

3. **Sample Size Influence**: The power of hypothesis testing is closely linked to sample size. Larger samples can detect smaller effect sizes and lower Type II errors, but they may also render minor effects statistically significant, leading to misinterpretations.

4. **Misinterpretation of Statistical Significance**: The significance of a statistical result does not necessarily imply practical or scientific relevance. Statistically significant outcomes may have limited practical value.

5. **Publication Bias**: The tendency to publish studies with significant findings leads to a skewed representation of scientific research. This is exacerbated by the frequent non-publication of studies with non-significant results, omitting valuable data.

6. **Alternative Approaches**: In response to these issues, alternative statistical methods like effect size reporting and confidence interval estimation have gained popularity. These provide more context than a binary significance test and can yield a more detailed understanding of data.

The ongoing debate about the utility and application of hypothesis testing suggests its future lies in a more nuanced approach, integrating its traditional usage with additional statistical methods for a comprehensive analysis.