Universidad Autónoma de Madrid

2023
FINAL PROJECT
TITLE

DATA ANALYSIS

Author: Tania Sánchez Díaz

Master's Degree in Pharmacological Research

INDEX

INDEX OF FIGURES

INDEX OF EQUATIONS

Possible titles:

- Understanding Type I and Type II Errors in Health Research: A Biotechnological Perspective
- Statistical Challenges in Health Research: Managing Type I and Type II Errors

1. HYPOTHESIS TESTING

Hypothesis testing is a crucial part of both empirical research and evidence-based medical practice. Therefore, proper formulation of the hypothesis is key to solving the research question. When talking about the scientific process, the first step is not the observation of the phenomenon, but rather creating a hypothesis that is subsequently verified through observations, experiments, and replications. A well-formulated hypothesis should be simple (one predictor and one outcome), specific and stated in advance, which will help to focus the research on the main objective and establish a firmer basis for the interpretation of the results (Banerjee *et al*., 2009).

Hypothesis testing serves as a statistical method offering a systematic approach for decision-making, utilizing a collection of probabilistic techniques instead of depending on personal judgment (Pereira & Leslie, 2009). Within this framework, statistical tests can be classified as parametric tests (t-test, ANOVA...) or non-parametric tests (Mann-Whitney U-test, Kruskal-Wallis test...). The choice of test depends on the study design and the nature of the data (Akobeng, 2016).

Through hypothesis testing, researchers use their data to assess whether it supports or refutes the predictions made in their research. This is done by employing two types of hypotheses: the null hypothesis and the alternative hypothesis (Bhandari, 2023). The **null hypothesis** holds that there is no association between the predictor variables and the outcome variables within the studied population. In contrast, the **alternative hypothesis** proposes that such an association does exist. It's important to note that the alternative hypothesis cannot be directly tested; rather, it is inferred to be true when the null hypothesis is rejected through statistical tests (Banerjee *et al*., 2009; Pereira & Leslie, 2009).

> To illustrate these concepts, here is an example:
>
> **Context**: a medical research team is investigating whether a new dietary supplement, Supplement Z, improves memory function in older adults.
>
> **Hypothesis**: "Supplement Z enhances memory function in older adults".
>
> In this case:
>
> **Null Hypothesis (H0):** "Supplement Z has no effect on memory function in older adults".
>
> **Alternative hypothesis (H1):** "Supplement Z improves memory function in older adults".

A hypothesis is either true or false (Banerjee *et al*., 2009) based on data from the results of a statistical test. As decisions are based on probabilities, there is always the possibility that the conclusion will be wrong. In this context, it is essential to understand the concepts of Type I and Type II errors (Table 1).

A Type I error occurs when a true null hypothesis is incorrectly rejected, often referred to as a 'false positive' On the other hand, a Type II error happens when a false null hypothesis is erroneously accepted, known as a 'false negative' (Bhandari, 2023). These errors represent the two primary risks of misinterpretation in hypothesis testing, underscoring the importance of careful statistical analysis and decision-making in research.

Table 1. *Type I and type II error (Bhandari, 2023).*

The null hypothesis is…	**TRUE**	**FALSE**
Rejected	Type I error α (false positive)	Correct decision (true positive)
Not rejected	Correct decision (true positive)	Type II error β (false negative)

1.1. TYPE I ERROR

A **Type I error** occurs when researchers mistakenly reject a null hypothesis that is actually true, believing there is a difference between treatment groups when there isn't (this is known as a false positive). Essentially, this means researchers incorrectly conclude one treatment is better than another when it's not. The chance of making this kind of error is the same as the significance level (alpha, α) set for the study (Akobeng, 2016).

The **significance level** is the value established at the start of a study for evaluating the statistical likelihood of achieving the results, known as the p-value (Bhandari, 2023). So, if researchers choose a 5% (0.05) significance level, it indicates they're accepting a 5% risk of being wrong when they think they've found a significant result. While a 5% Type I error rate is commonly accepted as standard, researchers have the option to adopt more stringent criteria, like a 1% error rate (p value < 0,01), for their studies (Akobeng, 2016).

When the p-value of a test is lower than the predefined significance level, it signifies that the results are statistically significant, supporting the alternative hypothesis. On the other hand, if the p-value is higher than the significance level, the results are regarded as statistically non-significant (Bhandari, 2023).

Example: Statistical Significance and Type I Error

In a research study, scientists are evaluating the effect of a new dietary supplement on blood pressure reduction. They compare a group that received the supplement with a group given a placebo. After conducting an analysis using an ANOVA test, they find a p-value of .025. Since this p-value is below the s*et al*pha level of .05, the results are deemed statistically significant, leading to the rejection of the null hypothesis.

However, the p-value of .025 implies that there is a 2.5% probability that the observed results could occur even if the null hypothesis were true - that is, the supplement has no real effect on blood pressure. This indicates that while the findings are significant, there remains a 2.5% risk of committing a Type I error, where the null hypothesis is incorrectly rejected.

The shaded area at the extreme end of the curve symbolizes alpha (α), also known in statistics as the **critical region** (Figure 1). This is the threshold set for determining statistical significance. If the results of a study fall within this critical region, they surpass the boundary of alpha, leading to the null hypothesis being rejected on the grounds of statistical significance. However, this rejection denotes a false positive - a Type I error - because, by the curve's definition, the null hypothesis is, in reality, correct (Bhandari, 2023).

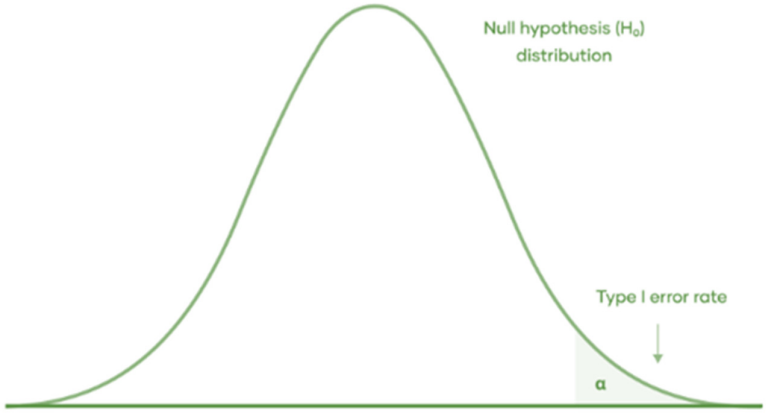

Figure 1. *Probability of making a Type I error (Bhandari, 2023).*

There are several factors that can elevate the risk of a Type I error in research studies. This is often due to the issue of multiple testing, which happens when researchers examine various endpoints and conduct

numerous statistical tests on the same dataset. The likelihood of finding a 'statistically significant difference' by chance increases with the number of tests performed on a dataset. In order to reduce the risk of Type I errors, researchers have three main strategies:

- Firstly, they can avoid multiple testing.
- Secondly, they can opt for a stricter p-value, like 0.01 instead of 0.05, for determining statistical significance. However, this approach could raise the chances of a false-negative result, or a Type II error, unless the study's sample size is increased to compensate for the lower alpha level. A
- Another method is to move away from labeling results as 'statistically significant' or 'not statistically significant.' Instead, researchers could describe their findings as estimates accompanied by confidence intervals (Akobeng, 2016).

The **false positive rate (FPR)**, which quantifies the frequency of type I errors, is directly tied to the calculation of Type I error. By measuring the proportion of false positives in a set of tests, the FPR provides a clear indication of the likelihood of committing a Type I error under specific testing conditions. This connection underscores the importance of accurately calculating and understanding the false positive rate, as it directly impacts the reliability and validity of statistical conclusions (Frost, 2023).

To illustrate these concepts, here is an example of how to calculate the **false positive rate** (type I error) for a particular set of conditions: 1000 test will be done with prevalence of real effects = 0,1; significance level = 0,05 (5% will incorrectly be significant); power = 80% (which means that 80% of the tests will detect that they are true positives).

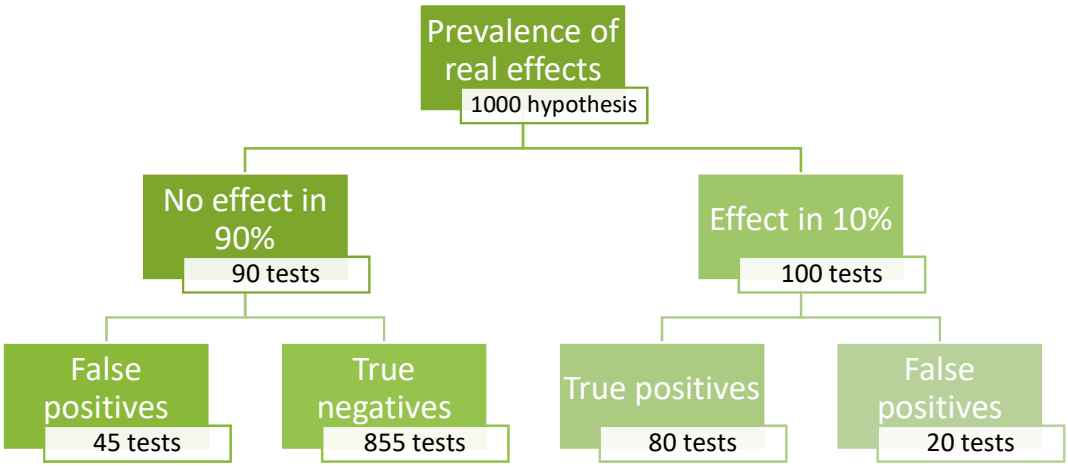

Figure 2. *Flowchart for calculation false positive rate (Frost, 2023).*

1.2. TYPE II ERROR

A **Type II error**, or false negative, refers to the incorrect acceptance of a null hypothesis when it is false. This error occurs when a study fails to detect a real effect or difference due to insufficient power, variability in the data, or inadequate sample size. It results in a 'false negative' conclusion, where researchers erroneously determine that no difference or effect exists when, in reality, it does. The probability of committing a Type II error is represented by the symbol beta (β), which quantifies the risk of overlooking a true effect. The magnitude of β is inversely related to the power of the study; a higher statistical power reduces the likelihood of a Type II error (Bhandari, 2023).

Conventionally, β is set at 20% (0.20), implying researchers are prepared to accept a 20% probability of inaccurately concluding no difference between groups. If the ramifications of a false-negative conclusion

are significant, researchers might opt to lower the acceptable level of β to, for instance, 10% or 5%. Understanding Type II errors is crucial, particularly as numerous medical studies categorically state 'no difference' between interventions without adequately considering the potential for Type II errors.

Statistical power, often simply termed 'power,' is the measure of a statistical test's capability to identify a genuine discrepancy between two distinct groups. The calculation of power is directly linked to the Type II error rate (β), and it is mathematically defined as the inverse of this rate (Equation 1).

Equation 1. *Statistical power equation (Akobeng, 2016).*

$$\text{Power} = 1 - \beta \quad \text{(Eq 1)}$$

Choosing the level of statistical significance (α), as well as β and consequently power (Eq 1), is a subjective decision that researchers must make before beginning a study. It is common practice to aim for a minimum power of 80% (0.8), which is typically recognized as sufficient for most research studies.

The determinants of a study's power are multifaceted, encompassing the prevalence of the outcome under investigation, the size of the expected effect, the architecture of the study, and critically, the number of participants involved. To ensure a study has the potential to accurately answer its research question, it must have an adequate sample size to achieve the requisite power (Akobeng, 2016).

Example: Statistical Power and Type II Error

Imagine you are conducting a study to measure the impact of a new educational program on student test scores. You perform a power analysis and find that given your sample size, your study has an 80% probability (statistical power) of detecting a 10% increase in test scores due to the program. A 10% increase implies that students in the program score 10% higher on average compared to those not in the program.

However, there is a risk of a Type II error if the program's true effect is less than a 10% increase. If the actual improvement is smaller, say 5%, your study might not have enough power to identify this smaller change. In this case, you could mistakenly conclude that the educational program has no effect, when it does have a modest one that your study wasn't able to detect due to insufficient statistical power.

There are different factors that affect the statistical power and, in consequence, the type II error:

- **Effect Size**: the larger the effect, the easier it is to detect.
- **Measurement Error**: both random and systematic errors in data collection can lower the power.
- **Sample Size**: bigger sample sizes decrease sampling error and enhance the power of the study.
- **Significance Level**: a higher significance level can lead to an increase in the study's power (Bhandari, 2023).

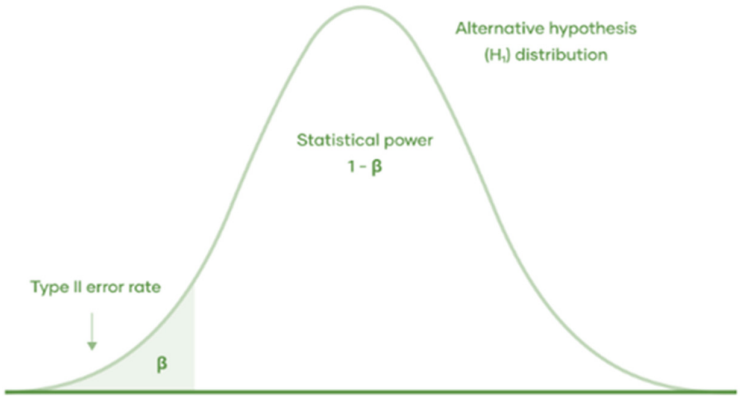

Figure 3. *Probability of making a Type II error (Bhandari, 2023).*

1.3. TYPE I ERROR VS TYPE II ERROR

While Type I and Type II errors are distinct concepts, they are interconnected in the sense that a decrease in one often leads to an increase in the other. For instance, lowering the significance level (say from 0.05 to 0.01) reduces the Type I error but makes it more challenging to refute the null hypothesis, thereby increasing the probability of a Type II error, where a real difference in the broader population might be overlooked.

Conversely, raising the Type I error rate by increasing the level of significance (for example, from 0.05 to 0.10) lessens the chances of wrongly rejecting the null hypothesis. This action, in turn, reduces the likelihood of making a Type II error, or failing to detect a true difference in the larger population (Kaur & Stoltzfus, 2017).

The graph accompanying this explanation, Figure 3, illustrates two curves:

1. The distribution of the null hypothesis represents potential outcomes assuming the null hypothesis is correct. A proper interpretation of any result within this distribution is to maintain the null hypothesis.
2. The distribution of the alternative hypothesis shows potential outcomes if the alternative hypothesis holds true. The correct interpretation here is to reject the null hypothesis.

The areas where these two distributions intersect indicate where Type I and Type II errors can occur. The area on the left denotes alpha, the Type I error rate, while the area on the right signifies beta, the Type II error rate.

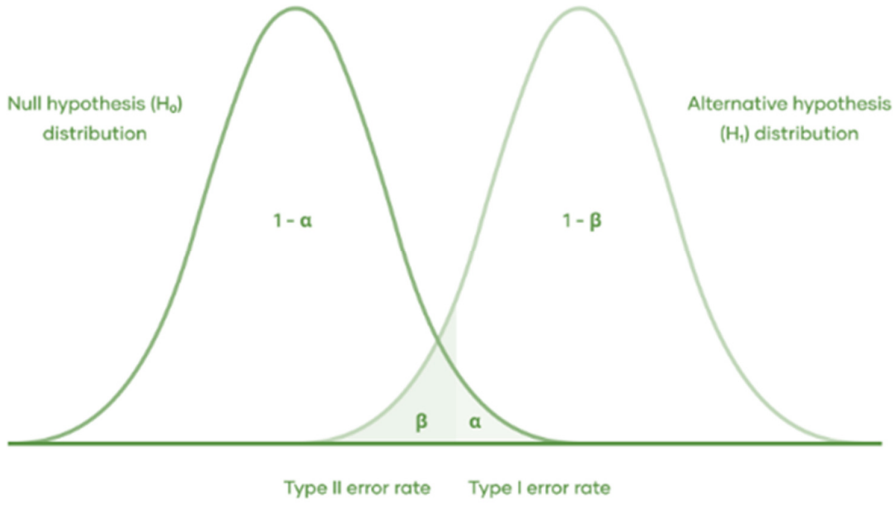

Figure 4. *Probability of making Type I and Type II errors (Bhandari, 2023).*

2. HYPHOTESIS TESTING APPLIED TO HEALTH STUDIES

In the current era of health and medical research, characterized by the collection of extensive data, researchers frequently conduct numerous hypothesis tests to draw multiple inferential conclusions. This process, however, presents challenges in maintaining the reliability of these conclusions due to the potential for statistical errors, especially when multiple tests are conducted. The primary errors of concern are Type I (false positives, where a true null hypothesis is incorrectly rejected) and Type II (false negatives, where a false null hypothesis is not rejected). To address these issues and balance the probability of errors, it's recommended that researchers apply adjustments to the significance levels of their tests. These adjustments help regulate error rates, reduce the likelihood of mistakenly identifying false positives, and

prevent the dismissal of true null hypotheses, thereby enhancing the validity and accuracy of the study's findings (Banerjee *et al*., 2009; Glickman *et al*., 2014).

To address these challenges, statistical methods like the **Bonferroni correction** and the **false discovery rate** have been developed. These methods are crucial tools in health studies, aiding researchers in adjusting their analyses to account for the risks of errors, especially in studies involving multiple comparisons. Understanding these concepts is essential for interpreting results accurately and ensuring that the conclusions drawn from medical research are both valid and reliable.

2.1. BONFERRONI CORRECTION

Consider a hypothetical 3-month study where two groups of patients are randomly assigned to receive either Drug A or Drug B (risperidone or haloperidol which are antipsychotic) to see which is better for certain mental health symptoms and brain functions. The study checks two types of mental health symptoms using two scales and examines five cognitive functions (attention and concentration, visual memory, verbal memory, working memory and ideational fluency), each with its own test. This makes seven different things to compare between the two drugs.

When comparing these outcomes, if there's no actual difference, there's normally a 5% chance of obtaining a misleading result that appears significant but isn't, known as the "false positive" risk. This risk increases with more comparisons; for instance, comparing five aspects might lead to a 23% chance of a false positive. To mitigate this, methods like the Bonferroni correction are used (Andrade, 2019).

The **Bonferroni correction** is a multiple-comparison correction method used in statistical analysis when conducting numerous related or unrelated tests at the same time. The adjustment is necessary because an alpha value that might be suitable for a single test may not be adequate when considering all tests together. To reduce the chance of false positive results the alpha threshold is lowered proportionally to the total number of comparisons (Wolfram, n.d.).

In research, the Bonferroni correction is often employed to adjust p-values during numerous statistical tests across various contexts. This approach is commonly applied in diverse experimental scenarios, including (1) comparing different groups at the start of a study, (2) exploring the connections between various variables, and (3) assessing multiple outcomes or endpoints in clinical trials. This highlights its significance in modern experimental methodologies.

The Bonferroni correction addresses the increased likelihood of a Type I error in multiple statistical tests. For example, using a p-value threshold of ≤ 0.05 across all tests might result in a false positive in 1 out of 20 trials. However, with 20 tests and the null hypothesis (Ho) true for all, the chance of at least one test being significant is not 5% but about 64%. The general formula for this error rate, described in Equation 2, adjusts the significance level to alpha/T to maintain an overall alpha level of 0.05.

Equation 2. *Error rate* (Armstrong, 2014).

$$\text{Error rate} = 1 - (1 - \alpha) / T \quad \text{(Eq 2)}$$

Despite its frequent use, the Bonferroni method has been debated. Some believe no correction is necessary, while others consider it essential. Criticisms include its potential to impede accurate statistical inference, its focus on the 'universal' null hypothesis, and its dependency on the number of tests. Additionally, reducing the probability of a Type I error increases the risk of a Type II error, possibly overlooking real differences. This becomes more significant with more tests, as the required p-value for significance decreases, reducing the test's power. Another debate is over the scope of tests for which the correction should apply, like whether it includes all tests in a report, a subset, or even tests from the same dataset in other reports (Armstrong, 2014).

2.2. FALSE DISCOVERY RATE (FDR)

In genome-wide studies, where typically thousands of hypothesis tests are performed at once, applying the traditional Bonferroni approach for multiple comparison correction can be overly stringent. This heightened caution against false positives often results in overlooking numerous significant findings. To maximize the detection of significant comparisons while still keeping a low rate of false positives, researchers increasingly rely on the False Discovery Rate (FDR). These method offer a more balanced approach to identifying meaningful results in large-scale data analysis (https://www.publichealth.columbia.edu/research/population-health-methods/false-discovery-rate).

The **False Discovery Rate (FDR)** is a statistical method applied in situations involving multiple hypothesis testing. Its primary function is to adjust for multiple comparisons ("Encyclopedia of Systems Biology," 2013).

The FDR (Equation 3) represents the expected fraction of tests where the null hypothesis is actually true. The primary aim of controlling the FDR is to establish significance levels for a group of tests in a manner that ensures the ratio of accurate null hypotheses among those deemed significant remains below a predetermined threshold (Glickman *et al.*, 2014).

> ***Equation 3.*** *Formula false discovery rate. The total count of null hypothesis rejections comprises both false positives (FP) and true positives (TP) (Benjaminit & Hochberg, 1995).*
>
> $$FDR = \frac{FP}{FP+TP}$$ (Eq 3)

The FDR control method has become increasingly prominent in fields requiring extensive testing, such as genomic research and micro-array data analysis. Its adoption extends to healthcare, where it's applied in evaluating healthcare providers and assessing clinical adverse event rates. FDR control offers significant advantages over traditional Bonferroni-type adjustments, primarily due to its increased power in statistical testing. This method enables researchers to make more reliable and well-calibrated inferences from large datasets, particularly in health research that involves mining large databases and analysing detailed health information.

Unlike the Bonferroni method, which can be overly conservative and thus limit the detection of true effects, FDR control is adept at identifying significant findings while maintaining a low rate of false positives. However, its application is not without challenges, especially in single test scenarios. Implementing FDR control in these contexts is complex, as it depends on the ability to estimate the probability of a null hypothesis being true before testing. While Bayes theorem could theoretically assist in applying FDR control by determining an appropriate significance level cutoff, in many single test scenarios where this assessment is difficult, not adjusting significance levels might be the most objective course of action. This underscores the need for careful consideration when applying FDR control to ensure its appropriate use in the specific context of the research design and objectives (Glickman *et al.*, 2014).

Question: Is the false positive rate the same as the false discovery rate?

No, the false positive rate represents the likelihood of incorrectly rejecting a true null hypothesis, whereas the false discovery rate measures the chance that a null hypothesis is actually correct even after it has been rejected (Glickman *et al.*, 2014).

3. CONCLUSIONS

In summary, this project has highlighted the crucial aspects of hypothesis testing in data analysis, particularly focusing on Type I and Type II errors and their implications in health studies. It underscored the delicate balance between these errors and the need for careful statistical decision-making. The introduction of methods like the Bonferroni correction and the False Discovery Rate (FDR) illustrated strategies to control error rates in multiple hypothesis testing scenarios, a common challenge in health research. While the

Bonferroni method is conservative, FDR provides a more nuanced approach, especially beneficial in large-scale data analysis. This exploration reaffirms the necessity of understanding statistical errors in hypothesis testing, an essential component for advancing accurate and reliable scientific research, particularly in health-related fields.

4. BIBLIOGRAPHY

Akobeng, A. K. (2016). Understanding type I and type II errors, statistical power and sample size. In Acta Paediatrica, International Journal of Paediatrics (Vol. 105, Issue 6, pp. 605–609). Blackwell Publishing Ltd. https://doi.org/10.1111/apa.13384

Andrade, C. (2019). Multiple Testing and Protection Against a Type 1 (False Positive) Error Using the Bonferroni and Hochberg Corrections. Indian Journal of Psychological Medicine, 41(1), 99–100. https://doi.org/10.4103/ijpsym.ijpsym_499_18

Armstrong, R. A. (2014). When to use the Bonferroni correction. In Ophthalmic & physiological optics : the journal of the British College of Ophthalmic Opticians (Optometrists) (Vol. 34, Issue 5, pp. 502–508). https://doi.org/10.1111/opo.12131

Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J., & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. Industrial Psychiatry Journal, 18(2), 127. https://doi.org/10.4103/0972-6748.62274

Bhandari, P. (2023, June 22). Type I & Type II Errors | Differences, Examples, Visualizations. Scribbr. Retrieved November 27, 2023, from https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Benjaminit, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. In J. R. Statist. Soc. B (Vol. 57, Issue I).

Encyclopedia of Systems Biology. (2013). In Encyclopedia of Systems Biology. Springer New York. https://doi.org/10.1007/978-1-4419-9863-7

False discovery rate. (2023, March 10). Columbia University Mailman School of Public Health. https://www.publichealth.columbia.edu/research/population-health-methods/false-discovery-rate

Frost, J. (2023, July 7). P-Values, error rates, and false positives. Statistics by Jim. [Internet] https://statisticsbyjim.com/hypothesis-testing/p-values-error-rates-false-positives/

Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. In Journal of Clinical Epidemiology (Vol. 67, Issue 8, pp. 850–857). Elsevier USA. https://doi.org/10.1016/j.jclinepi.2014.03.012

Kaur, P., & Stoltzfus, J. (2017). Type I, II, and III statistical errors: A brief overview. International Journal of Academic Medicine, 3(2), 268. https://doi.org/10.4103/ijam.ijam_92_17

Pereira, S. M. C., & Leslie, G. (2009). Hypothesis testing. Australian Critical Care, 22(4), 187–191. https://doi.org/10.1016/j.aucc.2009.08.003

Wolfram Research, Inc. (n.d.). Bonferroni Correction -- from Wolfram MathWorld. https://mathworld.wolfram.com/BonferroniCorrection.html