ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

Scuola di Ingegneria
Corso di Laurea Magistrale in Ingegneria Informatica
School of Engineering - Master's Degree in Computer Engineering

CHALLENGES IN STATISTICAL VALIDATION: INTRODUCING PRACTICAL SIGNIFICANCE PROBABILITY AS AN ALTERNATIVE

Thesis in
Data Mining M

Supervisor
Chiar.mo Prof. Ing. Claudio Sartori

Co-supervisor
Chiar.ma Prof.ssa Stefania Mignani

Presented by
Andrea Borghesi

Third Graduation Session
Academic Year 2023 – 2024

Keywords

Practical Significance
Statistical Tests
P-Value
Abandon Statistical Significance
Replication Crisis

Ai miei amati genitori.

-

To my beloved parents.

Introduction

For decades, researchers have relied on tools such as P-value and Null Hypothesis Significance Testing (NHST) to determine whether their findings are statistically significant. However, many challenges have emerged with these traditional methods. Common misunderstandings, arbitrary thresholds (e.g., p-value < 0.05), and oversimplified interpretations of complex evidence have contributed to issues such as the replication crisis [1][2][3]. These problems are not solely due to P-value misuse but also due to the misuse of other statistical methods [1][4]. A recent systematic review found that 31% of 1,579 Bayesian articles in psychology failed to specify the priors used in their analyses [5].

The P-value and NHST were introduced by Fisher (1925)[6] and Neyman & Pearson (1933)[7], respectively. Criticism of these methods is longstanding (e.g., Berkson 1942 [8]). In 2001, Sellke, Bayarri, and Berger demonstrated that a p-value of 0.05 could correspond to a false discovery rate of approximately 29% [9][10]. In 2018, 73 statisticians proposed to redefine statistical significance by lowering the p-value threshold to 0.005 [11], while others recommended 0.001 [10].

Statistical tests are important in many research fields, including medicine, psychology, economics, and technical areas like artificial intelligence. In machine learning (ML) studies, they are often used to judge how well models work and whether they can generalize. Verifying that ML models are reliable in real-world situations is crucial. Moreover, P-value, NHST and confidence intervals are commonly employed to compare how different ML algorithms perform and determine which one is best.

The primary issue is the widespread misinterpretation of statistical concepts such as P-value, confidence intervals, and statistical power. In 2015, the journal *Basic and Applied Social Psychology* (BASP) banned P-value [12], arousing mixed reactions [13]. In 2016, the American Statistical Association (ASA) released guidelines on P-value usage, urging researchers to avoid the term "statistically significant" [14]. In 2019, the paper *Retire Statistical Significance*[15],

signed by 800 authors, called for an end to the conventional use of P-value.

Despite efforts to find alternatives, such as confidence intervals, effect sizes, Bayesian methods, and new metrics (e.g., S-value [16], SGPV [17], MESP [18], EP [19]), none have fully replaced the P-value. As Cohen (1994) noted: *"Don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist."*[20].

In this study, we reviewed and analyzed key research papers to identify the most important critiques and recommendations. Our goal was to design a new metric called **Practical Significance Probability (PSP)**, which estimates the probability that an effect exceeds a predefined practical significance threshold. PSP is easy to learn and interpret, and shifts the focus from statistical significance to practical or scientific significance [21][22]. It is not a "magic alternative" to NHST. Instead, PSP should be used alongside other methods, as discussed in the final chapter of this thesis. Details of this method are provided in Section 2.2.
The PSP can be used in ML studies to check if a model's performance score exceeds a minimum level of practical relevance. As a result, the methods discussed in this thesis are not only relevant to traditional statistical analysis, but also play a key role in supporting the reliability and interpretability of ML findings.
In Chapter 4, we compare PSP to other statistical methods through an empirical analysis based on simulations and inspired by Goodman's research [18].

Thesis Structure

This thesis is structured as follows:

- **History of Statistical Significance**: We begin with a review of the history of statistical significance, the ASA's statement on p-values and the challenges in statistical interpretation and replicability.

- **Theoretical Background**: We then present common statistical methodologies and introduce the PSP. A detailed mathematical derivation of PSP is provided, along with its underlying assumptions and limitations.

- **Comparative Analysis of Statistical Validation Techniques**: This chapter compares existing statistical validation techniques, highlighting their strengths and weaknesses.

- **Empirical Analysis Supporting PSP**: Simulation studies and empirical comparisons demonstrating PSP's performance.
- **Code**: Reproducible code for simulations and a SciPy-like implementation of PSP.
- **Conclusions**: Final recommendations for reliable statistical analysis.

Table of Contents

List of Figures

Chapter 1

History of Statistical Significance

1.1 Fisher, Neyman-Pearson, Cohen

For decades, null hypothesis significance testing (NHST) has been a topic of debate among researchers. The foundations of this method can be traced back to Ronald Fisher, who had developed many of its key principles by 1925 [6]. A few years later, Jerzy Neyman and Egon Pearson introduced the concepts of type I and type II errors, along with the idea of setting a predetermined significance level, which led to the well-known hypothesis testing that have been used since then [7].
Despite its widespread adoption, NHST has faced criticism from early on. In 1938, Joseph Berkson published one of the first major challenges to its logic and usefulness [23][8], starting a discussion that continues to this day.

Fisher himself recommended not using the significance test alone, but also measuring the strength of the correlation between variables through analysis of variance.
Throughout the years, many other methodologies have been proposed, but only a few have been widely adopted by authors. A study [21] has found that the confidence interval [24] and Cohen's d [25] were the most common approaches along with NHST.
Cohen's d was the first statistical measure to be explicitly recognized as an effect size. The peculiarity of Cohen's work was not just defining d, but also offering practical guidelines for interpreting its magnitude. In his work, he described a medium effect (0.5) as something that a careful observer could detect without statistical analysis. A small effect (0.2), while noticeably less than medium, was still meaningful rather than negligible. A large effect (0.8), on the other hand, was positioned symmetrically above medium.

In 1994, Cohen's famous paper "The Earth Is Round ($p < .05$)" [20] highlighted the core issues regarding NHST and recommended abandoning the ritualistic p-value < 0.05. He also noted: "Don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist.".

1.2 ASA's Statement on P-Value

In 2016, the *American Statistical Association* (ASA) - the world largest community of statisticians - has published the "ASA Statement on P-value" [14], which is considered a milestone in the hypothesis testing discussion (cited more than 7300 times as of the time of this thesis).
The authors stated that "the statistical community has been deeply concerned about issues of reproducibility and replicability of scientific conclusions" and that the "misunderstanding or misuse of statistical inference is only one cause of the reproducibility crisis."
The statement suggests moving away from rigid declarations of "statistical significance." A P-value alone cannot determine the existence or importance of an association or effect. Similarly, we should stop using confidence intervals to make binary decisions based on whether a null value falls within the interval.

Along with the ASA's statement, another important paper entitled "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations" by Greenland et al. [1], gathered the most common misinterpretations of P-value, confidence intervals and statistical power. The main *misinterpretations* are the following:

- **The P value represents the probability that the test hypothesis is true.** Incorrect! The P-value assumes the hypothesis is true and only measures how much the data deviates from predictions, not the likelihood of the hypothesis itself.
- **A significant result ($P \leq 0.05$) means the hypothesis is false or should be rejected.** Incorrect! A low P-value suggests unusual data under the assumptions but does not prove the hypothesis false, it could result from errors or assumption violations.
- **A nonsignificant result ($P > 0.05$) means the hypothesis is true or should be accepted.** Incorrect! A large P-value does not confirm the hypothesis, random errors or faulty assumptions might cause it.
- **A large P-value is evidence in favor of the test hypothesis.** Incorrect! Any P value below 1 indicates that the tested hypothesis is not

the most compatible one; a larger P value suggests better compatibility.

- **Statistical significance implies scientific or substantive importance.** Incorrect! A low P-value only flags unusual data under model assumptions. Confidence intervals should be consulted to assess the practical importance of results.
- **The P-value is the probability of obtaining our data if the hypothesis is true.** Incorrect! The P-value depends on all model assumptions, including randomness and unbiased selection, not just the hypothesis.
- **If P $\leq$ 0.05 leads you to reject the hypothesis, then there's only a 5% chance your result is a false positive.** Incorrect! The 5% threshold only means that, over many tests, false rejections would occur 5% of the time under correct assumptions. Another study showed that a 0.05 p-value correspondss to a 29% false discovery rate [9].
- **P-value should be reported as inequalities, such as "P $<$ 0.02" or "P $>$ 0.05".** Incorrect! Exact P values provide more clarity and help in interpreting results accurately compared to vague inequalities.

The paper also presented the most common misinterpretations for confidence intervals and statistical power:

- **"A 95% confidence interval has a 95% chance of containing the true effect size."** Incorrect! A specific confidence interval represents a fixed range between two numbers, such as 0.72–2.88. The probability that this interval contains the true effect size is either 100% or 0%, depending on whether the true effect is within the interval. The 95% refers to the long-term frequency of intervals containing the true effect if computed from many studies, assuming the model's assumptions are correct.
- **"If two confidence intervals overlap, the difference between estimates is not significant."** Incorrect! Confidence intervals from different studies can overlap, yet a test for the difference between them could still yield P $<$ 0.05. It can be noted that if the 95% confidence intervals do not overlap, P will be less than 0.05 for the difference, assuming the same conditions used to compute the intervals. Conversely, if one confidence interval contains the point estimate of another, then P $>$ 0.05.
- **"A 95% confidence interval predicts that 95% of future study estimates will fall within this interval."** Incorrect! The 95% confidence level refers to the frequency with which newly observed intervals

will contain the true effect size. Despite these misinterpretations, many researchers consider confidence intervals more informative than P values, as they shift the focus from a single hypothesis to the range of effect sizes compatible with the data. When discussing practical implications of a study, we should consider all the possible effect sizes in the range of the confidence interval.

- **"If the P value exceeds 0.05 and the test has 90% power, then the chance of a false negative is 10%."** Incorrect! If the null hypothesis is false and one accepts it, the error rate is actually 100%, not 10%. The 10% figure refers only to how often the test would incorrectly accept the null hypothesis over many repetitions, assuming all other assumptions are true.

Another paper published on ASA entitled "Moving to a World Beyond p<0.05" [26] summarized the main rules to follow in a clear way:

- *Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the p-value passed some arbitrary threshold such as $p < 0.05$).*
- *Don't believe that an association or effect exists just because it was statistically significant.*
- *Don't believe that an association or effect is absent just because it was not statistically significant.*
- *Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.*
- *Don't conclude anything about scientific or practical importance based on statistical significance.*

ASA did not set strict rules on which statistical methods to use, but they provided important advice on how to interpret data correctly. To understand results properly, researchers should consider effect sizes and confidence intervals, which give more context about the strength and reliability of findings.
It is also important to think about the assumptions behind statistical methods. Every analysis is based on certain rules and conditions, and ignoring them can lead to mistakes. Additionally, P-value and confidence intervals alone cannot prove whether something is truly happening or not. Statistical tests only show probabilities based on the data, but real conclusions require extra information. One way to include this extra knowledge is through Bayesian methods, which

help make sense of data by combining it with what is already known.
ASA also encourages researchers to focus on collecting high-quality data and to use multiple ways of analyzing it instead of relying on just one method. Since real-world data is often messy and unpredictable, it is important to accept that uncertainty is a natural part of research. Results can change slightly each time an experiment is repeated, so researchers should always show how much uncertainty exists in their findings by including things like standard errors or confidence intervals.
ASA incorporates these concepts with the following sentence:

> "Accept uncertainty. Be thoughtful, open, and modest."

1.3 Challenges in Statistical Interpretation and Studies Replicability

The "replication crisis" derives from the fact that researchers often struggle to get the same results when they repeat a study. We indicate that with the term **replicability**: it means that if someone else does the same experiment, they should get similar results. More recently, another important concept called **reproducibility** has become a basic requirement in research. Reproducibility means that if someone has the same data and knows the steps used in the analysis, they should be able to get the same results. Bad statistical analysis have contributed to this crisis, although they are not the only reason [27].

A recent observational study has analyzed the statistical section of about 120,000 papers and studies using topic modeling techniques [3]. The researchers found that these sections often had recurring boilerplate text, reflecting a mechanistic approach to statistical reporting. In particular, around 13% of the papers included a variation of the phrase "a p-value < 0.05 was considered statistically significative". This sentence goes against the recommendations from the ASA on the correct use of P-value.

Using only the NHST to validate a research's finding is unreliable. A study examined the relationship between p-values and evidence against null hypotheses, demonstrating that a p-value of 0.05 can correspond to a **minimum false discovery rate** of approximately 29% (to be optimistic) [9]. In other words, a study that relies solely on a $P < 0.05$ has 1 in 3 chance of supporting a false claim.

As we have seen in the previous section, the replication crisis fault must

not be attributed only to P-value. In fact, also confidence intervals and statistical power are often misunderstood. Moreover, a recent systematic review found that 31% of 1,579 Bayesian articles in psychology failed to specify the priors used in their analyses [5].

In 2015, the journal *Basic and Applied Social Psychology* (BASP) banned P-value for the first time [12]. Also in this case, no strict replacement has been proposed as statical methods, the editors said: "BASP will require strong descriptive statistics, including effect sizes. We also encourage the presentation of frequency or distributional data when this is feasible. Finally, we encourage the use of larger sample sizes.".
This event has sparked mixed reactions. A 2019 study[13] examinating papers published after the ban has found that authors were making less proven claims. This, in turn, increases the risk of more non-reproducible effects. The study's authors argue that banning P-value could make publications worse.
Daniel Lakens (2021) pointed out that there's no strong evidence that removing P-values and hypothesis testing would make research better [28]. Hanson (1958) [29] found that research findings were more likely to be replicated when they followed clear confirmation rules, like using a 5% significance level. In his study, more than 70% of findings with such rules were later confirmed by other researchers, while less than 46% of findings without clear rules were confirmed.

Between the large number of papers that have contributed to the replication crisis, there are some that became famous for the consequences that they had. In 2006 a group of researchers published a paper claiming that they had build an algorithm that predicted which cancer patients would respond to chemotherapy [27]. When other statisticians have attempted to reproduce the study, they found a poorly conducted data analyses. Only 5 years later, in 2011, the original study was retracted.

Another well-known example comes from the 2010 paper "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance"[30], which claimed that adopting "high-power" poses for just a few minutes could significantly alter hormone levels (increasing testosterone and decreasing cortisol) and improve risk-taking behavior and feelings of power. In 2012, one of the authors, Amy Cuddy, presented a TED Talk based on this study, which became one of the most-watched talks ever, spreading the idea that simple changes in posture could dramatically affect success. However, in 2015, a replication study has been attempted by Ranehill et al.[31] on a larger sample and found no significant effect, directly challenging the original findings. A year later, Dana Carney, the first author of the original study, publicly stated that she no

longer stood by its conclusions [32].

1.4 From Statistical Significance to Practical Significance

In 2018, 73 statisticians proposed redefining statistical significance by lowering the p-value threshold to 0.005 [11]. Colquhoun recommended even a lower threshold of 0.001 [10]. In 2019, the papers "Abandon Statistical Significance" [33] (published on ASA) and "Retire Statistical Significance" [15] (signed by 800 authors) called for a drop of the NHST paradigm.

Over the years many alternatives have been proposed, but none of them have been widely adopted. We will analyze and compare them in Chapter 3. Ho et al. [34] suggested to use estimation plots in the descriptive analysis. Estimation plots show the actual size of differences between groups along with confidence intervals, making it easier to understand the results and their uncertainty.

Many authors agreed that effect sizes and confidence intervals should be included to validate the practical implications of a study.
While effect sizes are not synonymous with practical significance, they could represent a good starting point. Authors should find and report effect sizes whenever possible [22].
The concept of practical significance is not recent; Kirk in 1996 gave a good definition [21]:

> **Statistical significance** is concerned with whether a research result is due to chance or sampling variability; **practical significance** is concerned with whether the result is useful in the real world.

1.5 Practical Significance Probability (PSP): an overview

Alongside common effect size measures like Cohen's d, other approaches have been proposed to assess practical significance. Confidence intervals can also be useful for this purpose if the following principles are followed [35]:

- Considering both the upper and lower limits and their different practical implications.

- Not focusing on whether the interval includes the null value.
- Recognizing that the interval itself is an estimate subject to error.

Some newer methods define a threshold for the minimum effect size considered practically useful. Examples include equivalence testing [35] and MESP [36].

In this thesis, we analyze key critiques and recommendations to develop a new method that follows these principles:

- Combining effect size and uncertainty.
- Prioritizing practical relevance.
- Improving interpretability.
- Avoiding strict accept-or-reject decisions.

Building on these ideas and existing approaches, we introduce **Practical Significance Probability (PSP)**. PSP estimates the probability that an effect exceeds a predefined practical significance threshold. It is not meant to replace NHST but rather to complement other methods. Further details on PSP are provided in Section 2.2.

Chapter 2

Theoretical Background

In this chapter, we provide a concise theoretical description of the most common statistical methods and an exhaustive explanation of PSP.

2.1 Common Statistical Methodologies

2.1.1 Null Hypothesis Significance Testing (NHST) and P-Value

We have seen in the previous chapter that the **P-value** and **Null Hypothesis Significance Testing (NHST)** have been introduced by Fisher (1925) and Neyman-Pearson (1933) papers, respectively.
NHST is a method used to analyze data and determine if there is enough evidence to reject a default assumption called the **null hypothesis**. This default assumption, represented as H_0, usually states that there is no effect or no difference. The **alternative hypothesis**, H_1, suggests that an effect exists. To evaluate this, a test statistic is calculated from the data, which is then compared to what would be expected if H_0 were true.

A key part of NHST is the **significance level**, denoted as α, which is the threshold for rejecting H_0. Researchers commonly use $\alpha = 0.05$, meaning that if the probability of getting the observed data under H_0 is less than 5%, they reject H_0 in favor of H_1. However, this cutoff is arbitrary, and strict adherence to it can lead to misinterpretations and more recently a lower threshold of 0.005 or 0.001 have been recommended by other authors.

The P-value is a measure used in NHST to assess how well the data aligns with the null hypothesis. It represents the probability of obtaining a test result at least as extreme as the one observed, assuming that H_0 is correct. A small

P-value suggests that the observed data is unlikely under H_0. However, a large P-value does not confirm H_0; it simply indicates that the data is not strongly incompatible with it.
People often mistake the P-value for the probability that an effect θ is true given the data, $p(\theta|\text{data})$, which is what researchers actually want to know [21]. However, the P-value actually represents the probability of observing the data (or more extreme data) assuming the effect and model assumptions are correct, $\boldsymbol{p(\text{data}|\theta)}$. To quote Greenland et al. [1]:

> The P value is a statistical summary that measures the compatibility between observed data and what we would expect under the full model.

The P-value can be converted to a continuous value called Shannon information (**S-value**). This is done by applying the formula $s = -\log_2 p$ where p is the p-value. The S-value measures the amount of information that the test supplies against the hypothesis [16]. Higher S-values mean stronger evidence against the null hypothesis. Fricker [13] suggests to use S-value with confidence intervals instead of P-value.

2.1.2 Confidence Interval

A **confidence interval** is a way to estimate the possible range of an effect in a study. Instead of just testing whether an effect exists (like in NHST), a confidence interval gives us a range of values that are more in line with what was actually observed in the data. For example, if we compare two treatments and find a confidence interval of 10.0 to 20.0, this means that based on the data, the true effect is likely within this range, assuming our statistical model is correct. The key idea is that this interval includes values that are more compatible with the data than those outside the interval.
A 95% confidence interval means that if we repeated the same kind of study many times, 95% of the calculated intervals would contain the true effect size. However, it does not mean that any single confidence interval has a 95% chance of containing the true effect: intuitively, the probability that this interval contains the true effect size is either 100% or 0%, depending on whether the true effect is actually within the interval [1]. Confidence intervals give a fuller picture by showing a range of plausible values for the effect. This is why many journals now require them in research papers. By using confidence intervals, we can better understand the uncertainty in our estimates rather than just focusing on whether an effect is statistically significant.

2.1.3 Statistical Power

Statistical power tells us how likely a study is to detect a real effect if one truly exists. Before conducting a study, researchers estimate power to understand the probability that their test will find evidence against the null hypothesis (e.g., show a P-value below 0.05) when the alternative hypothesis is actually correct. For example, if a study has 80% power, this means that if there is a real effect, there is an 80% chance that the study will detect it. However, there is still a 20% chance (1 - power) that the study will miss the effect and fail to reject the null hypothesis. This is called a **Type II error** (or *beta error*).
Power is calculated before a study begins, using estimates of the expected effect size. Calculating power from observed data is just another way of looking at the P-value, so it does not provide new evidence about the effect.

Even when a study is designed with 80% power, real-world issues like low participant recruitment can reduce the actual power. Also, even if two separate studies each have 80% power, the chance that both will show statistically significant results is only 64% (0.80×0.80). This means that even well-powered studies can sometimes appear to contradict each other, leading to confusion in research findings.
Despite these limitations, power is still useful in planning studies and understanding why replication attempts may fail, even when an effect is real. In medical research, grant agencies commonly require sample sizes that yield statistical power of at least 80% [37].

2.1.4 Effect Size

Effect size tells us how big or meaningful an observed difference or relationship is, rather than just whether it exists. By focusing on magnitude, effect size helps us understand whether a finding has practical relevance. Several types of effect sizes can be used, depending on the nature of the data and the research question. Common examples include **Cohen's *d*** for comparing group means, the correlation coefficient (e.g., Pearson's r) for measuring the strength of linear relationships, and odds ratios in studies of categorical outcomes. By standardizing results, these metrics allow comparisons across different studies and different measurement scales.
Effect size is often reported together with a confidence interval, providing a range for the magnitude of the observed effect. This approach gives a clearer picture of both the strength of the evidence (through the interval) and the size of the effect itself.

Effect size estimates are especially important in meta-analysis, where findings from multiple studies are combined, and in power analyses, which rely on estimating plausible effect sizes to determine how many participants or observations are needed for a study to detect a meaningful difference.

A 2011 study[36], based on an empirical analysis, has found that a large effect size tends to correspond to low p-values, and small effect sizes tend to correspond to large p values. However, a p-value of 0.01 can correspond to effect sizes ranging from about 0.2 to 1, and an effect size close to 0.5 can correspond to p-values ranging from about 0.001 to 0.05.

2.1.5 Equivalence Testing and SGPV

When researchers want to show that an observed effect is too small to matter in practical terms, they can use **equivalence testing** (often indicated with TOST, i.e. two one-sided tests), which goes beyond simply determining if an effect differs from zero (as in NHST). In equivalence testing, a range of values around zero is specified to represent what is considered *not meaningfully different* from zero. If the entire confidence interval for a study's effect estimate falls within this predefined range, it suggests that the effect is so small it can be treated as practically equivalent to zero [35].

Second Generation P-Values (SGPV) [17][35] are another tool for assessing whether observed data sufficiently exclude meaningful effects. Like equivalence testing, SGPVs involve defining a range of effects that would be considered negligible. The SGPV then measures the overlap between this "null range" and the range of values supported by the data (often a confidence interval). If the confidence interval lies entirely within the null range, the SGPV equals 1. This implies the data are fully compatible with an effect so small that it is of no practical concern. If the confidence interval lies completely outside the null range, the SGPV equals 0. This means the data support an effect size that is larger than what we consider negligible. SGPVs between 0 and 1 signal that the current data do not allow a clear conclusion about equivalence versus a meaningful effect.

In practice, both the TOST and SGPV rely on comparing data to a specified smallest effect size of interest. In many cases, these two methods lead to very similar conclusions. However, the SGPV's utility can be limited if confidence intervals are asymmetric or broader than the equivalence range [35].

2.1.6 Bayesian Methods

Bayesian approaches offer an alternative framework to the traditional frequentist perspective by treating unknown parameters as random variables with probability distributions.
This framework begins with a **prior distribution**, which encodes the researcher's initial beliefs about the parameter (e.g., the true mean difference between two groups), before observing data. Once data are collected, the prior distribution is updated using **Bayes' theorem** to produce a **posterior distribution**, which reflects the updated beliefs after accounting for the observed results. Researchers then typically summarize uncertainty about the parameter through **credible intervals**. Unlike frequentist confidence intervals, these credible intervals can be interpreted as having a certain probability (e.g., 95%) of containing the true parameter value, assuming the model and priors are correct [38].

Bayesian analysis can be a useful alternative to traditional statistical methods like significance testing and equivalence testing, especially when researchers want to include prior knowledge in their analysis. Significance tests and confidence intervals alone cannot definitively prove whether an effect exists or not. Bayesian methods allow researchers to directly incorporate prior information into their statistical models[1], making them a preferred choice in many cases. However, full Bayesian analysis will probably never be adopted as a substitute for P-value because it is too complex for most users [39].

2.2 Introducing Practical Significance Probability (PSP)

After reviewing the state-of-art statistical tests (discussed in Chapter 3) and synthesizing insights from various studies and research papers, we developed a new approach to address some of their limitations. We propose a method that integrates effect size, uncertainty and practical relevance called **Practical Significance Probability (PSP)**. The PSP quantifies the probability that the true effect size exceeds a predefined **practical significance threshold** (**PST**) given the observed data. The PST represents the minimum effect size considered of practical importance and must be chosen by domain experts of the specific study.
The purpose of this proposal is not to replace existing statistical methods. As Cohen [20] said in 1994:

> Don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist.

Instead, the PSP offers a different approach that can help in understanding the significance of the results. It must be used alongside other statistical methods and a complete descriptive analysis, where results and assumptions specific to the study or research are well documented.

This approach provides a single, interpretable metric that addresses some of the limitations associated with the traditional P-value. With this method, we embrace the shift from statistical significance to practical significance (see Section 1.4) as recommended by many authors.

The PSP is calculated using the following formula:

$$PSP = 1 - \Phi\left(Z\right) \quad \text{with } Z = \frac{\delta - \hat{\theta}}{\mathrm{SE}_{\hat{\theta}}}$$

where

- $\hat{\theta}$ is the **observed effect size** (e.g., absolute mean difference),
- δ is the **practical significance threshold** (i.e., the smallest effect size of interest),
- $\mathrm{SE}_{\hat{\theta}}$ is the **standard error** of the estimated effect size,
- Φ is the **cumulative distribution function (CDF)** of the standard normal distribution.

The Z-score represents the number of standard errors by which the observed effect size exceeds the practical significance threshold. By calculating $1 - \Phi(Z)$, PSP provides the probability that the true effect size θ is greater than the practical significance threshold δ, given the observed data.

PSP uses the same basic formulas (z-scores, standard errors) as traditional **frequentist methods** but frames them in a way that looks like a Bayesian posterior probability statement: "the probability that θ exceeds δ, given the data.":

$$P(\theta > \delta | data)$$

Unlike a purely Bayesian approach, PSP does not require specifying a prior; instead, it fits within the **confidence-distribution framework** of frequentist statistics, which creates a distribution for θ by inverting standard test statistics.

The main advantages of PSP over traditional methods are:

- **Direct Interpretability**: PSP provides a probability that is directly interpretable in terms of practical significance, making it more meaningful for real-world applications.
- **Integration of Effect Size and Precision**: It combines the magnitude of the effect and the precision of the estimate into a single metric.
- **Avoidance of Dichotomization**: PSP moves away from the binary "reject/accept" decisions, promoting a more nuanced interpretation of statistical results.

2.2.1 Addressing Common Misinterpretations of P-value with PSP

P-values have long been a fundamental tool in statistical hypothesis testing; however, they are frequently misinterpreted, which can lead to incorrect conclusions. The PSP offers an alternative approach that addresses some of these misunderstandings by focusing on practical significance. This section explores how PSP addresses common misinterpretations associated with P-values.

One prevalent misinterpretation is the belief that the P-value represents the probability that the null hypothesis H_0 is true. This indicates a misunderstanding of the frequentist framework, where the P-value actually represents $P(data|H_0)$, not $P(H_0|data)$. In contrast, PSP calculates the probability that the true effect size θ exceeds a predefined practical significance threshold δ, given the observed data. By focusing on $P(\theta > \delta \mid data)$, PSP avoids making statements about the probability of H_0 being true or false, abandoning the concept of statistical significance.

Another common misconception is that a significant p-value implies that the null hypothesis is false and should be rejected, while a non-significant p-value means that the null hypothesis is true and should be accepted. This dichotomous interpretation oversimplifies the nuanced nature of statistical evidence. PSP addresses this issue by offering a continuous measure of evidence concerning practical significance without enforcing a rigid reject-or-accept decision.

Some researchers mistakenly interpret a large p-value as evidence in favor of the null hypothesis. However, a high p-value merely indicates a lack of sufficient evidence against the null hypothesis, not proof of its validity. PSP does not support this misinterpretation because it does not interpret low probabilities as evidence for the null hypothesis. This perspective helps prevent the erroneous conclusion that insufficient evidence against the null hypothesis

constitutes positive support for it.

There is also a tendency to confuse statistical significance with scientific or practical significance, assuming that a significant p-value indicates a result of practical relevance. P-values, however, do not measure the magnitude or importance of an effect. PSP focuses on practical significance by incorporating a predefined practical significance threshold into the analysis. By doing so, it ensures that statistical findings are directly aligned with real-world importance, bridging the gap between statistical significance and practical relevance.

Another frequent misunderstanding is the notion that a p-value less than or equal to 0.05 implies only a 5% chance of a Type I error, or false positive. This confuses the P-value with the significance level and overlooks the fact that the P-value does not provide the probability of making an error. While PSP reduces the likelihood of misinterpreting statistical significance as an error probability, users must still be cautious and understand that PSP quantifies the likelihood of a practically significant effect, not the rates of Type I or Type II errors.

Finally, p-values are often reported as inequalities (e.g., $p < 0.05$) instead of the exact level of evidence. PSP encourages the reporting of exact probability values, enhancing transparency and precision in statistical reporting. By providing specific PSP values, researchers can better interpret the strength of the evidence and make more informed decisions. This practice aligns with recommendations for improved statistical communication and helps avoid the arbitrary thresholding associated with the p-value.

In summary, the PSP addresses some of the common misinterpretations of P-values by offering a probability directly related to practical significance. By concentrating on the effect size and its real-world importance, PSP provides a meaningful and interpretable metric for statistical analysis.

2.2.2 Mathematical Derivation

To derive the PSP, we start by considering the *cumulative distribution function (CDF)* of a random variable X, which gives the probability that X takes a value less than or equal to a specific value x:

$$\Phi(x) = P(X \leq x)$$

The goal of PSP is to compute the probability that the true effect size θ

exceeds the practical significance threshold (δ):

$$P(\theta > \delta \mid data)$$

The observed effect size $\hat{\theta}$ is an estimate of the true effect size θ. Due to sampling variability, $\hat{\theta}$ varies around θ. Under the assumption that the sampling distribution of $\hat{\theta}$ is approximately normal, we indicate with $\hat{\theta}$ the normally distributed effect size around the true effect size θ with standard error $\text{SE}_{\hat{\theta}}$.

$$\hat{\theta} \approx N(\theta, \text{SE}_{\hat{\theta}}^2)$$

Since θ is a fixed but unknown parameter, we can consider the uncertainty around θ given our observed $\hat{\theta}$.

$$\theta \approx N(\hat{\theta}, \text{SE}_{\hat{\theta}}^2)$$

Thus, we can reframe the problem as:

$$P(\theta > \delta \mid data) = P(\delta < \theta \mid \hat{\theta})$$

To standardize the variables and utilize the standard normal distribution, we define:

$$Z = \frac{\theta - \hat{\theta}}{\text{SE}_{\hat{\theta}}} \qquad \text{and} \qquad z = \frac{\delta - \hat{\theta}}{\text{SE}_{\hat{\theta}}}$$

Here, z is a numerical value representing how many standard errors the practical significance threshold is away from the observed effect size. The steps of the derivation are as follows:

1. Start from the probability

$$P(\theta > \delta) = P(\theta - \hat{\theta} > \delta - \hat{\theta})$$

2. Standardize the inequality

$$P\left(\frac{\theta - \hat{\theta}}{\text{SE}_{\hat{\theta}}} > \frac{\delta - \hat{\theta}}{\text{SE}_{\hat{\theta}}}\right) = P(Z > z)$$

3. Use the CDF of the standard normal distribution

$$\Phi(z) = P(Z \leq z) \quad \rightarrow \quad P(Z > z) = 1 - \Phi(z)$$

4. Therefore, the PSP is calculated as

$$PSP = 1 - \Phi(z)$$

2.2.3 Assumptions

In this section, we review the assumptions on which the PSP relies on:

1. **Normality of the Sampling Distribution**: The PSP method assumes that the sampling distribution of the estimated effect size ($\hat{\theta}$) is approximately normal. This assumption is generally justified by the Central Limit Theorem for large sample sizes. However, for small samples or when the data are skewed, the normality assumption may not hold, potentially leading to inaccurate PSP values.

2. **Independence of Data**: The data are assumed to be independently distributed. Violations of independence, such as clustered, correlated, or time-series data, can affect the validity of the standard error ($\text{SE}_{\hat{\theta}}$) and, consequently, the PSP.

3. **Accurate Estimation of Standard Error**: The reliability of the PSP calculation depends on the accurate estimation of the standard error of the effect size. Misestimations due to heteroscedasticity or other factors can lead to misleading PSP values.

4. **Predefined Practical Significance Threshold (δ)**: The practical significance threshold should be established before data analysis and grounded in domain-specific knowledge. Post-hoc selection of δ can introduce bias and inflate the probability of Type I errors.

These assumptions are also made when computing the P-value for NHST. More specifically, the fourth assumption could be applied similarly for the alpha value.

2.2.4 Limitations

1. **Ignores Type I and Type II Error Rates**: The PSP focuses on the probability that the true effect size exceeds the PST, but does not directly control for Type I (false positive) or Type II (false negative) error rates.

2. **Sensitivity to Sample Size and Normality**: In small samples, the Central Limit Theorem may not ensure normality of the sampling distribution, making the PSP unreliable. Non-normal data distributions require alternative approaches or transformations to meet the normality assumption.

3. **PST specific for a study**: The choice of the PST may vary between studies or analysts. This subjectivity can make it difficult to compare PSP values across different studies or contexts.

4. **Exclusion of Prior Information**: The PSP method does not incorporate prior knowledge or Bayesian updating mechanisms, which can be valuable in certain research settings.

5. **Potential Misinterpretation**: The PSP provides a probability that the true effect size exceeds a threshold, which may be misinterpreted as the probability that the effect is practically significant. Users must be cautious to interpret PSP within the statistical framework and not overextend its implications.

Chapter 3

Comparative Analysis of Statistical Validation Techniques

In this chapter, we compare several statistical methods used in research. Our goal is to highlight each method's strengths and weaknesses, making it easier to choose the appropriate approach for different research questions. We will review the methods we have described in the previous chapter:

- Null Hypothesis Significance Testing (NHST) and P-Values
- Confidence Intervals
- Effect Size Measures
- Equivalence Testing and Second Generation P-Values (SGPV)
- Bayesian Factors
- Practical Significance Probability (PSP)

The table 3.1 below summarizes some of the key aspects of each statistical validation technique.
In the next chapter we provide an empirical analysis to better analyse the practical implications of each method.

Method	Easy to Learn	Easy to Interpret	Include Uncertainty	Express Effect Size	Include Minimum Practical Effect Size	Binary Test	Sample Size Dependent
NHST P-value	Yes	Yes	Yes	No	No	Yes	Yes
Confidence Interval	Yes	Yes	Yes	Yes	No	No*	Yes
Cohen's d	Yes	Yes	No	Yes	No	No	No
Equivalence Testing /SGPV	Yes	No	Yes	No	Yes	Yes	Yes
Bayes Factors	No	No	Yes	No	No	Yes	No
PSP	Yes	Yes	Yes	No	Yes	No	Yes

Table 3.1: Comparison of Statistical Methods on Different Criteria

* CI can be used in a dichotomous decision-making framework (e.g., if a 95% CI excludes zero, one might conclude statistical significance), but this approach has been criticized.

This table makes it clear that no single method meets every need. For example, traditional P-value and NHST are easy to learn and interpret, but they push us toward a binary outcome ("significant" or "not significant") without telling us how large or practically important the effect might be. Confidence intervals address uncertainty and can hint at effect size, but they do not automatically define what is "practically" meaningful. Cohen's d is a direct measure of effect size, but it does not inherently capture the uncertainty around that estimate.
Equivalence testing can account for a minimum practical effect size, but it uses a dichotomous "equivalent or not" decision rule. Bayes factors let us compare evidence for or against a hypothesis in a continuous way, but they can still be used in a yes-or-no choice once a certain threshold is reached and they do not automatically include a notion of practical significance. Moreover, they can be harder to interpret or more complicated to learn.
In contrast, PSP method is designed to keep things simple (like P-value) while explicitly relying on a minimum practical effect size and reflect uncertainty, without forcing a strict yes-no decision. Of course, as with all methods, it still

has limitations, so it's best to combine multiple approaches to assess more robustness.

To conclude, this table highlights the principles guiding our design of PSP. We wanted a method that was both easy to learn and to interpret, so non-statisticians can adopt it easily as P-value. At the same time, we wanted to capture uncertainty and rely on a minimum practical significance threshold, allowing us to talk directly in terms of practical utility. Finally, we designed it to be a continuous value between 0 and 1, and not be used as a dichotomous approach to assess any kind of significance.

Chapter 4

Empirical Analysis supporting PSP

In this chapter, we first present three examples of PSP in action, then we conduct an empirical analysis to compare it with other methods by simulating experiments.

4.1 Anecdotes with PSP

In this section, we propose three examples that show how to use PSP and its comparison with P-value.

(1) Basic PSP Usage
Suppose a clinical trial evaluates a new drug intended to lower blood pressure. The practical significance threshold (δ) is set at a reduction of 5 mmHg, deemed clinically meaningful. The trial results show an average reduction of 6 mmHg. The standard error of the effect size is 2 mmHg:

$$Z = \frac{5-6}{2} = -0.5 \qquad \Phi(-0.5) = 0.3085 \qquad PSP = 0.6915$$

There is approximately a 69% probability that the true effect size exceeds the clinically meaningful threshold of 5 mmHg.

(2) PSP and P-value in Disagreement
A nutrition researcher is evaluating a new weight loss supplement. The practical significance threshold (δ) is set at a weight loss of **3 kg** over a 12-week period. Both the treatment group n_1 and the control group n_2 have 60 participants ($n_1 = n_2 = 60$). The results of the research are:

	mean weight loss	standard deviation
Treatment group:	$\bar{X}_1 = 4$ kg	$s_1 = 5$ kg
Control group:	$\bar{X}_2 = 1.5$ kg	$s_2 = 4.5$ kg

The observed effect size is $\hat{\theta} = \bar{X}_1 - \bar{X}_2 = 2.5$ kg. The standard error can be computed with the Welch's approximation:

$$SE_{\hat{\theta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \approx 0.8684$$

Then, we can calculate the PSP as:

$$z = \frac{\delta - \hat{\theta}}{\mathrm{SE}_{\hat{\theta}}} = \frac{3 - 2.5}{0.8684} \approx 0.5758 \qquad PSP = 1 - \Phi(z) \approx 0.28$$

To make a comparison, we calculate also the p-value:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE} \approx 2.879 \qquad df = n_1 + n_2 - 2 = 118$$

The p-value for this problem is approximately 0.005. Even if we obtained a small p-value, the practical impact may be minimal as the PSP is only 28%.

(3) Example from Roberts textbook
We consider a well-known example originally given in a textbook by Roberts [40][19]. Two manufacturers, denoted by A and B, are suppliers for a component. We are concerned with the lifetime of the component and want to choose the manufacturer that affords the longer lifetime. Manufacturer A supplies 9 units for lifetime testing. Manufacturer B supplies 4 units. The test data give the sample means 42 and 50 hours, and the sample standard deviations 7.48 and 6.87 hours, for the units of manufacturer A and B respectively:

$$\begin{array}{lll} n_A = 9 & X_A = 42 & s_A = 7.48 \\ n_B = 4 & X_B = 50 & s_B = 6.87 \end{array}$$

The two-tailed p-value was 0.0923, while the one-tailed p-value was 0.0462. If we consider 1 hour as the practical significance threshold (δ), the PSP value is:

$$SE_{\hat{\theta}} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \approx 4.2445$$

$$z = \frac{\delta - \hat{\theta}}{\mathrm{SE}_{\hat{\theta}}} = \frac{1 - |50 - 42|}{4.2445} \qquad PSP = 1 - \Phi(z) \approx 0.95$$

The PSP is high ($\approx 95\%$), suggesting that Manufacturer B's components have a significantly longer lifetime than Manufacturer A's, beyond the threshold deemed practically important. However, this could require further investigation because both P-value and PSP could be inaccurate due to the small size of the samples.

4.2 Simulation Study and Empirical Evidence

4.2.1 Overview of Prior Simulation Study

In 2019, a study published in *The American Statistician* [18] proposed and tested a simulation-based approach to compare the performance of several statistical methods. In addition to the most common ones, such as the P-value with different alpha values, confidence intervals, and effect sizes, the researchers introduced a new approach called MESP (Minimum Effect-Size plus P-value).

The MESP rejects the null hypothesis only if both the following conditions are met: (1) the effect size is meaningfully large, and (2) the NHST rejects the null hypothesis with an alpha value of 0.05. This hybrid approach combines a minimum significance threshold with the traditional NHST. The authors also introduced the term *Minimum Practically Significant Distance* (MPSD), which aligns with the concept used in equivalence tests and SGPV, and that we refer to as the "practical significance threshold" in this study for the PSP.

The authors aimed to step back from discussions of theoretical grounds about P-value utility as evidence for a hypothesis and instead sought empirical evidence to help address the issue. Similar efforts have been made by other researchers in the past [41][36], showing that one of the best options could be to triangulate various methods' results to maximize confidence. A more recent simulation study [42], conducted to assess the success of P-value-based inferences, concluded that the P-value should be used as a cue alongside other statistical techniques, such as effect size and Bayes factors.

Inspired by the study published in *The American Statistician* [18], we replicated their simulation-based approach. The authors set three objectives:

1. To explore whether P-values can have evidential value.

2. To examine the nature and limitations of that value.

3. To empirically compare that evidential value with possible alternative approaches, including their proposed MESP method.

They created a dataset of 10,000 simulated experiments. Each experiment included key parameters, as listed in Table 4.1, with values randomly drawn from predefined ranges.
For each simulated case, the null hypothesis with a null mean of 100 ($H_0 : \mu = 100$) was tested using the following methods:

1. **NHST:** Reject the null hypothesis if the p-value is lower than 0.05.

Table 4.1: Parameters used in the simulated experiments

Parameter	**Predefined Range**
Sample Size	5–100
True Population Mean	75–125
True Population Standard Deviation	4–60
MPSD	2–20

2. **NHST (small α):** Reject the null hypothesis if the p-value is lower than 0.005 (as recommended in the paper *Redefine Statistical Significance* [11], signed by a coalition of 72 methodologists).

3. **Distance-Only Method:** Reject the null hypothesis if the effect size (absolute difference between the mean sample and the null mean) exceeds the MPSD.

4. **Interval-Based Method:** Reject the null hypothesis if there is no overlap between the *thick null interval*, bounded by $null\ mean \pm MPSD$, and a 95% confidence interval centered on the mean of the observed sample.

5. **MESP Method:** Reject the null hypothesis if both the conventional NHST (p-value less than 0.05) and the distance-only method reject it.

Each of these methods is evaluated by comparing its decision about the null hypothesis to a method based on a *full-knowledge null hypothesis rejection*. This benchmark rejects the null hypothesis only when the true effect size (the difference between the actual population mean and the null hypothesis mean) is greater than the MPSD.

For a method inference to be considered correct (or successful), its decision to reject or not reject the null hypothesis must match the decision made by the full-knowledge method.

The results of the 10,000 experiments are summarized in Figure 4.1. The horizontal axis categorizes the experiments by three power levels (less than 0.3, between 0.3 and 0.8, and higher than 0.8), while the vertical axis shows the percentage of inference success.

For tests with high nominal power (e.g., large samples or small population variance), the NHST had the worst true Type I error rates, as observed in "Panel 1". This indicates that the P-value test failed by not accepting the null hypothesis in about half of the cases, potentially leading to false positives.

Despite all the criticisms, the authors concluded that P-values provide some evidential information relevant to an inference about a population mean.

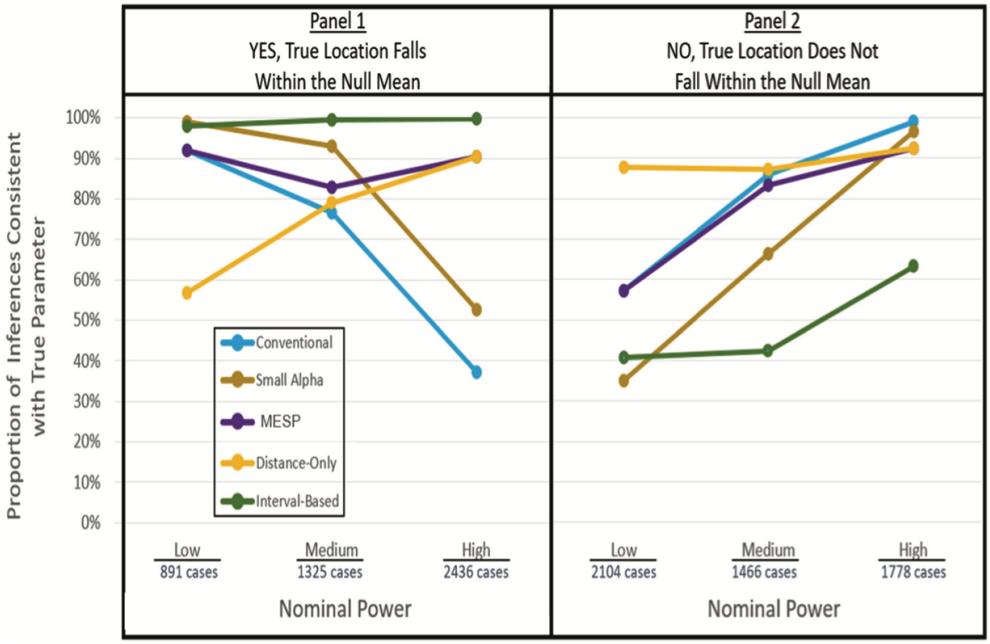

Figure 4.1: Inference success by power level. Image from [18].

They also stated that the NHST should not be regarded as a definitive or sole justification for research conclusions, but instead should be interpreted properly. All methods compared in the study demonstrated strengths and weaknesses, with none emerging as the optimal solution. Finally, they concluded that the MESP recognizes the heuristic nature of the α in the NHST and incorporates the crucial criterion of effect size.

4.2.2 Revisiting the Simulation with PSP

In this study, we replicated the same simulation-based approach from the research summarized in the previous section, but with the following modifications:

- *Discard experiments with power lower than 0.60.* Such experiments have a high probability of Type II errors (false negatives), meaning they often fail to detect a true effect even when one exists. These experiments provide weak evidence for comparing statistical methods due to their noise from small sample sizes or high variability. For this reason, we focus on high-powered experiments (> 0.60), as these reduce the amount of noise and yield more reliable comparisons. It would be unusual to consider the results coming from an experiment with a small number

of samples and high variability. Thus, for this study, we are not really interested in how statistical methods perform in those situations. We chose the threshold of 0.60 to include a broader range of experiments compared to the traditional cutoff of 0.8 for strong power.

- *Balance the number of experiments where the null hypothesis should be rejected.* We constrained our simulation algorithm to generate an equal number of experiments where the true population mean falls within and outside the null hypothesis equivalence interval. In this way, we have balanced metrics and evaluations of the statistical methods.

- *Incorporate the confusion matrix for each statistical method to compare false positives and false negatives.* One persistent issue with NHST is the misinterpretation of p-values as the probability that results are due to chance, contributing to a high rate of false positives in scientific literature. Colquhoun [39] advocates for integrating *False Positive Risk* (FPR) into statistical practice to address this issue. These metrics could help us understand how to reduce false positives and improve reproducibility.

- *Include PSP as an alternative to P-value.* Unlike p-values, which provide a dichotomous "statistical significance" decision, the PSP estimates the probability that an effect size surpasses a practical significance threshold. PSP is not a test and is not directly comparable with NHST. For this study, we defined a *PSP alpha* (PSP_{α}) representing the minimum probability required to reject the null hypothesis. While we agree with many authors on abandoning dichotomous testing approaches like NHST, we made an exception here to compare PSP's performance with traditional methods. Finally, we rename the Minimum Practically Significant Difference (MPSD) to the Practical Significance Threshold (PST), the meaning and the purpose remain unchanged.

- *Include the Least Difference in Means.*[43] This is another method recently proposed with the focus on practical significance. With respect to the PSP, this is a full Bayesian method and it is based on credible intervals.

- *Use a sample size greater than 30.* To enhance the reliability of statistical comparisons, we set a minimum sample size of 30 for the simulated experiments. This threshold is guided by the Central Limit Theorem, which ensures that the sampling distribution of the mean approximates normality as sample size increases. Additionally, larger sample sizes reduce variability and make it easier to detect true effects, resulting in more reliable outcomes. While 30 is not a strict rule, it provides a practical heuristic. By focusing on experiments with adequate sample

sizes, we reduce noise and obtain more robust evaluations of the statistical methods.

In Table 4.2, we present the columns generated for each experiment.

Column	**Explanation**
`n`	Sample size.
`true_pop_mean`	True population mean.
`true_pop_std`	True population standard deviation.
`observed_mean`	Observed sample mean.
`observed_effect_size`	Observed effect size, calculated as the difference between the observed mean and null hypothesis mean.
`observed_std`	Observed sample standard deviation.
`pst`	Practical significance threshold, defining the minimum effect size considered practically meaningful.
`power`	Statistical power of the test.
`p_value`	p-value of the hypothesis test.
`interval`	Confidence interval for the estimate, presented as a tuple of lower and upper bounds.
`PSP`	Probability of a substantial practical effect.
`cohen_d`	Cohen's d, an effect size measure calculated as (true population mean - null mean) / true population standard deviation.
`full_knowledge_reject_null`	Boolean indicating whether the null hypothesis is rejected based on true population values compared to `pst`.
`NHST_0.05`	Boolean indicating rejection of the null hypothesis under a significance level of $\alpha = 0.05$.
`NHST_0.005`	Boolean indicating rejection of the null hypothesis under a stricter significance level of $\alpha = 0.005$.

Column	Explanation
`MESP_0.05`	Boolean indicating rejection of the null hypothesis when p-value ≤ 0.05 and absolute observed effect size $\geq$ `pst`.
`confidence_interval_test`	Boolean indicating whether the confidence interval excludes the null hypothesis mean $\pm$ `pst`.
`least_diff_in_means_test`	Boolean indicating if the least difference in means is higher than the practical significance threshold.
`PSP_test_0.8`	Boolean indicating whether the probability of a substantial practical effect (PSP) is at least 0.8.

Table 4.2: Description of columns generated for each experiment.

We used the same predefined ranges as those in the original paper, as reported in Table 4.1, except for the sample size as explained above, and simulated 50,000 experiments with a minimum power threshold of 0.60. We chose a $PSP_\alpha = 0.8$ that could represent a balanced value. In Section 4.2.5 we have analyzed the behavior of the PSP method when using different levels of PSP_α.

4.2.3 Exploratory Analysis of Simulations Parameters

In this section, we aim to examine the distributions of the parameters used to simulate the experiments, checking for outliers or inconsistencies. In the original paper's simulation [18], the parameter distributions were flat, meaning the values were uniformly distributed within the specified range. However, in our study, as shown in the Image 4.2 below, we do not observe the same flat behavior. This deviation is due to our minimum power threshold, which filters out experiments with particularly high standard deviation and low effect size. Consequently, the histograms of the true population mean and true population standard deviation are not flat.

Similarly, in Image 4.3, we observe a lower number of experiments with low absolute effect size.

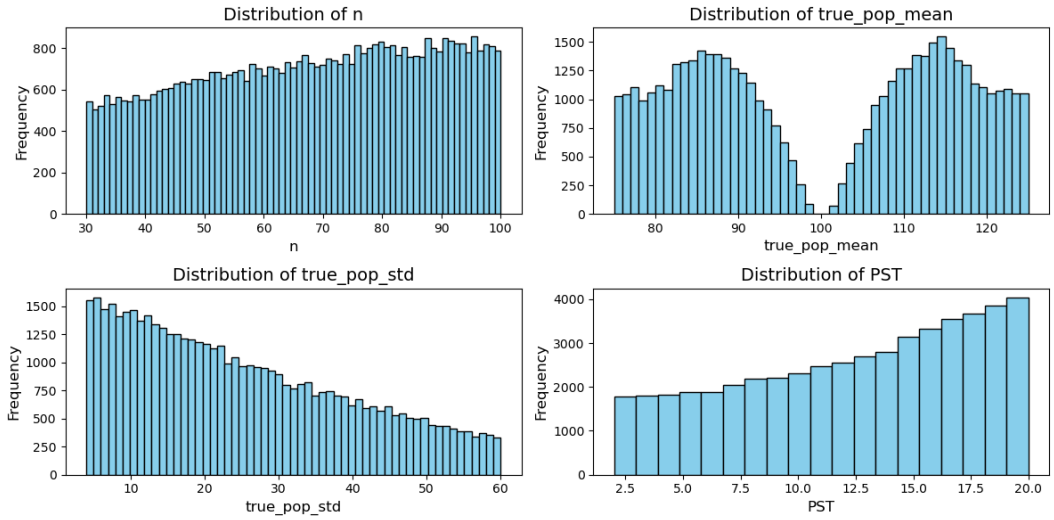

Figure 4.2: Parameter distributions in our study.

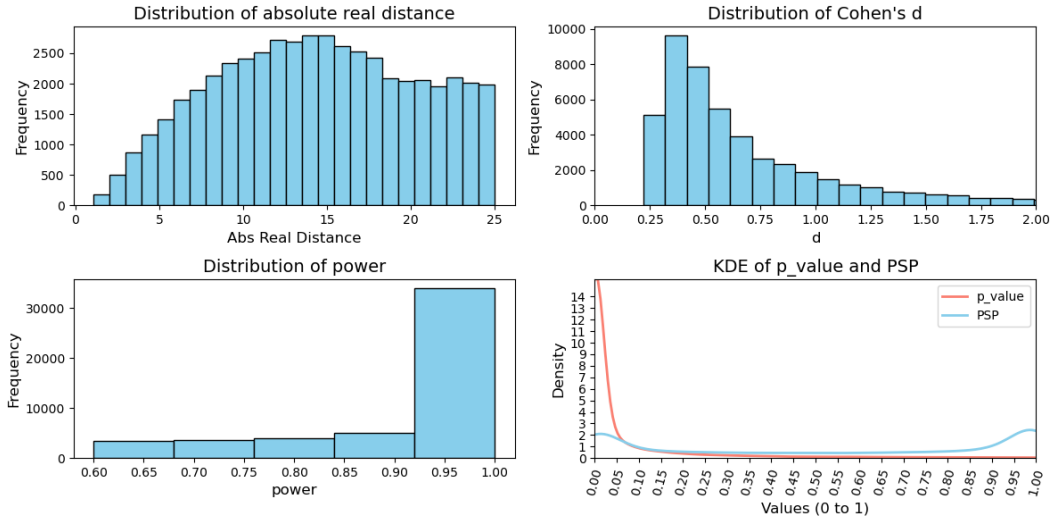

Figure 4.3: Distribution of other columns in the dataset.

The fourth image presents the kernel density estimate (KDE) plot for the p-values and PSP distributions. The p-values are more clustered near zero, while the PSP distribution is more spread out, showing concentration in the two extremes: between 0 and 0.05 and between 0.95 and 1.0.

4.2.4 Simulations Results Analysis

Once all the experiments were simulated, we conducted the statistical tests and compared each decision to reject the null hypothesis with the `full_knowledge_reject_null`, which indicates whether the null hypothesis should have been rejected. From this comparison, we calculated confusion matrices for each method, as shown in Figure 4.4.

Below there is a summary table generated from the confusion matrices, which also includes the *false positive rate* (FPR) and the *F1-score* (F1). These are defined as follows:

$$\text{FPR} = \frac{FP \cdot 100}{TP+FP}$$

$$\text{precision} = \frac{TP}{TP+FP}, \quad \text{recall} = \frac{TP}{TP+FN}, \quad F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}}$$

Test	TP	FP	TN	FN	TP%	FP%	TN%	FN%	FPR	F1
NHST_0.05	20545	18714	6286	4455	41.09	37.43	12.57	8.91	47.67	0.64
NHST_0.005	16467	13626	11374	8533	32.93	27.25	22.75	17.07	45.28	0.60
MESP_0.05	19638	4149	20851	5362	39.28	8.30	41.70	10.72	17.44	0.81
confidence_interval_test	12597	161	24839	12403	25.19	0.32	49.68	24.81	1.26	0.67
least_diff_in_means_test	12663	166	24834	12337	25.33	0.33	49.67	24.67	1.29	0.67
PSP_test_0.8	17920	1269	23731	7080	35.84	2.54	47.46	14.16	6.61	0.81

Figure 4.5: Summary Table generated from the confusion matrices.

From this analysis, we made the following observations:

- The NHST with an α level of 0.05 has shown a high tendency for false positives (37%) and a correspondingly high false positive rate (47%). This aligns with findings from the original study [10], as shown in Figure 4.1, where it is evident that the NHST wrongly rejected the null hypothesis in 63% of cases for high-power experiments. Colquhoun [10] demonstrated that using an alpha significance level of 0.05 in NHST does not ensure that the Type I error rate (false positive rate) is close to 5%. Instead, he found that the actual false positive rate can be substantially higher, depending on factors such as the prior probability of the hypothesis being true and the statistical power of the test. Specifically, Colquhoun showed that with typical conditions, a p-value of 0.05 could correspond to a false positive rate of at least 29%, and in some scenarios, it could be much higher.

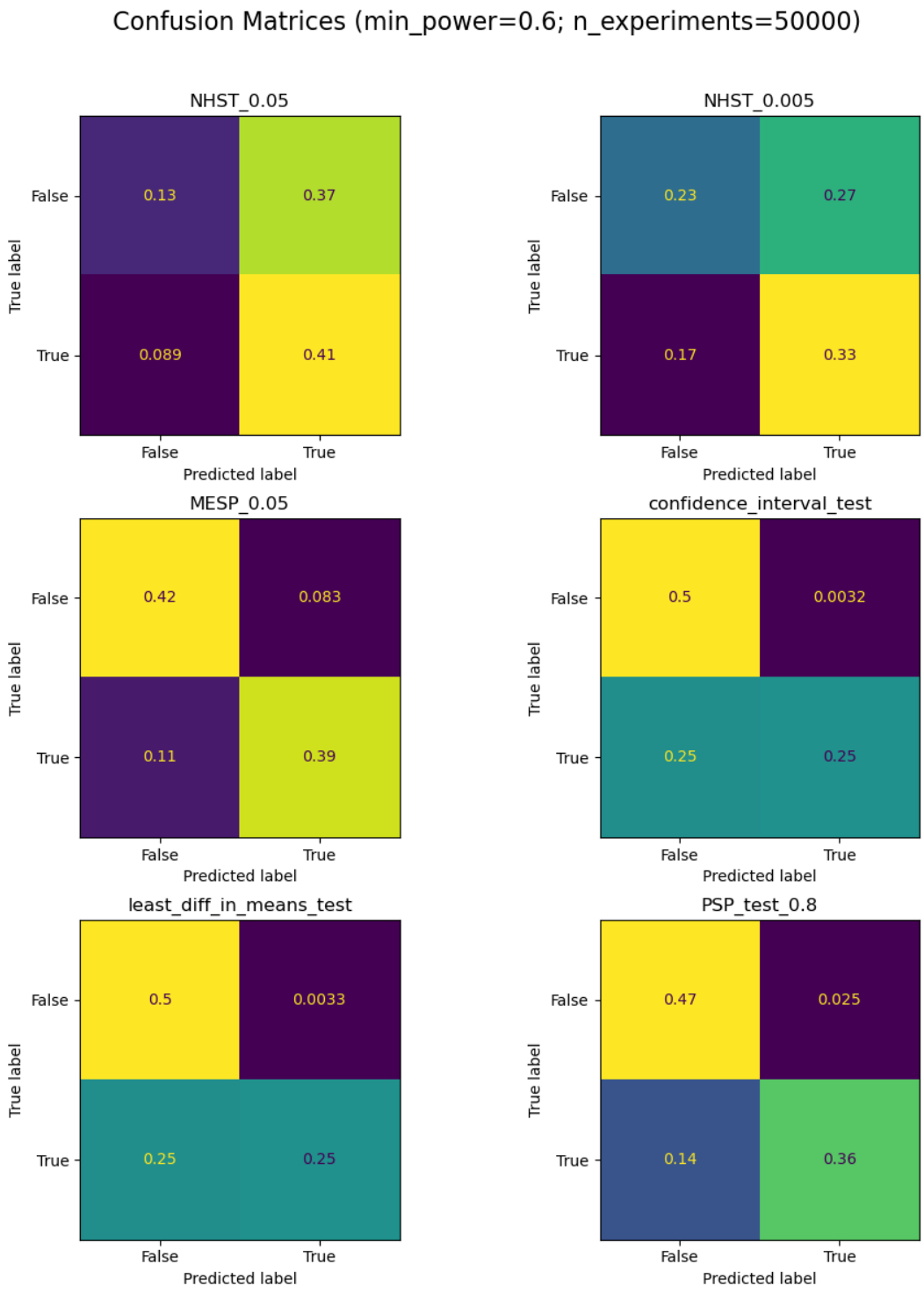

Figure 4.4: Confusion matrices for each method.

- Reducing the α level to 0.005 lowered the false positive percentage to 27%, but this value remains considerably higher compared to other methods. The false positive rate (45%) is still similar to the traditional NHST with an α of 0.05.

- The interval-based method achieved the lowest false positive percentage (0.03%), but had the highest false negative percentage ($\approx$25%). While this method is highly accurate when rejecting the null hypothesis, it often fails to detect an effect when one exists, limiting its practical utility. The interval-based test and the Least Difference in Means test have very similar values as they are highly correlated (see Figure 4.6).

- The MESP and the PSP achieved the best F1-scores, both around 81%. While the MESP method performed better in terms of false negative (10.7% vs 14% for PSP), the PSP had a better FPR (6.6% versus 17.4% for MESP). Based on these results, we conclude that the PSP method is more reliable and robust against false positives.

It is important to remember that the MESP is a hybrid approach that incorporates P-value, making it focused on determining statistical significance by answering a binary yes-or-no question. In contrast, the PSP provides a probability, avoiding the need for a dichotomous decision on statistical significance. Despite their different objectives, the two methods often produced similar results in this study, as illustrated by the correlation data in Figure 4.6.

Finally, in Table 4.3, we list the most common "decision patterns". The key observations are:

- The second row shows that in more than 20% of experiments, the NHST (with an α level of either 0.05 or 0.005) incorrectly rejected the null hypothesis, falsely indicating statistical significance. Additionally, the fourth row reveals that in 8.5% more experiments, the NHST with a α level of 0.05 made the same mistake.

- The fifth row shows that in 7.3% of experiments, the interval-based methods failed to reject the null hypothesis. This supports the idea that these methods tend to be more conservative compared to the others.

- The sixth row shows that in 5.2% of experiments, all the methods incorrectly accepted the alternative hypothesis.

- Row 14 shows that for 1.2% of experiments, PSP was the only method that correctly rejected the null hypothesis.

Column 1	Column 2	Correlation
full_knowledge_reject_null	NHST_0.05	0.09
full_knowledge_reject_null	NHST_0.005	0.12
full_knowledge_reject_null	MESP_0.05	0.62
full_knowledge_reject_null	confidence_interval_test	0.57
full_knowledge_reject_null	least_diff_in_means_test	0.57
full_knowledge_reject_null	PSP_test_0.8	0.68
NHST_0.05	NHST_0.005	0.64
NHST_0.05	MESP_0.05	0.50
NHST_0.05	confidence_interval_test	0.16
NHST_0.05	least_diff_in_means_test	0.17
NHST_0.05	PSP_test_0.8	0.34
NHST_0.005	MESP_0.05	0.40
NHST_0.005	confidence_interval_test	0.32
NHST_0.005	least_diff_in_means_test	0.32
NHST_0.005	PSP_test_0.8	0.37
MESP_0.05	confidence_interval_test	0.50
MESP_0.05	least_diff_in_means_test	0.50
MESP_0.05	PSP_test_0.8	0.77
confidence_interval_test	least_diff_in_means_test	0.99
confidence_interval_test	PSP_test_0.8	0.64
least_diff_in_means_test	PSP_test_0.8	0.64

Figure 4.6: Correlation between statistical methods.

	Full Knowledge Reject NH	**Reject NH**	**DON'T Reject NH**	**Count**
1	True	full_knowledge_reject_null NHST_0.05 NHST_0.005 MESP_0.05 confidence_interval_test least_diff_in_means_test PSP_test_0.8		10895 (21.8%)
2	False	NHST_0.05 NHST_0.005	full_knowledge_reject_null MESP_0.05 confidence_interval_test least_diff_in_means_test PSP_test_0.8	10324 (20.6%)
3	False		full_knowledge_reject_null NHST_0.05 NHST_0.005 MESP_0.05 confidence_interval_test least_diff_in_means_test PSP_test_0.8	6255 (12.5%)
4	False	NHST_0.05	full_knowledge_reject_null NHST_0.005 MESP_0.05 confidence_interval_test least_diff_in_means_test PSP_test_0.8	4241 (8.5%)
5	True	full_knowledge_reject_null NHST_0.05 NHST_0.005 MESP_0.05 PSP_test_0.8	confidence_interval_test least_diff_in_means_test	3656 (7.3%)
6	True	full_knowledge_reject_null	NHST_0.05 NHST_0.005 MESP_0.05 confidence_interval_test least_diff_in_means_test PSP_test_0.8	2603 (5.2%)
7	False	NHST_0.05 NHST_0.005 MESP_0.05	full_knowledge_reject_null confidence_interval_test least_diff_in_means_test PSP_test_0.8	2069 (4.1%)
8	True	full_knowledge_reject_null NHST_0.05 MESP_0.05 PSP_test_0.8	NHST_0.005 confidence_interval_test least_diff_in_means_test	2061 (4.1%)

	Full Knowledge Reject NH	**Reject NH**	**DON'T Reject NH**	**Count**
9	True	full_knowledge_reject_null NHST_0.05 NHST_0.005 MESP_0.05	confidence_interval_test least_diff_in_means_test PSP_test_0.8	1283 (2.6%)
10	True	full_knowledge_reject_null NHST_0.05 MESP_0.05	NHST_0.005 confidence_interval_test least_diff_in_means_test PSP_test_0.8	1184 (2.4%)
11	False	NHST_0.05 NHST_0.005 MESP_0.05 PSP_test_0.8	full_knowledge_reject_null confidence_interval_test least_diff_in_means_test	1095 (2.2%)
12	True	full_knowledge_reject_null confidence_interval_test least_diff_in_means_test	NHST_0.05 NHST_0.005 MESP_0.05 PSP_test_0.8	1065 (2.1%)
13	False	NHST_0.05 MESP_0.05	full_knowledge_reject_null NHST_0.005 confidence_interval_test least_diff_in_means_test PSP_test_0.8	811 (1.6%)
14	True	full_knowledge_reject_null PSP_test_0.8	NHST_0.05 NHST_0.005 MESP_0.05 confidence_interval_test least_diff_in_means_test	579 (1.2%)
15	True	full_knowledge_reject_null NHST_0.05 NHST_0.005	MESP_0.05 confidence_interval_test least_diff_in_means_test PSP_test_0.8	520 (1.0%)
16	True	full_knowledge_reject_null NHST_0.05 MESP_0.05 confidence_interval_test least_diff_in_means_test PSP_test_0.8	NHST_0.005	415 (0.8%)
17	True	full_knowledge_reject_null NHST_0.05	NHST_0.005 MESP_0.05 confidence_interval_test least_diff_in_means_test PSP_test_0.8	387 (0.8%)

Table 4.3: Results of hypothesis testing scenarios.

4.2.5 Understanding PSP_α Threshold Behavior

In our empirical analysis, we used a PSP_α of 0.8 as the threshold for rejecting the null hypothesis. We then compared the PSP test performance with the other statistical tests. In this section, we explore how varying the PSP_α from 0.5 to 0.99 affects the results, and we measure how each PSP test with a different threshold is correlated with the other tests.

Figure 4.7 shows the confusion matrices for different PSP_α values. As expected, increasing PSP_α lowers both the false positive rate and the true positive rate. A threshold between 0.75 and 0.8 seems like a balanced choice, while a PSP_α higher than 0.9 makes the test more conservative, producing confusion matrices similar to those of the interval-based methods.
We can see that in Figure 4.8, where the correlation between each PSP_α test and each other statistical test is calculated. For $PSP_\alpha > 0.9$, the correlation with the `confidence_interval_test` and the `least_diff_in_means_test` increases (ranging from 0.83 to 0.91).
On the other hand, PSP has the highest correlation with the full knowledge method for low PSP_α, but this comes at the cost of a higher false discovery rate (around 16%).
The PSP test with low PSP_α is also highly correlated with the MESP test. This is probably because both methods rely on comparing the observed effect size with the practical significance threshold.

Method	full_knowledge_reject_null	NHST_0.05	NHST_0.005	MESP_0.05	confidence_interval_test	least_diff_in_means_test
PSP Alpha						
0.500000	0.72	0.31	0.28	0.90	0.49	0.50
0.550000	0.72	0.31	0.29	0.88	0.51	0.52
0.600000	0.72	0.32	0.31	0.86	0.53	0.54
0.650000	0.71	0.32	0.32	0.84	0.55	0.56
0.700000	0.71	0.33	0.33	0.81	0.58	0.58
0.750000	0.70	0.33	0.35	0.79	0.60	0.61
0.800000	0.68	0.34	0.37	0.77	0.64	0.64
0.850000	0.66	0.34	0.38	0.74	0.67	0.68
0.900000	0.63	0.34	0.39	0.70	0.73	0.74
0.950000	0.58	0.32	0.40	0.64	0.83	0.84
0.970000	0.55	0.29	0.40	0.59	0.90	0.91
0.990000	0.48	0.26	0.39	0.51	0.84	0.83

Figure 4.8: Correlation between each PSP_α test and the other statistical methods.

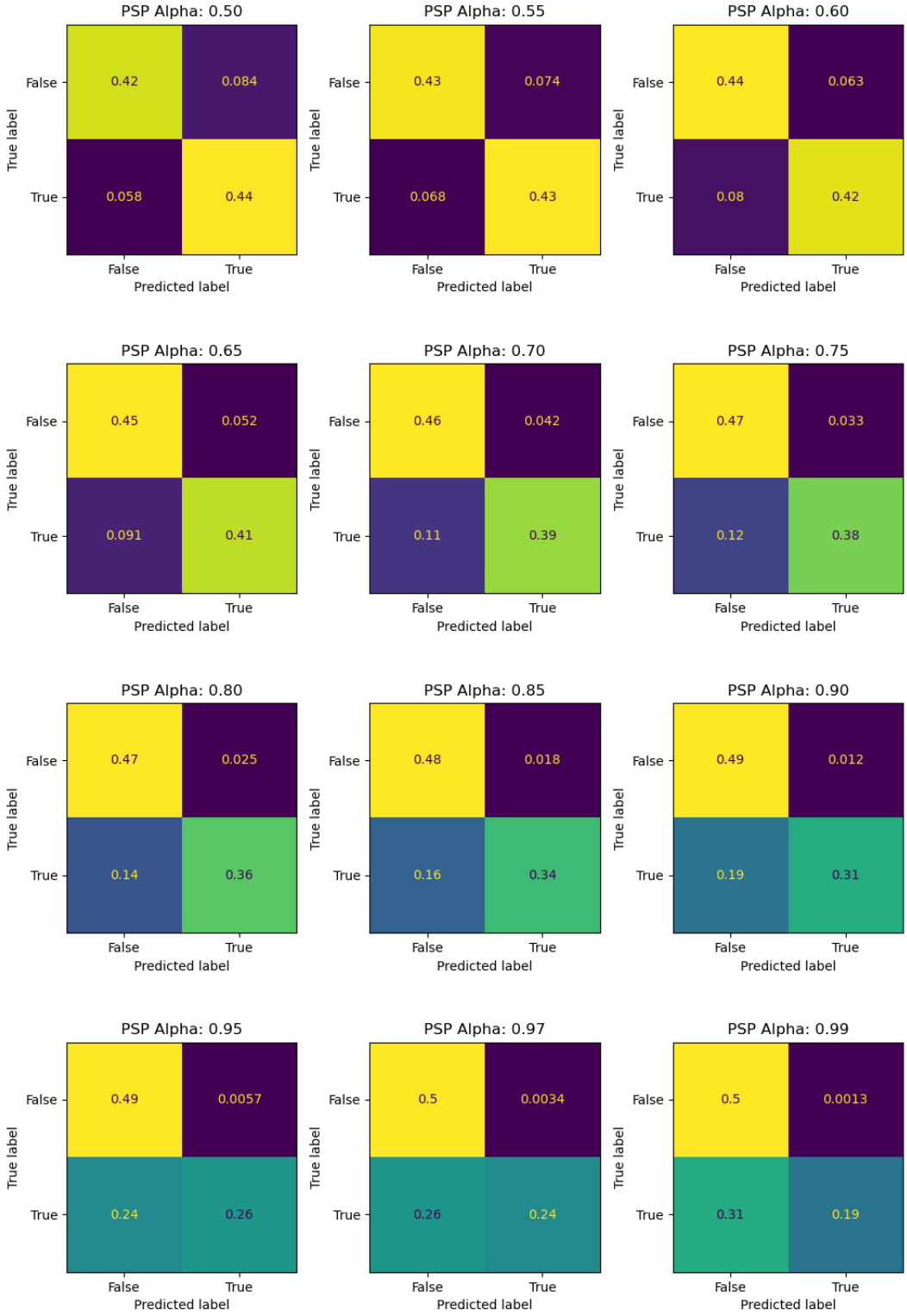

Figure 4.7: Confusion matrices for each PSP_α.

Chapter 5

Code

All the code used in this project is stored in a public GitHub repository[1], where the main scripts for the simulations, data analysis, and plotting can be found. The repository contains separate functions for computing p-values, confidence intervals, and other metrics.

In Listing 1 on the next page, the `PSP` function implementation is shown, which follows a structure similar to `scipy.stats` functions. This design choice helps keep the interface user-friendly and consistent with the SciPy library, which is one of the most commonly used to compute the P-value.

This snippet illustrates how the `PSP` function calculates a probability value (from 0 to 1) representing how much the observed effect surpasses a chosen practical significance threshold (`pst`). As in other SciPy implementations, the user can specify the method for computing the standard error. Typically, if the control and treatment groups share the same variance, the Student t-test formula is used; otherwise, Welch's approximation is applied. In some cases (as in the psychology field), Welch's method is often more robust and is only slightly worse than the Student one when the variances are about equals [44]. Further details and additional code components are available online.

[1] https://github.com/Borgo99/empirical-analysis-supporting-PSP

```python
def PSP(control_sample, treatment_sample, pst, equal_var=True):
    if len(control_sample) != len(treatment_sample):
        raise ValueError("Control and treatment samples must have the
        ↪  same length.")
    n = len(treatment_sample)
    control_mean = np.mean(control_sample)
    treatment_mean = np.mean(treatment_sample)
    control_std = np.std(control_sample, ddof=1)
    treatment_std = np.std(treatment_sample, ddof=1)
    observed_effect_size = treatment_mean - control_mean

    if equal_var:
        # 1) Calculate pooled standard deviation:
        sp = math.sqrt(
            ((n - 1) * control_std**2 + (n - 1) * treatment_std**2)
            / (2*n - 2)
        )
        # 2) Standard error of the difference:
        stde = sp * math.sqrt(2.0 / n)
    else:
        # Welch's approximation for unequal variance:
        stde = math.sqrt(
            control_std**2 / n + treatment_std**2 / n
        )

    psp = 1 - stats.norm.cdf((abs(pst) - abs(observed_effect_size)) /
    ↪  stde)
    return psp
```

Listing 1: PSP Function Code

Chapter 6

Conclusions and Future Works

6.1 Best Practices in Statistical Testing

In this thesis, we have explored many studies that discuss the benefits and drawbacks of different statistical methods. Over the years, several authors have questioned the use of the P-value and Null Hypothesis Significance Testing (NHST) to decide what is called "statistical significance." The American Statistical Association, along with the 800 authors of the paper "Retire Statistical Significance" [15], has strongly advised people not to use this term. One main reason is that the null hypothesis, which states that an effect is exactly zero, is often not realistic, so rejecting it is not particularly meaningful [21][28].

Researchers have proposed many possible replacements for NHST, including equivalence tests and new Bayesian techniques, but none of these methods has completely taken its place. At the start of this thesis, we mentioned Cohen, and now we wish to recall his statement from 1994 [20], which now sounds almost as a theorem:

> "Don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist."

In other words, there is no single method that will always work best. The strongest approach is to use and compare several methods, so that we can benefit from the strengths of each one.

It is also important to note that "P-values behave exactly as they should" [16]. The main issue lies in their misinterpretation [1]. The P-value, defined as $p(\text{data} \mid \theta)$, does not actually tell us $p(\theta \mid \text{data})$, which is what researchers often want to know [21]. In many cases, P-values have been used to answer the wrong question, and this misuse comes from bad practices, not from the

concept of the P-value itself.

When reporting results, some authors suggest that P-value should be treated as a continuous measures [33]. Others recommend using the S-value [16][13]. Even if researchers choose to report the P-value, they should do it with caution. Very small P-values (for example, $p < 0.001$ [10]) could be more meaningful compared to a p-value of 0.05 that has an high false discovery rate[9], but even a very small P-value only shows that the observed data do not align well with the assumed model. It does not reveal which specific assumption is incorrect. A small or large P-value may come from an incorrect hypothesis, violations in the study protocol, or selective reporting of findings [1].

A statistic is an estimate of an unknown population parameter, derived from a random subsample of that population. If data were available for the entire population, there would be no uncertainty in sampling. To assess model uncertainty, it is important first to identify what assumptions go into the model. These include formal requirements of the statistical model as well as choices made by the researcher, such as the selection of samples. Second, researchers should check how valid these assumptions are. Third, they should analyze how key results change when the model is altered in different ways [37]. According to “Moving to a world beyond p < 0.05” [26], researchers should "Accept uncertainty. Be thoughtful, open, and modest.". This viewpoint is also emphasized in the paper "Retire Statistical Significance", where the 800 authors advise to embrace uncertainty, for example by renaming confidence intervals as "compatibility intervals". This renaming helps remind us to avoid overconfidence. They also recommend describing the practical implications of any values within these intervals.

These findings are also important for AI and ML contexts, where comparing different models' performance and evaluating their practical implications is a key aspect. In ML, large datasets can lead to low p-values, which might overestimate the differences between models. Relying only on p-values and confidence intervals may not provide a full picture. Therefore, a statistical section should include other measures that consider also the practical implications.

A complete and robust statistical section should address several key ideas:

- Effect sizes can offer important information about the magnitude and practical importance of a finding [36]. Even when using precise P-value, it is crucial to interpret them alongside confidence intervals and effect sizes [1].

- Confidence intervals generally give a more direct insight into the size of an effect because they reflect both the estimate and its associated uncertainty [45]. Nevertheless, neither confidence intervals nor significance tests alone can confirm with absolute certainty whether an effect truly exists or not [1]. Therefore, we should also pay attention to confidence interval misinterpretations.

- Bayesian methods incorporate prior information directly into the model, making them useful for more refined analyses [1]. However, a full Bayesian analysis might be too complex to replace P-value as a standard practice in many fields. It may be more practical in studies where professional statisticians can guide the process [39].

- An increasingly popular way to present results is through estimation plots, which place the focus on effect sizes and their confidence intervals. By explicitly showing the uncertainty and the difference between groups, these plots help researchers and readers make more informed decisions [34].

- Practical significance addresses whether a result is meaningful in real-world settings [21]. Including a method that accounts for a minimum practically meaningful effect size can quickly highlight the actual impact of a study's findings. Our proposed Practical Significance Probability method can be effectively used for this purpose.

6.2 Conclusions on PSP

In this study, we have introduced the Practical Significance Probability (PSP) as a complementary approach that shifts attention from purely statistical significance to practical importance. PSP gives an intuitive probability that the true effect surpasses a chosen practical threshold, making it easier to judge the real-world relevance of a result.
Our empirical analysis and simulation study highlight several benefits of PSP. By setting a threshold (PSP_α) of 0.8 to reject the null hypothesis, we achieved an F1-score of about 81% and kept the false positive rate around 6.6%. This performance compares well to other methods, suggesting that PSP is less prone to overstating evidence than traditional NHST.
However, PSP should not be treated as another "reject-or-accept" test beyond this specific evaluation. We agree with the idea of moving away from strictly binary conclusions. The PSP is an easily understood probability measure that should be used alongside other methods, as recommended in the previous section.

6.3 Future Work

The extensive application of PSP across various fields could help determine its generalizability. Comparative studies using real-world datasets can help understand how PSP behaves in different contexts, identify domain-specific considerations, and inspire further innovations in statistical methodology. In ML, future research might focus on integrating PSP into model selection and validation workflows, providing a more nuanced evaluation of performance differences that matter in practice.

Ringraziamenti

Ringrazio i miei genitori, questo traguardo è dedicato a loro ed è frutto dei loro sforzi e sacrifici. Mi hanno sempre fatto sentire amato e sono, e continueranno a esserlo, un modello e un esempio da seguire. Grazie per aver sempre creduto in me.

Ringrazio la mia dolce Giulia per essermi stata vicina fin dall'inizio e per avermi amato anche quando siamo stati distanti. Voglio condividere questo traguardo con lei perché è stata fondamentale nel permettermi di raggiungerlo.

Ringrazio anche tutta la mia bellissima famiglia, i miei nonni e i miei amici di sempre.

Bibliography

[1] Senn S.J. Rothman K.J. et al. Greenland, S. Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31, 04 2016.

[2] James L. Rogers, Kenneth I. Howard, and John T. Vessey. Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3):553–565, 1993.

[3] Nicholas M. White, Thanya Balasubramaniam, Rajath Nayak, and Adrian G. Barnett. An observational analysis of the trope "a p-value of < 0.05 was considered statistically significant" and other cut-and-paste statistical methods. *PLoS ONE*, 17(3):e0264360, 2022.

[4] Jorge N. Tendeiro and Henk A. L. Kiers. A review of issues about null hypothesis bayesian testing. *Psychological Methods*, 24(6):774–795, 2019.

[5] Rens van de Schoot, Simon D. Winter, Olivia Ryan, Mariska Zondervan-Zwijnenburg, and Sarah Depaoli. A systematic review of bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2):217–239, 2017.

[6] R. A. Fisher. Statistical methods for research workers. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics*, Springer Series in Statistics. Springer, New York, NY, 1992.

[7] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[8] Joseph Berkson. Tests of significance considered as evidence. *Journal of the American Statistical Association*, 37(219):325–335, 1942.

[9] Thomas Sellke, M. J. Bayarri, and James O. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

[10] David Colquhoun. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1(3):140216, November 2014.

[11] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafò, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, Jan 2018.

[12] Chris Woolston. Psychology journal bans p values. *Nature*, 519:9, 2015.

[13] Ronald D. Fricker, Kirsten Burke, Xiaoyue Han, and William H. Woodall. Assessing the statistical analyses used in basic and applied social psychology after their p-value ban. *The American Statistician*, 73(sup1):374–384, 2019.

[14] Ronald L. Wasserstein and Nicole A. Lazar. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

[15] Valentin Greenland, Amrhein and Blake McShane. Retire statistical significance. *Nature*, 567:305–307, 03 2019.

[16] Sander Greenland. Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, 73(sup1):106–114, 2019.

[17] Jeffrey D. Blume, Robert A. Greevy, Victoria F. Welty, Jennifer R. Smith, and William D. Dupont. An introduction to second-generation p-values. *The American Statistician*, 73(sup1):157–167, 2019.

[18] W. M. Goodman, S. E. Spruill, and E. Komaroff. A proposed hybrid effect size plus p-value criterion: Empirical evidence supporting its use. *The American Statistician*, 73(sup1):168–185, 2019.

[19] Hening Huang. Exceedance probability analysis: a practical and effective alternative to t-tests. *Journal of Probability and Statistical Science*, 20:80–97, 10 2022.

[20] Jacob Cohen. The earth is round (p<. 05). *American Psychologist*, 49:997–1003, 12 1994.

[21] Roger Kirk. Practical significance: A concept whose time has come. *Educational and Psychological Measurement - EDUC PSYCHOL MEAS*, 56:746–759, 10 1996.

[22] Michael J. Peeters. Practical significance: Moving beyond statistical significance. *Currents in Pharmacy Teaching and Learning*, 8(1):83–89, 2016.

[23] Joseph Berkson. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33(203):526–536, 1938.

[24] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236:333–380, 1937.

[25] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition, 1988.

[26] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a world beyond "p $<$ 0.05". *The American Statistician*, 73(sup1):1–19, 2019.

[27] Roger Peng. The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3):30–32, 2015.

[28] Daniël Lakens. The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, 16(3):639–648, 2021.

[29] Russell C. Hanson. Evidence and procedure characteristics of "reliable" propositions in social science. *American Journal of Sociology*, 63:357–370, 1958.

[30] Dana R. Carney, Amy J. Cuddy, and Andy J. Yap. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, 21(10):1363–1368, 2010.

[31] Eva Ranehill, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A. Weber. Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26(5):653–656, 2015.

[32] Dana R. Carney. My position on "power poses". 2016. Accessed: 2025-02-18.

[33] Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019.

[34] Justin Ho, Taha Tumkaya, Sumit Aryal, Hyungwon Choi, and Adam Claridge-Chang. Moving beyond p values: Data analysis with estimation graphics. *Nature Methods*, 16:565–566, 2019.

[35] Daniël Lakens and Marie Delacre. Equivalence testing and the second generation p-value. *Meta-Psychology*, 4, 2020.

[36] R. Wetzels, D. Matzke, M. D. Lee, J. N. Rouder, G. J. Iverson, and E. J. Wagenmakers. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3):291–298, May 2011.

[37] Andrea A. Anderson. Assessing statistical results: Magnitude, precision, and model uncertainty. *The American Statistician*, 73(sup1):118–121, 2019.

[38] John K. Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018.

[39] David Colquhoun. The false positive risk: A proposal concerning what to do about p-values. *The American Statistician*, 73(sup1):192–201, 2019.

[40] N. A. Roberts. *Mathematical Methods in Reliability Engineering*. McGraw-Hill Book Co. Inc., New York, 1964.

[41] G. S. Howard, S. E. Maxwell, and K. J. Fleming. The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and bayesian analysis. *Psychological Methods*, 5(3):315–32, Sep 2000.

[42] J. I. Krueger and P. R. Heck. The heuristic value of p in inductive statistical inference. *Frontiers in Psychology*, 8:908, Jun 2017.

[43] Bruce A. Corliss, Yaotian Wang, Heman Shakeri, and Philip E. Bourne. The least difference in means: A statistic for effect size strength and practical significance, 2022.

[44] M. Delacre, D. Lakens, and C. Leys. Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, 30(1):92–101, April 2017.

[45] Dong Kyu Lee. Alternatives to p value: Confidence interval and effect size. *Korean Journal of Anesthesiology*, 69(6):555–562, 2016.