Journal of Machine Learning Research 25 (2024) 1-42 Submitted 1/24; Revised 6/24; Published 9/24

Estimating the Replication Probability of Significant Classification Benchmark Experiments

Daniel Berrar DANIEL.BERRAR@OPEN.AC.UK
Machine Learning Research Group
School of Mathematics and Statistics
The Open University
Milton Keynes MK7 6AA, United Kingdom
and
Department of Information and Communications Engineering
School of Engineering
Tokyo Institute of Technology
2-12-1-S3-70 Ookayama, Meguro-ku, Tokyo 152-8550, Japan

Editor: Fei Sha

Abstract

A fundamental question in machine learning is: "What are the chances that a statistically significant result will replicate?" The standard framework of null hypothesis significance testing, however, cannot answer this question directly. In this work, we derive formulas for estimating the replication probability that are applicable in two of the most widely used experimental designs in machine learning: the comparison of two classifiers over multiple benchmark datasets and the comparison of two classifiers in k-fold cross-validation. Using simulation studies, we show that p-values just below the common significance threshold of 0.05 are insufficient to warrant a high confidence in the replicability of significant results, as such p-values are barely more informative than the flip of a coin. If a replication probability of around 0.95 is desired, then the significance threshold should be lowered to at least 0.003. This observation might explain, at least in part, why many published research findings fail to replicate.

Keywords: benchmarking, p-value, replication, reproducibility, significance test

1. Introduction

Machine learning is, to a large extent, an experimental science where much of the progress is due to empirical evidence (Langley, 1988; Pineau et al., 2021; Gundersen et al., 2022). For example, benchmark experiments are widely used to compare the performance of supervised learning algorithms (Drummond and Japkowicz, 2010; Benavoli et al., 2017; Berrar, 2022; Wainer, 2023). To evaluate a new classifier, it has become common practice to compare it with established classifiers and to assess the differences in performance with a statistical significance test (Drummond and Japkowicz, 2010; Henderson et al., 2018; Cockburn et al., 2020; Berrar, 2022; Stapor et al., 2021).

Why is significance testing nowadays so widely used in machine learning? One reason might be that the reviewer guidelines of some of the top-tier journals and conferences

explicitly request that reviewers check whether a significance test was carried out. Significance tests are also often recommended in machine learning tutorials (Lones, 2024). Clear "wins," ideally supported by a significance test, of a novel algorithm over established ones are typically required for a research paper to be accepted at a top venue (Sculley et al., 2018). There is also certainly a genuine desire by researchers to underpin their interpretations with a statistical test as a, presumably, objective and rigorous method providing reassurance about the validity of the conclusions drawn (Demšar, 2006).

Yet what exactly does it mean when a result is "statistically significant?" Commonly, a result or effect is considered significant if the p-value from a null hypothesis significance test (NHST) is below 0.05. This p-value is defined as the probability of (potential) results that are as extreme as, or even more extreme than, the actually observed result, given that the null hypothesis is true and given the investigator's stopping and testing intentions. Let the null hypothesis be $H_0 : X \sim f(x, \theta)$, where X denotes data, and $f(x, \theta)$ is a probability density with parameter θ (Berger and Delampady, 1987; Bayarri and Berger, 2000). Then

$$p\text{-value} = \mathbb{P}(T \geq t(x_{\text{obs}}) \mid H_0, I) \tag{1}$$

where the observed data is x_{obs}, and $T = t(X)$ is a test statistic to investigate the compatibility of the null hypothesis with the data. Here, t is a statistical function, for example, the mean. I denotes the stopping and testing intentions.

The p-value is truly an intricate measure, as it takes into account hypothetically observable, but actually unobserved results. A little known fact about the p-value is that it also depends on how the investigator thought about the experiment and that even unrealized intentions influence its calculation. Although this might seem bizarre, it is a logical consequence of the definition of the p-value (for an instructive example, see, e.g., (Berrar, 2022), Section 3.2. Note that the notion of statistical significance depends on a dichotomous interpretation of the p-value: if the p-value is smaller than a threshold (typically, 0.05), the result is significant; if not, not. It is preferable, however, to interpret the p-value as a continuous measure that quantifies the degree of compatibility of the observed data with the null hypothesis. Nonetheless, $p < 0.05$ is often one of the decisive factors for the acceptance of a paper (McShane et al., 2019; Drummond and Japkowicz, 2010) and widely adopted in many disciplines, including computer science (Cockburn et al., 2020).

Statistical significance is also widely understood to imply that a result is "real" or "unlikely to be due to chance," despite the numerous warnings (e.g., by Schmidt and Hunter (2016); Goodman (2008, 1999); Greenland et al. (2016); Stapor et al. (2021)) against this misinterpretation that has become known as the *fallacy of the transposed conditional* or *prosecutor's fallacy*.

Another interpretation of "significance" is related to the concept of *replication*, which is at the core of the scientific enterprise. An effect or result is generally regarded as having been replicated successfully if the effect (or result) is statistically significant in the same direction in both the initial study and the follow-up study (Miller, 2009). For example, a better treatment effect due to the administration of a drug A (compared to another drug B) is regarded as having been replicated if the initial study shows that A is significantly better than B, and if also the follow-up study shows that A is significantly better than B.

In Section 2, we will provide a formal definition and detailed analysis of the *replication probability*, which is a central concept of this study. The replication probability can be un-

derstood in two different ways, which Miller (2009) calls the "individual" and "aggregate" replication probability. The aggregate replication probability refers to the proportion of significant results in the pool of all follow-up experiments from different researchers in a given domain who test different null hypotheses. By contrast, the individual replication probability is defined as "the long-run proportion of significant results within exact replications of a particular initial study." (Miller, 2009)[p.618]. For example, when we observe that two classifiers perform significantly differently on a number of benchmark datasets, we are interested in the individual replication probability, that is, whether we would obtain again a significant result in the same direction if we were to repeat the study. Here, a significant result "in the same direction" means that classifier A is significantly better than B in the replication, provided that A was also significantly better than B in the initial experiment. We will focus on the individual replication probability, as we believe that it is of primary interest in the context of machine learning. Clearly, once a researcher has obtained a significant finding, it is natural to ask about the chances that this finding will replicate if the same experiment is repeated. By "replication," we mean that the replication (or follow-up) experiment is carried out like the initial experiment, that is, using the same learning algorithms, the same sizes of training and test sets, data resampling protocol (e.g., 10-fold cross-validation), performance metrics, statistical test, implementation, seeding of the random number generator, etc. We assume that there are two important differences between the initial experiment and the replication, though. First, the replication might be carried out by an independent, different investigator. Second, the data in the replication are new random samples from the same population as in the initial experiment.

Let us now consider the following example, which describes a typical benchmark study in machine learning:

> "Two classifiers, A and B, are compared on 44 different benchmark datasets from the UCI repository. The difference in performance (for example, accuracy) is assessed based on a suitable significance test. Suppose that A performed significantly better than B, with a p-value of 0.01. What can we now say about the probability that A will again significantly outperform B in a follow-up study of the same design, that is, in an exact replication study?"

We would like to invite the reader to briefly ponder this question—what is the most accurate answer?

1. The replication probability is about 0.50.
2. The replication probability is about 0.99.
3. The replication probability is about 0.75.
4. The replication probability is at least 0.90.
5. The replication probability is at most 0.10.
6. The p-value is uninformative about the replication probability.

As we will see later, the best answer is (3). Answer (6) is not correct because the p-value does provide some information about the replication probability, but it does so only indirectly and under some assumptions. The p-value of 0.01 from the initial study would generally be interpreted as "highly significant" (Nuzzo, 2014; McShane et al., 2019) and therefore,

presumably, indicate that the null hypothesis of equal performance is most certainly not true. Consequently, there should be a very high chance (perhaps $1 - 0.01 = 0.99$?) of replicating the significant result. This reasoning, however, amalgamates the prosecutor's fallacy and the replication fallacy (Carver, 1978) and is not correct. Remarkably, even answer (4) is incorrect.

Oakes (1986) posed a conceptually similar problem to 70 academic psychologists, and 60% of them thought that an initial p-value of 0.01 implies that the chances are 99% that a replication would yield again a significant result. Another study confirmed that this misconception is widely held among academic psychologists (Gigerenzer, 2018). It is tempting to speculate that this misconception is not much less prevalent in other scientific disciplines. Nuzzo (2014) notes that most researchers would misinterpret $p = 0.01$ as a 99% chance of a successful replication. Another study by Lai et al. (2012) showed that researchers in psychology, medicine, and even statistics tend to severely underestimate the variability of the p-value in replication experiments.

The upshot of these studies is that significant p-values are widely misinterpreted as being indicative of a high chance of replication. In fact, p-values do not directly quantify the replication probability. Under some assumptions, however, it is possible to estimate the replication probability based on p-values. The replication probability gives a direct answer to a central question that the p-value cannot provide:"Given that I obtained a significant result, what are the chances that I will obtain it again in a replication study?" This question is the fundamental motivation for the present work.

1.1 Novel Contributions

Our novel contributions can be summarized as follows. First, we derived formulas for the replication probability for two of the most common experimental designs in supervised learning: (i) the comparison of two classifiers over multiple different benchmark datasets, (ii) and the comparison of two classifiers in k-fold cross-validation. We carried out a number of experiments to empirically validate these formulas.

Second, our research contributes to the ongoing debate on the reproducibility crisis and the appropriate threshold for the p-value. Previously, it was suggested to lower the significance threshold from 0.05 to 0.005 for claims of novel discoveries (Benjamin et al., 2018; Ioannidis, 2018). Our results, however, provide a rationale for an even lower threshold of approximately 0.003.

Third, we show that barely significant results, i.e., with p-values in the range $[0.01, 0.05)$, fail to replicate with high probability. Remarkably, a p-value just below the common significance threshold of 0.05 is not much more informative than the flip of coin. This observation has potentially far-reaching implications for empirical research in machine learning and editorial guidelines for authors.

1.2 Related Work

At the heart of the scientific enterprise is the process of stating a hypothesis, specifying a research protocol, carrying out experiments, collecting data, analyzing the data, interpreting the findings, and drawing conclusions. Clearly, it is of paramount importance that the experiments are repeatable, so that the observations and conclusions can be consid-

ered valid. The notions of "replicability" and "reproducibility" therefore play a pivotal role. In the machine learning literature, however, there is currently no clear consensus on these notions (Plesser, 2018; Bouthillier et al., 2019; Gundersen, 2021; Sonnenburg et al., 2007); sometimes, they are used interchangeably. Gundersen et al. (2022) defines reproducibility as "[...] the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators, and a reproducibility experiment is the experiment conducted by independent researchers to confirm research findings." By stating that "[r]eproducibility requires changes; replicability avoids them," Drummond (2009) draws a clear distinction between these two notions: reproducibility means that the initial results can be corroborated via *different* methods or experiments, whereas replicability means that the *same* experiments can be carried out with the same methods. Drummond's understanding of reproducibility is similar to what Bouthillier et al. (2019)[p.727] call *inferential reproducibility* (i.e., "a finding or a conclusion is reproducible if one can draw it from a different experimental setup." [p.727]). They state that this type of reproducibility is partly based on statistical significance. Bouthillier et al. (2019) further distinguish between *methods reproducibility* (a method is reproducible if by reusing the original code, we can obtain the same results) and *results reproducibility* (i.e., a result is reproducible if a reimplementated method leads to statistically similar values). Bouthillier et al. (2019) speculate that in machine learning, reproducibility mostly refers to methods reproducibility, for which code sharing is the commonly proposed solution. The Association for Computing Machinery (2020) distinguishes between *repeatability* (the same researchers can reliably repeat their own experiments), *replicability* (different researchers can obtain the same results using the same experiments and data that another team used), and *reproducibility* (different researchers use different experiments and data to reach the same conclusions). Pineau et al. (2021) distinguish between reproducible and replicable research as follows: an experiment is reproducible if the same data, the same analytical tools, and the same experimental designs are used to reach the same conclusion, whereas different data (but sampled from similar distributions) are used in replicable work. This notion of replicability is practically identical to the definition that we adopt in this article.

Over recent years, replicability and reproducibility have been of particular concern in many scientific disciplines, specifically psychological (Klein et al., 2018) and biomedical research (Eisner, 2018). Alarmingly, a survey by the journal *Nature* reported that more than 70% of researchers have tried—and failed—to reproduce another scientist's experiments (Baker and Penny, 2016). The reproducibility crisis also affects AI (Hutson, 2018) and machine learning (Semmelrock et al., 2023). To address this crisis, various changes to the academic practice have been proposed, such as sharing of code and data, as well as study pre-registration (Nosek et al., 2018). Pre-registration has also been suggested in machine learning (Berrar and Dubitzky, 2019; Cockburn et al., 2020).

One might assume that replicability and reproducibility are less of a problem in machine learning than in other disciplines, given that in general, in silico experiments can be more easily re-run than wet lab experiments. Furthermore, sharing of data, code, and even workflows has long been common practice in machine learning (Sonnenburg et al., 2007; Vanschoren et al., 2013). There are also initiatives like the Machine Learning Reproducibility Challenge that invite the research community to reproduce the computational experiments in accepted papers. However, Gundersen et al. (2022) reported that many results pub-

lished at the top venues in machine learning were not reproducible. They described various factors relating to study design, algorithms, implementation, observation, evaluation, and documentation that contribute to this irreproducibility. Pineau et al. (2021) identified three main challenges that stand in the way of proper replicability in machine learning. First, new methods are often not better than established ones when a more exhaustive hyperparameter search is performed (Melis et al., 2018) or when a different random initialization is chosen (Bouthillier et al., 2019). Second, the experimental protocols often lack sufficient details to reproduce the results. Third, proper statistical analysis is not always conducted to corroborate the experimental findings. Similarly, Henderson et al. (2018) emphasize the need for proper significance testing and tighter standardization of experimental reporting.

Whereas these studies make a case in favor of significance testing for bolstering reproducibility, there have also been opposing viewpoints, arguing that the current practice of significance testing is in fact a key contributing factor to the reproducibility crisis (Nuzzo, 2014; Gibson, 2020), including in machine learning (Berrar and Dubitzky, 2019). Null hypothesis significance testing (NHST) has indeed caused many heated debates in the statistics community for more than six decades, and there is still no consensus on the proper role of the p-value in research practice (Wasserstein et al., 2019), with a spectrum of viewpoints ranging from praising NHST (Hagen, 1997) to calling for the abandonment of significance testing (McShane et al., 2019). Numerous articles have reported that the p-value is widely misinterpreted (Stang et al., 2010; Goodman, 2008), for instance, as a Bayesian posterior probability (Senn, 2002). Still, there is also the argument that the p-value can be a meaningful measure, provided that it is used properly and sensibly (Mulaik et al., 2016; Lakens, 2021; Colquhoun, 2017; Harrington et al., 2019).

In this work, we are concerned with the chances of replicating a significant result. Goodman (1992) addressed the same problem, but only for the case that the observed difference has a normal distribution and under the strong assumption that the standard errors of the observed effect under the null and the alternative hypothesis are equal. In our study, we do not make these assumptions. Goodman's analysis was not underpinned by experimental results, but it convincingly demonstrated that the replication probability can be substantially lower than expected. Miller (2009), too, investigated the probability of replicating a significant effect; however, Miller (2009) concluded that neither the individual nor the aggregate replication probability could be determined reliably and therefore advised against their use. Colquhoun (2017) investigated the connection between the p-value and the reproducibility by deriving a formula for the probability that the null hypothesis is true despite a significant p-value, which is referred to as the false positive risk (FPR). Although this is an interesting probability that adds to the interpretation of the p-value, it does not tell us the probability that a replication experiment will again give a significant result. Sackrowitz and Samuel-Cahn (1999) and Cumming (2006, 2008) investigated the probability of obtaining a p-value for a given population effect and reported a large variation of replicated p-values; thus, it is quite likely that the p-value from a replication is different from the p-value of the initial experiment. Cumming (2006) warns that researchers should therefore not place too much weight on any single p-value.

2. Theoretical Analysis of Replication Probability

In Fisherian significance testing, there is only *one* hypothesis, the null hypothesis H_0, whereas in Neyman-Pearsonian hypotheses testing, there are *two* hypotheses, the null hypothesis H_0 and the alternative hypothesis H_1. In the context of significance testing, we may nonetheless talk about "the alternative hypothesis" when we refer to "not H_0."

The probability of obtaining a significant result, given that there really is a difference in performance (i.e., the null hypothesis of equal performance is false), is the same as the power of the experiment, power $= \mathbb{P}(H_0 \text{ rejected} \mid H_1 \text{ true})$. The type I error is a pre-experimentally fixed error rate, $\alpha = \mathbb{P}(H_0 \text{ rejected} \mid H_0 \text{ true})$. The type II error rate is $\beta = \mathbb{P}(H_0 \text{ not rejected} \mid H_1 \text{ true})$, so power $= 1 - \beta$. The p-value and power belong to two different schools of thought: the former belongs to the Fisherian significance testing, whereas the latter belongs to the Neyman-Pearsonian hypothesis testing (Berrar, 2017). But the true, unknown effect size has an influence on both of them. Note that the p-value is not an error probability, unlike α (Hubbard and Bayarri, 2003).

Power depends on three factors (Fraley and Marks, 2007; Schmidt and Hunter, 2016). First, the larger the type I error, α, the smaller the type II error, β, and therefore the larger the power, everything else being equal. Second, the larger the population effect size (in our context, the true difference in performance between A and B), the larger the power. Third, the larger the sample size n, the higher the precision with which we measure the effect size, and hence the power is larger. In an exact replication, n and α are the same as in the initial experiment. The true effect size is also fixed, as it is a population parameter. Therefore, the power in the replication is exactly the same as in the initial experiment. In other words, the probability of obtaining a significant result in the replication is exactly the same as the probability of obtaining a significant result in the initial experiment. One can think of the initial and follow-up experiment as two independent realizations from the infinite pool of possible replications; the order in which they are carried out is irrelevant. In fact, the initial experiment could be regarded as a replication of the follow-up experiment.

But doesn't the p-value reveal something about the chances of successfully replicating the initial experiment? For example, before we run the experiment comparing classifiers A and B, we are agnostic about the outcome—A could be better than B, or B could be better than A, or their performance could be the same. Once we carry out the experiment and observe that A significantly outperforms B with, say, $p = 0.01$, the aggregate replication probability indeed changes because each new outcome changes the total number of all results that enter its calculation (Miller, 2009). However, the individual replication probability—which is of interest here—depends only on the power and is therefore constant. Whatever the p-value from the initial experiment is, it does not *change* that probability. But the p-value does *inform* us about what that probability might be (see also (Miller, 2009)).

As we will see later, to calculate the replication probability analytically, it is necessary to know the true *effect size*, i.e., the amount or magnitude of something of interest (Cumming, 2012), for example, the true difference in performance between two classifiers, or the true probability that one classifiers performs better than the other. In practice, the true effect size is virtually always unknown. The standard approach then is to use the observed effect size as the best estimate for the true effect size (Goodman, 1992; Miller, 2009). The effect size can be defined in different meaningful ways; for example, when we test the null

hypothesis $H_0 : \theta = 0.5$ (i.e., the probability that one classifier performs better than the other one is 0.5), the effect size can be stated as $\delta = |\theta - 0.5|$, and it is estimated as $\hat{\delta} = |\hat{\theta} - 0.5|$, where $\hat{\theta}$ is the observed success rate (i.e., how often A was better than B, divided by the number of comparisons). Another meaningful effect size is the true difference δ in performance in the population of datasets over which the classifiers were compared.

If the null hypothesis of equal performance is true, then the probability of a significant result in the initial experiment is α. With a two-tailed test, the probability of a significant result in the follow-up experiment is $\alpha/2$, since half of the significant results go in the wrong direction. In keeping with common practice, we consider $p < 0.05$ a significant result. We define the *true replication probability of a significant result* as follows.

Definition 1. True replication probability of a significant result.
Let $\mathcal{M}$ denote a suitable statistical model to assess the statistical significance of the results of an initial experiment E_1. The *true replication probability* under the model $\mathcal{M}$ is the conditional probability that a replication of an initial experiment, E_{1s}, with a significant result gives again a significant result in the same direction. Let E_1 denote the initial experiment and E_2 denote its replication. Then

$$\mathcal{P}_{\mathcal{M}} = \mathbb{P}(E_{2s} \mid E_{1s}) \tag{2}$$

where the index s indicates significance in the same direction. E_2 is an exact replication of E_1, except that different data from the same population were used.

In this definition, the term "suitable statistical model" refers to a statistical significance test that is appropriate for the problem at hand. For example, to assess the significance of the differences in performance between two classifiers in k-fold cross-validation, the resampled variance-corrected t-test is a suitable model, whereas the standard t-test is not (Nadeau and Bengio, 2003).

The replication probability depends on the statistical model because there often exist different statistical tests that are suitable, usually with different power or different assumptions. In the real world, the true replication probability is unknown, but it can be estimated empirically by using the relative frequency of all those replications that turn out to be significant. In the remainder of this article, we refer to this probability as the *empirical replication probability of a significant result*, which we define as follows.

Definition 2. Empirical replication probability of a significant result.
Let $\mathcal{M}$ denote a suitable statistical model to assess the statistical significance of the results of an initial experiment E_1. The *empirical replication probability* under the model $\mathcal{M}$ is defined as

$$\mathcal{F}_{\mathcal{M}} = \frac{|\mathbf{E}_{2s}|}{|\mathbf{E}_2|} \tag{3}$$

where $\mathbf{E}_{2s}$ denotes the set of all replications that are significant in the same direction as E_{1s}, and $\mathbf{E}_2$ is the set of all replications (significant or not), and $|\cdot|$ denotes the number of elements in a set.

Equation 3 is conceptually equivalent to what Miller (2009)[p.618] called the "individual replication probability."

In practice, however, we usually have only *one* study at hand, that is, the benchmark study that we carried out. If we obtain a significant result, it is reasonable to ask about the chances that we observe again a significant result in a follow-up study. Clearly, replicating the initial study multiple times with different datasets from the same population would be ideal, but this is of course generally not feasible. The challenge, therefore, is to estimate the replication probability analytically based on only the information from the initial study and under some model assumptions. This leads to the following definition.

Definition 3. Estimated replication probability of a significant result.
Let $\mathcal{M}$ denote a suitable statistical model to assess the statistical significance of the results of an initial experiment E_1. The *estimated replication probability* under the model $\mathcal{M}$ is the conditional probability that a replication of an initial experiment, E_{1s}, with a significant result gives again a significant result in the same direction. Let E_1 denote the initial study and E_2 denote its replication. Then

$$\widehat{\mathcal{P}}_{\mathcal{M}} = \mathbb{P}(E_{2s} \mid E_{1s}, \mathcal{A}) \tag{4}$$

where $\mathcal{A}$ denotes the assumptions regarding the effect size and the distribution under the alternative hypothesis; the index s indicates significance in the same direction. E_2 is an exact replication of E_1, except that different data from the same population were used.

Next, we derive formulas for the replication probability that are applicable in two of the most widely used study designs in machine learning: (i) the comparison of two classifiers over multiple datasets, and (ii) the comparison of two classifiers in cross-validation.

2.1 Comparison of Two Classifiers Over Multiple Datasets

For the comparison of two classifiers over multiple datasets, we consider three different statistical models: a binomial model, a Bayesian model, and a model based on the Wilcoxon signed rank test (short, Wilcoxon model). We consider the binomial and the Bayesian model mainly for introductory purposes, as these models assume only wins and losses and ignore ranks; they are therefore of limited practical use. In practice, the Wilcoxon signed rank test is preferable (Benavoli et al., 2017).

2.1.1 Replication probability under a binomial model

Let A and B be two classifiers whose performance is compared over N different benchmark datasets. In the simplest scenario, the statistical model assumes that no equal performance of A and B on the same dataset is observable; hence, there is a clear "winner" for each

dataset. Assuming a binomial model, we can now derive a point estimate for the replication probability and the limits of an approximate prediction interval.

Lemma 1. Estimated replication probability based on the binomial model. Let $\hat{\theta} = \frac{x}{N}$ denote the observed success rate for the comparison of two classifiers, A and B, over N benchmark datasets, where x is the number of datasets for which A performed better than B (i.e., a success or win). It is assumed that $\hat{\theta}$ is the best estimate for θ. The point estimate of the replication probability based on the binomial model is then

$$\widehat{\mathcal{P}}_b = \sum_{i=s}^{N} \text{Bin}(i; N, \hat{\theta}) \tag{5}$$

where s is the smallest number of successes for a significant result in the same direction under the null hypothesis, and Bin is the binomial mass function.

An approximate $(1-c)100\%$ prediction interval for the replication probability is given by

$$\left[\sum_{i=s}^{N} \text{Bin}(i; N, \hat{\theta}_{\text{low}}), \quad \sum_{i=s}^{N} \text{Bin}(i; N, \hat{\theta}_{\text{high}})\right] \tag{6}$$

where $\hat{\theta}_{\text{low}}$ and $\hat{\theta}_{\text{high}}$ are the lower and upper limits, respectively, of the $(1-c)100\%$ Clopper-Pearson confidence interval for θ.

Proof. The null hypothesis is stated as $H_0 : \theta = 0.5$, where θ denotes the true probability that one classifier performs better than the other on the population of datasets from which the N benchmark datasets are a random sample. Under the assumed statistical model, the number i of successes (or wins) of A follows a binomial distribution, with probability mass function $\text{Bin}(i; N, \theta) = \binom{N}{i}\theta^i(1-\theta)^{N-i}$. The test is two-sided, since a priori A could be better than B or vice versa.[1] Without loss of generality, it is assumed that in the initial experiment, A significantly outperformed B. To obtain again a significant result in the same direction as in the initial experiment, at least s wins of A must be observed in the replication. Assuming that the observed success rate $\hat{\theta}$ is the best estimate for θ, the probability of observing at least s successes is that shown in Equation 5.

Let $\hat{p} = \frac{r}{n}$ denote a proportion estimated from a sample of size n. An exact $(1-c)100\%$ Clopper-Pearson confidence interval for the population proportion p is then given by

$$\left[\frac{r}{r + (n-r+1)F_{1-\frac{1}{2}c;2(n-r+1),2r}}, \frac{(r+1)F_{1-\frac{1}{2}c;2(r+1),2(n-r)}}{(n-r) + (r+1)F_{1-\frac{1}{2}c;2(r+1),2(n-r)}}\right] \tag{7}$$

where $F_{1-\frac{1}{2}c;\nu_1,\nu_2}$ is the quantile function of the F-distribution with ν_1 and ν_2 degrees of freedom (Clopper and Pearson, 1934). A $(1-c)100\%$ Clopper-Pearson confidence interval

1. All tests in the present work are two-sided.

for θ has the limits $\hat{\theta}_{\text{low}}$ and $\hat{\theta}_{\text{high}}$, so the success rate could plausibly be as low as $\hat{\theta}_{\text{low}}$ or as high as $\hat{\theta}_{\text{high}}$, which leads to Equation 6. ■

Example 1. Let us assume that an experiment involves 44 benchmark datasets. Suppose that A performs better than B on exactly 29 out of 44 datasets (Figure 1). The two-tailed p-value is $2\sum_{i=29}^{44} \text{Bin}(i; 44, 0.5) = 0.0488$, which is a significant result, assuming that the common threshold of 0.05 is used. Now, what is the probability of observing a significant win of A over B in an exact replication? With 29 successes in 44 trials in the initial experiment, the success probability is estimated as $\hat{\theta} = \frac{29}{44} = 0.66$. To obtain again a significant result that goes in the same direction, we need to observe at least 29 successes in the replication experiment. Assuming that the success probability is 0.66 (or, equivalently, that the effect size is 0.16), the probability of observing 29 or more successes in the replication is therefore $\sum_{i=29}^{44} \text{Bin}(i; 44, 0.66) = 0.5746$. This is already quite a remarkable result: despite the significant p-value of 0.048 from the initial experiment, the chances that the results replicate are only about 57%. Thus, the predictive value of the p-value of 0.048 is not much higher than that of a coin toss.

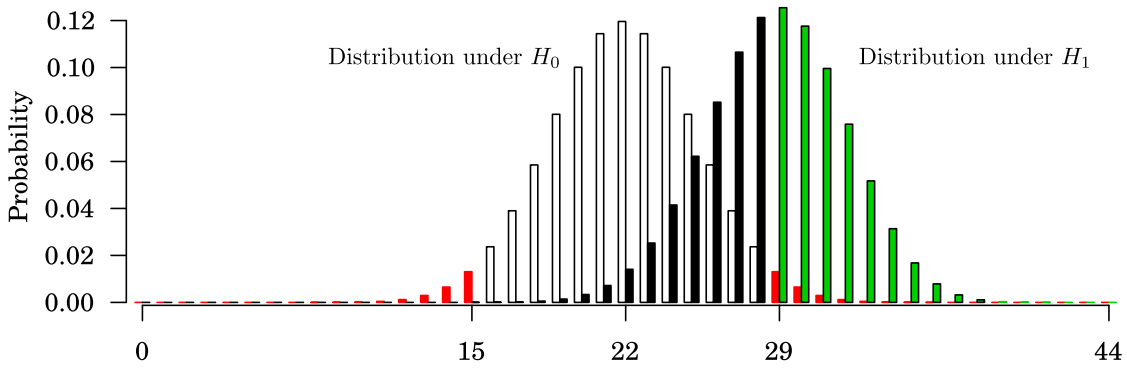

Figure 1: Binomial distribution of successes for A under the null hypothesis H_0 of no difference (white bars) and the alternative hypothesis H_1 with an estimated success probability of $\hat{\theta} = \frac{29}{44}$. For $H_0 : \theta = 0.5$ and $\alpha = 0.05$, significant outcomes are $\{0...15\}$ and $\{29...44\}$ (red bars). For the success rate of $\frac{29}{44} = 0.66$, the replication probability is $\sum_{i=29}^{44} \text{Bin}(i; 44, 0.66) = 0.57$ (green bars).

For 29 successes in 44 trials, the exact 95%-confidence interval for the true success probability is $[0.501, 0.795]$. So in the worst case, the replication probability could be as low as $\sum_{i=29}^{44} \text{Bin}(i; 44, 0.501) = 0.0250$, and in the best case, it could be as high as $\sum_{i=29}^{44} \text{Bin}(i; 44, 0.795) = 0.9890$. This interval is certainly too wide to be of any practical use.

2.1.2 Replication probability under a Bayesian model

The estimated replication probability based on a Bayesian model is a simple extension of Lemma 1.

Lemma 2. Estimated replication probability based on a Bayesian model. Let $\hat{\theta} = \frac{x}{N}$ denote the observed success rate for the comparison of two classifiers, A and B, over N benchmark datasets, where x is the number of datasets for which A performed better than B (i.e., a success). It is assumed that $\hat{\theta}$ is the best estimate for θ. Assuming a uniform prior distribution for the success probability, the point estimate of the replication probability based on a Bayesian model is

$$\widehat{\mathcal{P}}_B = \sum_{i=s}^{N} \text{Bin}(i; N, \hat{\theta}_m) \tag{8}$$

where s is the smallest number of successes for another significant result in the same direction under the null hypothesis, and $\hat{\theta}_m = \frac{1+x}{2+N}$.

An approximate $(1-c)100\%$ prediction interval for the replication probability is given by

$$\left[\sum_{i=s}^{N} \text{Bin}(i; N, \hat{\theta}_{\text{mlow}}), \quad \sum_{i=s}^{N} \text{Bin}(i; N, \hat{\theta}_{\text{mhigh}})\right] \tag{9}$$

where $\hat{\theta}_{\text{mlow}}$ and $\hat{\theta}_{\text{mhigh}}$ are the lower and upper limits, respectively, of the $(1-c)100\%$ highest density interval (HDI) of the posterior distribution of the success probability.

Proof. Assuming a uniform prior distribution for the success probability, θ, of $\text{Beta}_{\text{prior}}(\alpha = 1, \beta = 1)$, the posterior distribution is $\text{Beta}_{\text{post}}(\alpha = 1 + x, \beta = 1 + N - x)$ after observing x successes or wins of A out of N comparisons. This posterior has the mean $\hat{\theta}_m = \frac{\alpha}{\alpha+\beta} = \frac{1+x}{2+N}$. For another significant result in the same direction as in the initial experiment, at least s wins of A need to be observed. This leads to Equation 8.

The interval of the most credible values for the success probability is the Bayesian posterior highest density interval (HDI). A 95%-HDI covers 95% of the posterior distribution, and any value inside the interval has a higher credibility than any value outside of the interval (Kruschke, 2018). Using the limits of the $(1-c)100\%$ HDI as the worst-case and best-case estimates for θ, respectively, we obtain the limits as show in Equation 9. ■

Example 2. For 29 successes (i.e., "wins" of A) and 15 failures (i.e., "wins" of B), the posterior distribution is $\text{Beta}_{\text{post}}(\alpha = 1 + 29, \beta = 1 + 15)$, assuming a uniform prior. The probability that A performs better than B on a new, randomly selected dataset is the mean of the posterior beta distribution, $\frac{\alpha}{\alpha+\beta} = \frac{30}{46} = 0.652$. The point estimate for the replication probability is therefore $\sum_{i=29}^{44} \text{Bin}(i; 44, 0.652) = 0.5311$. We obtain a 95%-HDI of $[0.515, 0.785]$ for θ. This means that the success rate could be as low as 0.515 or as high as 0.785; consequently, the replication probability could be as low as $\sum_{i=29}^{44} \text{Bin}(i; 44, 0.515) =$

0.0384 or as high as $\sum_{i=29}^{44} \text{Bin}(i; 44, 0.785) = 0.983$. Like before, this interval is too wide to be of any practical use.

Figure 2 shows the point estimates for the replication probabilities based on the binomial model and simple Bayesian model, together with their upper and lower limits, for all 16 possible significant p-values from the comparison of two classifiers over 44 datasets where A is better than B (assuming that only clear wins and losses are possible).

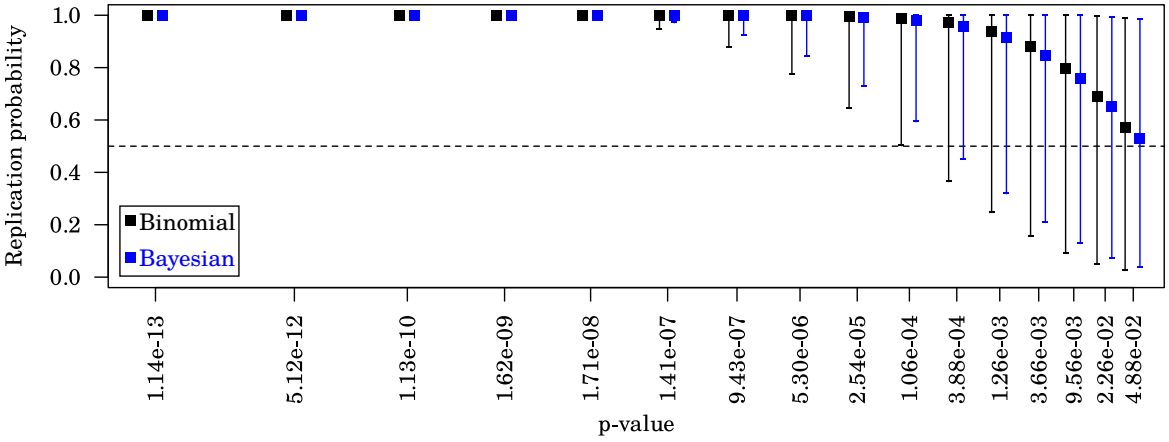

Figure 2: Estimated replication probability as a function of all significant p-values for the comparison of two classifiers over $N = 44$ benchmark datasets. The point estimates (black squares) are based on a binomial test. The Bayesian point estimates (blue squares) assume a uniform prior. Vertical lines represent 95% prediction intervals.

There are three important observations. First, the point estimates based on the Bayesian model are slightly lower than the estimates based on the binomial model, and with slightly narrower intervals. Second, for the p-value of 0.0488 (corresponding to 29 successes), both estimates for the replication probability are quite close to 0.5. This indicates that a p-value just below the significance threshold of 0.05 does not give much confidence that the initial significant result replicates—it may or may not, and the chances are just above 50%. Third, for p-values of 0.000388 or larger, the intervals of plausible values for the replication probability are too wide to be of any practical use. Note that these intervals include 0.5. Even for the p-value of 0.00366, the replication probability based on the binomial model could be as low as 0.155. Only for extremely small p-values ≤ 0.000106, corresponding to 35 or more wins for A, the lower limit is above 0.5.

What if the difference between A and B is *not* significant in the initial experiment? For example, let us assume that A outperforms B on exactly 24 out of 44 datasets. It seems that A performs slightly better than B, but the result is not significant with a p-value of 0.6516. Could the replication reveal that A *is* in fact significantly better? With the point estimate for the success probability of $\hat{\theta} = \frac{24}{44} = 0.5455$, the probability of seeing a significantly better performance of A in the replication is $\sum_{i=29}^{44} \text{Bin}(i; 44, 0.5455) = 0.0856$. But the upper limit of the 95%-CI for θ is 0.6961, so the replication probability might be as high as

0.7606. For the Bayesian upper limit, the probability that A performs significantly better than B is 0.7057. Although the results from the initial experiment (24 wins for A) are in quite good agreement with the null hypothesis, the chances that we will see a significantly better performance of A in the replication could be as high as about 71-76%.

2.1.3 Replication probability under the Wilcoxon model

Both the binomial and the Bayesian model are arguably too simple, as they consider only binary outcomes (win and loss) and do not take into account the differences in performances across different datasets. The Wilcoxon signed rank test (Wilcoxon, 1945; Sheskin, 2007) is a widely used alternative (Benavoli et al., 2016). Briefly, the test first calculates the difference in performance per dataset. Those datasets for which both classifiers achieve the same performance are discarded from further analysis. The differences are then ranked from smallest to largest, and each difference obtains a signed rank. The null hypothesis is stated as follows: the median of the differences in scores, δ, is 0; that is, $H_0 : \delta = 0$. If the null hypothesis is true, then the sum of the absolute ranks of the positive scores, W^+, should be the same as the sum of the absolute scores of the negative ranks, W^-. If, however, the null hypothesis is not true, then these sums should be different. This corresponds to a two-tailed test. The Wilcoxon statistic is a discrete random variable, and by considering its probability distribution, an exact p-value can be calculated, provided that there are no tied ranks. Assuming that n_e is sufficiently large, the Wilcoxon statistic is approximately normally distributed and is standardized as shown in Equation 10,

$$Z_w = \frac{W_L - \frac{1}{4}n_e(n_e+1) - \frac{1}{2}}{\sqrt{\frac{n_e(n_e+1)(2n_e+1)}{24} - \frac{\sum t^3 - \sum t}{48}}} \sim \mathcal{N}(0,1) \tag{10}$$

where n_e is the effective sample size, that is, $n_e = N - d$, where N is the number of datasets and d is the number of datasets for which A and B performed the same. The term $-\frac{1}{2}$ in the numerator is a continuity correction term. The term $\frac{1}{48}(\sum t^3 - \sum t)$ in the denominator is a correction term for ties; here, t indicates the number of tied ranks (Sheskin, 2007, p.233).

Using the Wilcoxon statistic and following the same logic as in Lemma 1 and 2, we can now derive an estimate of the replication probability under the Wilcoxon model.

Proposition 1. Estimated replication probability for the comparison of two classifiers over multiple datasets.
Let Z_w denote the test statistic of the Wilcoxon signed rank test (Equation 10) for the comparison of two classifiers, A and B, over N benchmark datasets. It is assumed that the observed value of Z_w is a good estimate of the true effect. The point estimate of the replication probability is then

$$\widehat{\mathcal{P}}_w = 1 - F\left(z_{1-\frac{1}{2}\alpha}, Z_w, \hat{s}^*\right)$$

$$= 1 - \int_{-\infty}^{1-\frac{1}{2}\alpha} \frac{1}{\sqrt{2\pi}\hat{s}^*} \exp\left\{-\frac{1}{2}\left(\frac{z - Z_w}{\hat{s}^*}\right)^2\right\} dz \tag{11}$$

where F is the cumulative distribution function of the normal distribution with mean Z_w; $\hat{s}^*$ is the bootstrapped standard deviation of the sampling distribution of Z_w; $z_{1-\frac{1}{2}\alpha}$ is a quantile of the standard normal distribution; α is the significance level. For example, for the conventional level of 5%, $z_{0.975} = 1.96$.

The limits of an approximate $(1-c)100\%$ prediction interval for $\widehat{\mathcal{P}}_w$ are

$$\left[1 - F\left(z_{1-\frac{1}{2}\alpha}, Z_w - z_{1-\frac{1}{2}c} \cdot \hat{s}^*, \hat{s}^*\right), \quad 1 - F\left(z_{1-\frac{1}{2}c}, Z_w + z_{1-\frac{1}{2}c} \cdot \hat{s}^*, \hat{s}^*\right)\right] \tag{12}$$

Proof. Assuming that the observed value of Z_w is a good estimate of the true effect, the estimated mean of the distribution under the alternative hypothesis is Z_w. As the significance test in the initial experiment was two-sided, a test statistic larger than or equal to $z_{1-\frac{1}{2}\alpha}$ is required in the replication experiment. As the standard deviation of Z_w under the alternative hypothesis is unknown, it is estimated via bootstrapping. This leads to the formulation of Equation 11. As Z_w could be plausibly as low as $Z_w - z_{1-\frac{1}{2}c} \cdot \hat{s}^*$ or as high as $Z_w + z_{1-\frac{1}{2}c} \cdot \hat{s}^*$, the limits of an approximate $(1-c)100\%$ prediction interval for the replication probability are as shown in Equation 12. ■

Example 3. Suppose that we observe $Z_w = 1.96$ in the initial experiment, which corresponds to a two-sided p-value of 0.05 (red area in Figure 3a). For simplicity, we assume that the distribution under the alternative hypothesis is normal with the same standard deviation as that under the null hypothesis, i.e., $s = 1$. If the distribution under the alternative hypothesis is indeed centered at $Z_w = 1.96$, then the probability is 0.5 that we observe a test statistic of at least 1.96 in the replication study (Figure 3a, blue area), with a 95%-confidence interval of $[0, 3.92]$. So at best, the replication probability is 0.975, and at worst, it is 0.025 (Figure 3b).

However, there is no reason to assume that the standard deviation under the alternative hypothesis is the same as that under the null; in fact, it most likely isn't. Suppose that we observe Z_w in a benchmark experiment of N datasets. Then we create b bootstrap samples for each dataset by random sampling with replacement, train and apply the classifier to

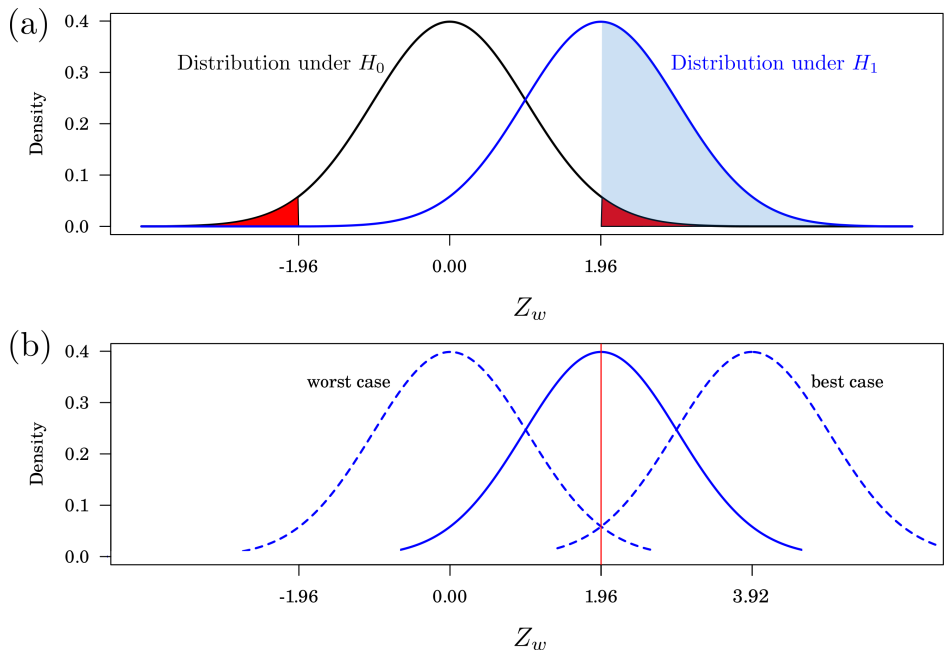

Figure 3: (a) Distribution of the test statistic Z_w under the null hypothesis H_0 of no difference (black bell-shaped curve) and the alternative hypothesis H_1 (blue bell-shaped curve). The mean of the distribution under H_1 is estimated as $\mu_z = 1.96$. Each red area is 0.025, and the blue area is 0.5. (b) Assumed distribution of Z_w under the alternative hypothesis, with estimated worst case ($Z_w = 0$) and best case ($Z_w = 3.92$). For illustrative purposes, the same standard deviation is assumed for both distributions.

the corresponding test sets, and then calculate the bootstrapped estimate of the standard deviation of Z_w (Algorithm 1).

Algorithm 1: Bootstrapped estimate of the standard deviation of Z_w.

```
N ← number of benchmark datasets
b_max ← 300                    // Initialize max number of bootstrap samples per dataset.
for b in 1 to b_max do
    for i in 1 to N do
        create b^th pair of bootstrap samples (T^b, W^b) from i^th pair of training set T_i and test set
          W_i by random sampling with replacement
        train A and B on T^b
        apply A and B to W^b and record performance
    calculate Z_w^b (Equation 10)
ŝ* ← standard deviation of Z_w^b
```

2.2 Comparison of Two Classifiers in k-fold Cross-validation

Two classifiers, A and B, are compared in k-fold cross-validation, and A significantly outperforms B. The replication probability can then be estimated as follows.

Proposition 2. Estimated replication probability in k-fold cross-validation. Let T denote the observed t-value from the variance-corrected resampled t-test (Nadeau and Bengio, 2003) for the comparison of two classifiers, A and B, in k-fold cross-validation, where A performs significantly better than B. It is assumed that T is a good estimate for the mean of the distribution under the alternative distribution. The point estimate of the replication probability is then given by

$$\widehat{\mathcal{P}}_t = 1 - \int_{-\infty}^{t_{\nu,1-\alpha/2}} f_a(t,\mu)\ dt \tag{13}$$

where $f_a(t,\mu)$ is the probability density function under the alternative hypothesis for the non-central t-distribution with $\nu = k-1$ degrees of freedom and non-centrality parameter $\mu = T$, and $t_{\nu,1-\frac{1}{2}\alpha}$ is the quantile of Student's t-distribution for probability of $1-\frac{1}{2}\alpha$.

Let $q_{\frac{1}{2}c}$ and $q_{1-\frac{1}{2}c}$ denote the $\frac{1}{2}c$ and $1-\frac{1}{2}c$ quantiles of the non-central t-distribution under the alternative hypothesis, respectively. A $(1-c)100\%$ prediction interval for the replication probability is then given by

$$\left[1 - \int_{-\infty}^{t_{\nu,1-\frac{1}{2}c}} f_a(t,q_{\frac{1}{2}c})\ dt,\ \ 1 - \int_{-\infty}^{t_{\nu,1-\frac{1}{2}c}} f_a(t,q_{1-\frac{1}{2}c})\ dt\right] \tag{14}$$

Proof. The variance-corrected t-statistic (Nadeau and Bengio, 2003) is calculated as

$$T = \frac{\frac{1}{k}\sum_{i=1}^{k}(a_i - b_i)}{\sqrt{\left(\frac{1}{k} + \frac{|\mathbf{V}|}{|\mathbf{T}|}\right)s^2}}\ \sim\ t_{k-1} \tag{15}$$

where k denotes the number of cross-validation folds; a_i and b_i are the observed performance values (e.g., accuracy) of A and B on the i^{th} validation set; $|\mathbf{V}|$ and $|\mathbf{T}|$ are the sizes of the validation and training sets, respectively; and s^2 is the variance of the performance differences. Under the alternative hypothesis, T follows a non-central t-distribution with $\nu = k-1$ degrees of freedom, where k is the number of cross-validation folds. Assuming that the distribution under the alternative hypothesis has mean T, the probability of observing again a significant result in the same direction as in the initial experiment is the area under the alternative distribution from $t_{\nu,1-\frac{1}{2}\alpha}$ to infinity, which leads to Equation 13. The mean under the alternative hypothesis could plausibly be as low as $q_{\frac{1}{2}c}$ or as high as $q_{1-\frac{1}{2}c}$, which leads to Equation 14. ■

Example 4. Let us assume that in 10-fold cross-validation, we observe a variance-corrected t-statistic of 2.262, which corresponds to a two-sided p-value of 0.05. With $\mu = 2.262$ as the non-centrality parameter for the distribution under the alternative hypothesis and with $\nu = k-1 = 9$ degrees of freedom, the point estimate of the probability of observing again a significant result is

$$\widehat{\mathcal{P}}_t = 1 - \int_{-\infty}^{2.262} f_a(t, \mu = 2.262)\ dt = 0.5235$$

The 2.5% and 97.5% quantiles of the non-centrality t-distribution with $\mu = 2.262$ are 0.305 and 5.489, respectively. The 95% prediction interval for the replication probability is therefore $[0.046, 0.998]$. Figure 4 shows the estimated replication probability as a function of the p-value.

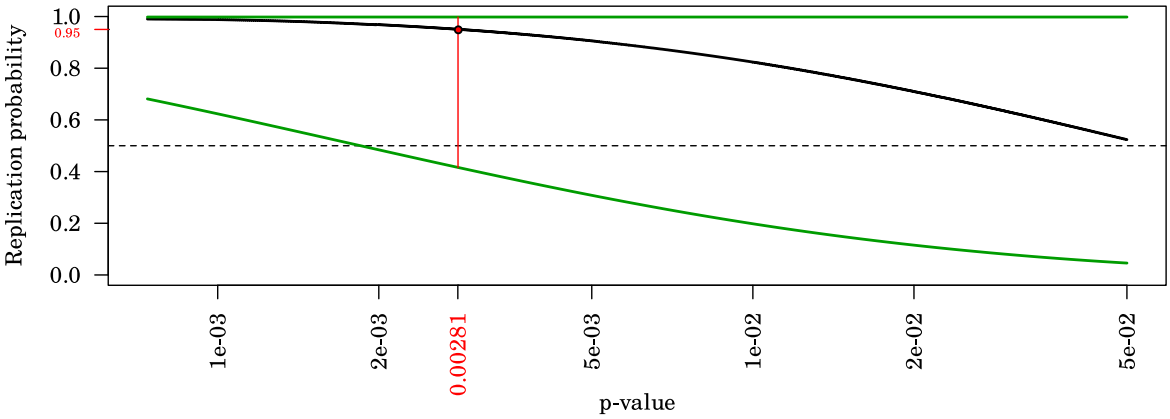

Figure 4: Point estimate $\widehat{\mathcal{P}}_t$ of the replication probability (black solid line) as a function of the p-value based on t-statistics with $\nu = 9$ degrees of freedom (Equation 13). Approximate 95% prediction intervals are marked by green lines (Equation 14). For the p-value of 0.00281, the point estimate of the replication probability is 0.95, with a 95%-prediction interval of $[0.417, 1.000]$. The p-value axis is in log-scale.

Figure 4 suggests that if a replication probability of at least 0.95 is desired, then the p-value should be 0.00281 at most.

3. Empirical Analysis of Replication Probability

The goal of the empirical analysis is to validate the theoretical findings; specifically, we are interested in the following two questions: (*i*) Are the point estimates of the replication probability and the empirical results in relatively good agreement? (*ii*) For which p-value is the replication probability (both theoretical and empirical) close to 0.95?

The main challenge is to create a large number of experiments with varying effect sizes that are associated with smaller and smaller p-values. Each significant experiment can then be repeated multiple times to generate the empirical distributions of p-values, which allows a comparison between the estimated and empirical replication probabilities. We can also find out how small a p-value should be so that the chances of replication are 95%. As this necessitates a repeated sampling from known populations, we used synthetic datasets with known parameters, which we describe below in detail.

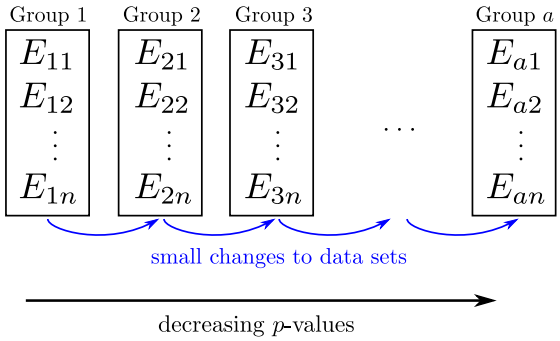

Figure 5: Template of the general study design comprising $i = 1...a$ groups of experiments. Each group i consists of n experiments E_{ij}, $j = 1...n$. Within each group, the experiments are replications of each other. The groups are generated by making small incremental changes to the datasets so that from group i to $i + 1$, smaller and smaller p-values can be observed.

The key idea for the design of our studies is schematically represented by Figure 5. Group 1 comprises n experiments E_{11} to E_{1n}, and each of these experiments can be regarded as a replication of any other experiment from that same group. Then, the idiosyncrasies of the data (e.g., the number of predictive features) were slightly changed so that the p-values decreased. The resulting experiments are contained in group 2. This procedure continued up to group a, which comprises the "most significant" experiments. So here, by "study" we mean a collection of experiments. One experiment is a comparison of two classifiers on datasets with specific characteristics.

The rationale for this somehow unusual design is the following. In a trial-and-error approach, we compared the classifiers on various synthetic datasets with different properties, such as their dimensions (i.e., number of cases and number of features), mean and standard deviations of positive and negative cases, etc. We observed that by modifying these properties, we could create combinations of datasets and learning algorithms that led to increasingly larger performance differences, ranging from non-significant to highly significant.

We carried out four studies. Studies #1 and #2 compare the performance of two classifiers over multiple datasets. Studies #2 and #3 compare the performance of two classifiers in k-fold cross-validation. These studies will be described in the following subsections; for details, see Algorithms 2, 3, and 4 in the Appendices A, B, and C.

As learning algorithms, we used the naive Bayes (NB) algorithm and the support vector machine (SVM), both implemented with the R library `e1071` (Meyer et al., 2022), and random forests (RF), implemented with the R library `randomForest` (Liaw and Wiener, 2002). The concrete algorithms, however, are irrelevant for the analysis. What matters is that we can create groups of experiments with results that range from non-significant to highly significant. The default hyperparameters were used, and no further optimization was performed. All models and experiments were implemented on a standard PC (Intel Core i7-7700T CPU, 2.90GHz × 8, 32GB RAM).

3.1 Comparison of Two Classifiers Over Multiple Datasets

3.1.1 Study #1: Comparison of NB and SVM over 20 datasets

In the first study, we compared NB (classifier A) and SVM (classifier B) in $a = 21$ experimental groups. Each group comprises $n = 1000$ experiments. In each consecutive experimental group, the composition of the datasets was changed in such a way that the performance difference between the classifiers became slightly larger, on average. This means that smaller and smaller p-values could be observed in higher groups (Figure 6). In each group, each experiment involved $N = 20$ synthetic datasets. Figure 7a shows the different sizes of these datasets per experimental group.

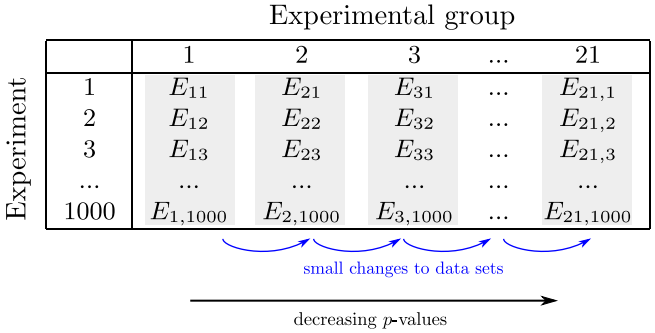

Figure 6: Overview of study #1 comprising 21 experimental groups, each with 1000 experiments. The experiments within the same group are replications of each other. Small changes to the datasets make the experiments in subsequent groups "more significant" than the experiments in the preceding groups, from non-significant (group 1) to highly significant (group 21).

We now describe the incremental changes from group to group (for details, see Algorithm 1, Appendix A). In the first experiment of the first group, E_{11}, the first training set consists of $n_1 = 290 + 10 \cdot 1 = 300$ cases, the second training set consist of $n_2 = 290 + 10 \cdot 2 = 310$ cases, and so on; the 20^{th} training set consists of 490 cases (line 10, Algorithm 2, Appendix A). The number of features is $d_i = i + 10$, that is, 11 for the first training set, $d = 12$ for the second training set, and so on; the 20^{th} and last training set has 30 features. Half of the cases in each training set are positive, the other half are negative. The feature values of the positive cases were sampled from a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$, with $\mu_i = 0.3 + \frac{1}{200}(N+1-i)$ and $\sigma_i = 1 + \frac{1}{50}(N+1-i)$. So for example, the mean and standard deviation for the first training set are $\mu_1 = 0.3 + \frac{1}{200}(20+1-1) = 0.4$ and $\sigma_1 = 1 + \frac{1}{50}(20+1-1) = 1.4$, respectively; for the second training set, $\mu_2 = 0.3 + \frac{1}{200}(20+1-2) = 0.395$ and $\sigma_2 = 1 + \frac{1}{50}(20+1-2) = 1.38$, and so on (lines 12, 13, and 16, Algorithm 2, Appendix A). The feature values for the negative cases were randomly sampled from a standard normal distribution $\mathcal{N}(0, 1)$ (line 22, Algorithm 2, Appendix A). For each training set, a corresponding test set was generated by random sampling from the same distributions. The classifiers were trained on the same training sets and then evaluated on the same test sets (line 28, Algorithm 2, Appendix A). For the performance differences, the success rate and a p-value were then calculated. Here,

(a)

Experimental group

Data set	1	2	* 3	4	5	* 6	7	8	9	* 10	11	12	13	14	15	* 16	17	18	19	20	21
1	300	350	400	450	500	550	600	650	700	750	800	850	900	950	1000	1050	1100	1150	1200	1250	1300
2	310	360	410	460	510	560	610	660	710	760	810	860	910	960	1010	1060	1110	1160	1210	1260	1310
3	320	370	420	470	520	570	620	670	720	770	820	870	920	970	1020	1070	1120	1170	1220	1270	1320
4	330	380	430	480	530	580	630	680	730	780	830	880	930	980	1030	1080	1130	1180	1230	1280	1330
5	340	390	440	490	540	590	640	690	740	790	840	890	940	990	1040	1090	1140	1190	1240	1290	1340
6	350	400	450	500	550	600	650	700	750	800	850	900	950	1000	1050	1100	1150	1200	1250	1300	1350
7	360	410	460	510	560	610	660	710	760	810	860	910	960	1010	1060	1110	1160	1210	1260	1310	1360
8	370	420	470	520	570	620	670	720	770	820	870	920	970	1020	1070	1120	1170	1220	1270	1320	1370
9	380	430	480	530	580	630	680	730	780	830	880	930	980	1030	1080	1130	1180	1230	1280	1330	1380
10	390	440	490	540	590	640	690	740	790	840	890	940	990	1040	1090	1140	1190	1240	1290	1340	1390
11	400	450	500	550	600	650	700	750	800	850	900	950	1000	1050	1100	1150	1200	1250	1300	1350	1400
12	410	460	510	560	610	660	710	760	810	860	910	960	1010	1060	1110	1160	1210	1260	1310	1360	1410
13	420	470	520	570	620	670	720	770	820	870	920	970	1020	1070	1120	1170	1220	1270	1320	1370	1420
14	430	480	530	580	630	680	730	780	830	880	930	980	1030	1080	1130	1180	1230	1280	1330	1380	1430
15	440	490	540	590	640	690	740	790	840	890	940	990	1040	1090	1140	1190	1240	1290	1340	1390	1440
16	450	500	550	600	650	700	750	800	850	900	950	1000	1050	1100	1150	1200	1250	1300	1350	1400	1450
17	460	510	560	610	660	710	760	810	860	910	960	1010	1060	1110	1160	1210	1260	1310	1360	1410	1460
18	470	520	570	620	670	720	770	820	870	920	970	1020	1070	1120	1170	1220	1270	1320	1370	1420	1470
19	480	530	580	630	680	730	780	830	880	930	980	1030	1080	1130	1180	1230	1280	1330	1380	1430	1480
20	490	540	590	640	690	740	790	840	890	940	990	1040	1090	1140	1190	1240	1290	1340	1390	1440	1490

one randomly selected experiment from group 3

(b)

Data set	NB	SVM	Success
1	0.818	0.810	1
2	0.812	0.810	1
3	0.843	0.810	1
4	0.802	0.765	1
5	0.841	0.825	1
6	0.820	0.809	1
7	0.848	0.841	1
8	0.811	0.809	1
9	0.802	0.804	0
10	0.773	0.769	1
11	0.830	0.816	1
12	0.747	0.767	0
13	0.804	0.808	0
14	0.809	0.775	1
15	0.800	0.783	1
16	0.785	0.802	0
17	0.782	0.784	0
18	0.793	0.788	1
19	0.778	0.762	1
20	0.759	0.778	0
avg. acc.	0.803	0.796	Sum = 14
			$\hat{\theta} = 14/20$

binomial p-value p = 0.115
Wilcoxon p-value p = 0.064

(c) All 1000 experiments from group 3

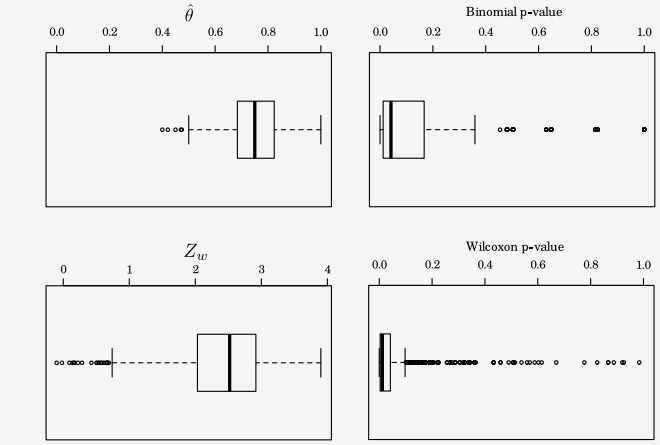

Figure 7: (a) Study #1 consists of 21 experimental groups, each comprising 1000 experiments. The experiments within the same group are replications of each other. Each experiment includes 20 datasets of different sizes, which are shown in the top table; for example, the second dataset of each experiment in group 3 contains 410 cases. (b) This table shows the result from *one* randomly selected experiment from group 3. The data refer to the classification accuracies on the test sets. NB achieved a higher accuracy than SVM on 14 out of 20 datasets. The observed success rate is therefore $\hat{\theta} = 0.7$ (for this particular experiment). The binomial p-value is 0.115, and the Wilcoxon p-value is 0.064 (for this particular experiment). (c) In each group, each experiment is executed 1000 times, with different sampled training and test sets. The boxplots show the observed success rates, $\hat{\theta}$, the binomial p-values, the Wilcoxon statistic Z_w, and the Wilcoxon p-values for the experiments in group 3.

by "success" we mean that NB performed better than SVM on one of the 20 test sets. This procedure was then repeated 1000 times (line 5, Algorithm 2, Appendix A) to generate 1000 replications of the first experiment, leading to 1000 p-values in group 1.

The experiments in the second group were carried out in exactly the same way, except that we slightly increased the size of all datasets: the first dataset had 350 cases, the second dataset had 360 cases, and so on; the 20th dataset had 540 cases. Each experiment was repeated 1000 times. We proceeded analogously for all 21 groups (Figure 7a). The experiments are described in Algorithm 2, Appendix A.

Figure 8 shows the training and test sets of *one* experiment from group 3 in detail.

We will illustrate the calculation of the replication probabilities using the experiments from group 3 (cf. Figure 7). Here, we observe, on average, $N\bar{\hat{\theta}} = 20 \cdot 0.7522 \approx 15$ "wins" of NB out of 20 trials. Under the null hypothesis of equal performance, 15 successes out of 20 trials is just significant, with $p = 2\sum_{i=15}^{N=20} \text{Bin}(i, N, \theta = 0.5) = 0.0414$.

What can we say about the population of experiments from which group 3 is one sample? Note that this population does *not* include the experiments from the other groups shown in Figure 7a. Instead, the population comprises the infinite set of experiments that are *like* the experiments from group 3, but with different randomly sampled cases per training and test set. The 1000 experiments in group 3 represent one random sample from this population for which NB happens to perform slightly better than SVM, on average: the p-value is 0.0414, a marginally significant result.

In a real-world benchmark study, only *one* experiment would be carried out and evaluated based on a significance test (e.g., a comparison of NB and SVM over 20 benchmark datasets from the UCI repository). If the result is significant, then the question is what are the chances of replicating this result. So suppose now that we have at hand *one* such experiment. Let this experiment be from the pool of $k = 1000$ experiments from group 3, and we observe 15 successes in 20 trials ($\hat{\theta} = 15/20 = 0.75$, $p = 0.0414$). What are the chances of obtaining again a significant result in an exact replication? This probability can be estimated by the relative frequency of those experiments in which NB outperformed SVM significantly (i.e., at least 15 times in 20 benchmark datasets). Here, we can simply count these successes. Following Definition 2, we can calculate the empirical replication probability as follows: in 1000 experiments from group 3, NB outperformed SVM significantly $s_{ba} = 537$ times (under the binomial or Bayesian model). Therefore, the observed empirical replication probability for the experiments in group $a = 3$ is

$$\mathcal{F}_{a=3} = \frac{s_{ba} - 1}{k_{\max} - 1} = \frac{537 - 1}{1000 - 1} \approx 0.5365.$$

Assuming the binomial model, the point estimate of the replication probability (Equation 5) is

$$\widehat{\mathcal{P}}_{b,a=3} = \sum_{i=15}^{N=20} \text{Bin}(i, N, \hat{\theta} = 0.75) = 0.6172$$

With the Bayesian model assuming a uniform prior, the estimate of the success rate is $\hat{\theta} = (15+1)/(20+2) \approx 0.7273$, so the point estimate of the replication probability (Equation 8) is

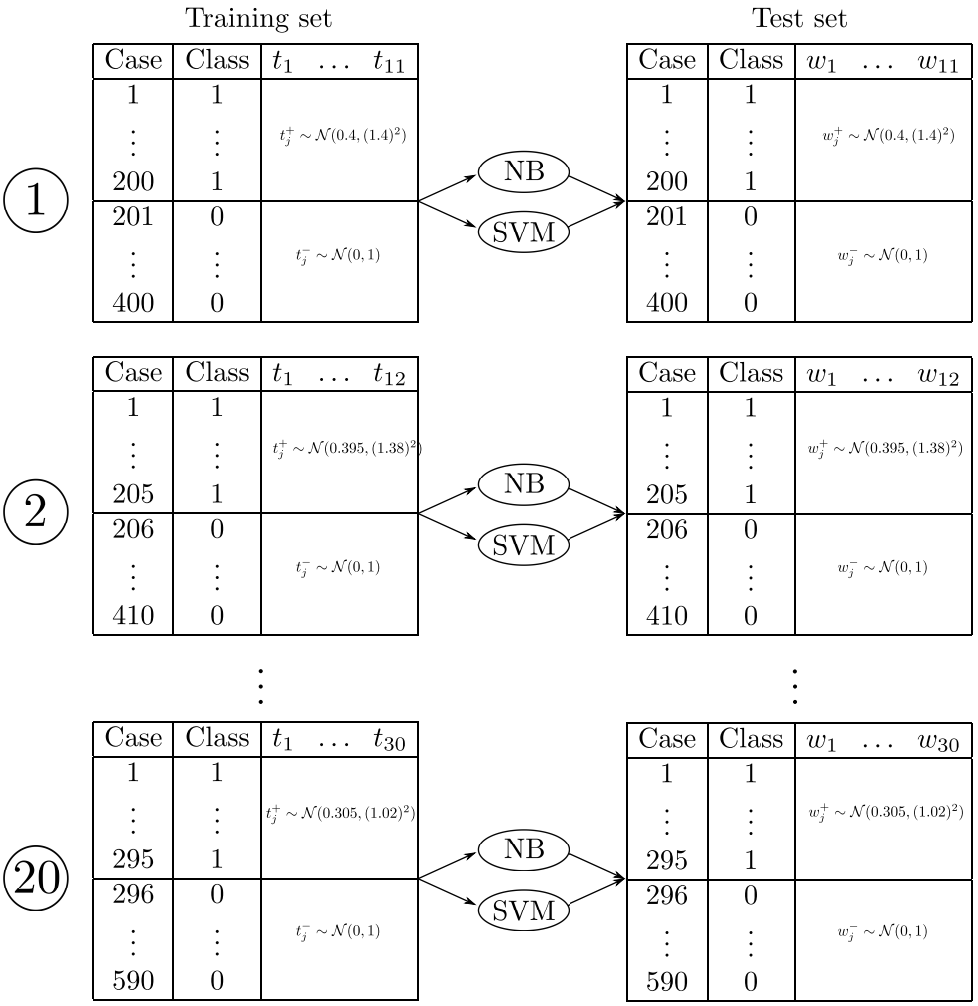

Figure 8: Setup of an experiment from group 3: Comparison of NB and SVM over $N = 20$ different benchmark datasets, ranging from $n_1 = 400$ cases to $n_{20} = 590$ cases (with 50% positives and 50% negatives). The number of features ranges from 11 to 30. The feature values of all negative cases are randomly sampled from a standard normal distribution. The feature values of the positive training cases, t_j^+, and positive test cases, w_j^+, are randomly sampled from $\mathcal{N}(0.4, (1.4)^2)$ in the first dataset; from $\mathcal{N}(0.395, (1.38)^2)$ in the second dataset; and from $\mathcal{N}(0.305, (1.02)^2)$ in the last dataset. Both models are trained on the training data, applied to the same test data, and then evaluated. This process is repeated 1000 times, leading to 1000 experiments E_{3i}, with $i = 1..1000$. For details, see Algorithm 2, Appendix A.

$$\widehat{\mathcal{P}}_{B,a=3} = \sum_{i=15}^{N=20} \text{Bin}(i, N, \hat{\theta} = 0.7273) = 0.5246$$

The estimated replication probability under the Bayesian model is remarkably close to the empirical replication probability. Note that these probabilities are only marginally larger than 0.5.

We proceeded analogously for all remaining groups. Figure 9 shows the replication probabilities (with approximate intervals) based on the binomial and the Bayesian model as a function of the p-value for 15, 16, 17, and 18 (rounded) "wins" of NB over SVM in 20 datasets.

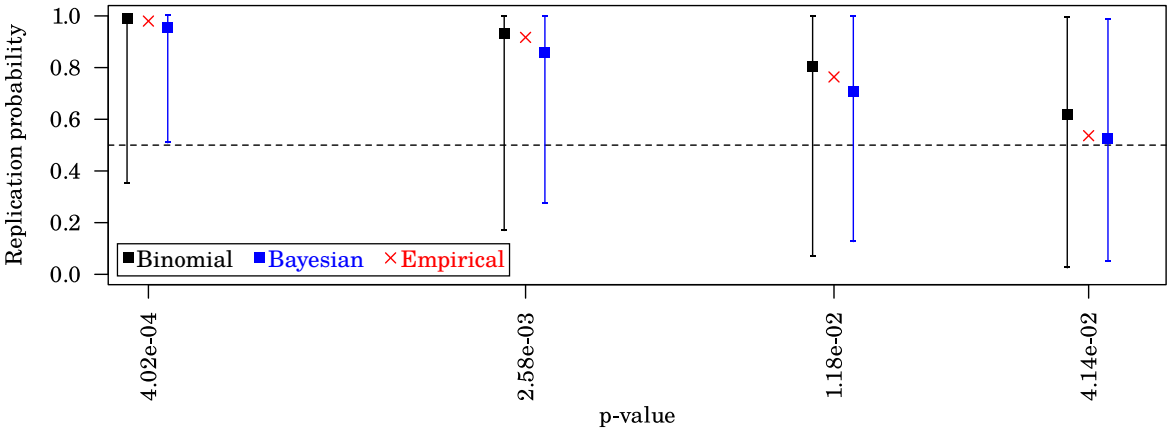

Figure 9: Replication probability as a function of the binomial p-value for benchmark studies with $N = 20$ datasets. The empirical replication probability is marked by the red $\times$. The point estimate of the replication probability under the binomial model, $\widehat{\mathcal{P}}_b$, is marked by the solid black square. The point estimate of the replication probability under the Bayesian model, $\widehat{\mathcal{P}}_B$, is marked by the solid blue square. Vertical lines indicate 95% prediction intervals. The groups that produced these p-values are 3, 6, 10, and 16 (Figure 7a, marked by *). The p-value axis is in log-scale.

As Figure 9 shows, the replication probabilities under the binomial model slightly overestimate the empirical replication probabilities, whereas the replication probabilities under the Bayesian model slightly underestimate them.

Under the Wilcoxon model, the empirical replication probability is calculated as

$$\mathcal{F}_{wa} = \frac{s_{wa} - 1}{k_{\max} - 1} \tag{16}$$

where s_{wa} is the number of significant values of Z_w (with positive sign for significance in the same direction) in group a. For the results shown in Figure 7b, we obtain a p-value of $p = 0.064$. But when we repeat the experiment 1000 times (Figure 7c), we see far more

significant results than with the binomial model. This is not surprising, given that the Wilcoxon test is more powerful. Consequently, the empirical replication probability (under the Wilcoxon model) is much higher. For the experiments in group $a = 3$, we observed $s_{wa} = 778$ times a significant result. Therefore, the empirical replication probability (under the Wilcoxon model) is

$$\mathcal{F}_{w,a=3} = \frac{778 - 1}{1000 - 1} \approx 0.778.$$

Assuming the Wilcoxon model, the point estimate of the replication probability is derived according to Equation 11, with the average of $\bar{Z}_w = 2.437$ and the bootstrapped estimate of the standard deviation, $\hat{s}^* = 0.779$, as

$$\widehat{\mathcal{P}}_{w,a=3} = 1 - F(z_{0.975}, \bar{Z}_w, \hat{s}^*) = 0.730.$$

We proceeded again analogously for the remaining groups. Figure 10 shows the replication probability as a function of the Wilcoxon p-value for all 21 groups. For $p = 0.00274$ (green), the empirical replication probability is $\mathcal{F}_{w,a=7} = 0.9570$, and the estimated replication probability is $\widehat{\mathcal{P}}_{w,a=7} = 0.9665$. The mean absolute deviance between the empirical and analytical probabilities is 0.009792 (range of $3.907 \cdot 10^{-13}$ to 0.04810). For comparison, without bootstrapping (and assuming a standard deviation of 1 for Z_w under the alternative hypothesis), the mean absolute deviance is 0.07318 (range of 0.02496 to 0.10708).

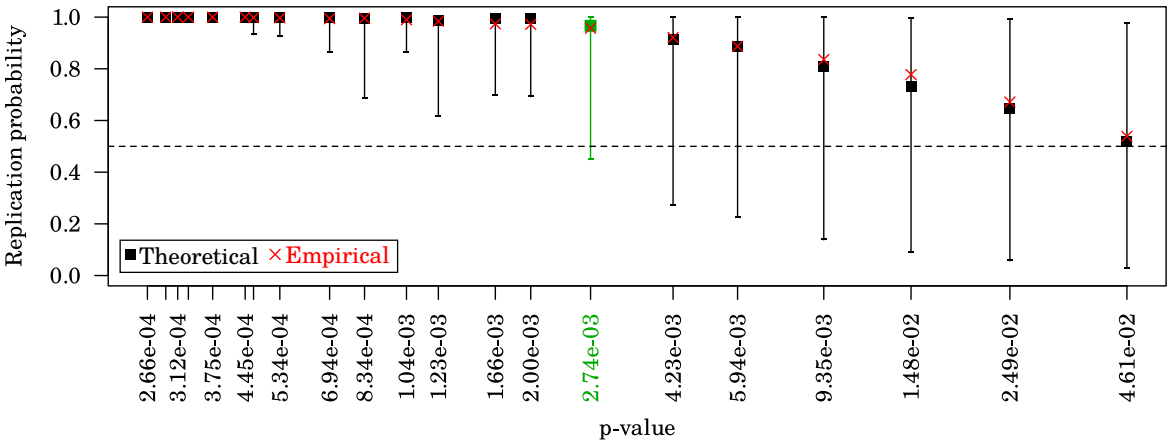

Figure 10: Replication probability as a function of the Wilcoxon p-value for the 21 experimental groups with $N = 20$ datasets each. The empirical replication probability is marked by the red $\times$. The point estimate of the replication probability is marked by a solid square. Vertical lines indicate approximate 95%-prediction intervals. The p-value axis is in log-scale.

In summary, the replication probability based on the Wilcoxon model estimates the empirical replication relatively well, and the agreement becomes better for smaller p-values.

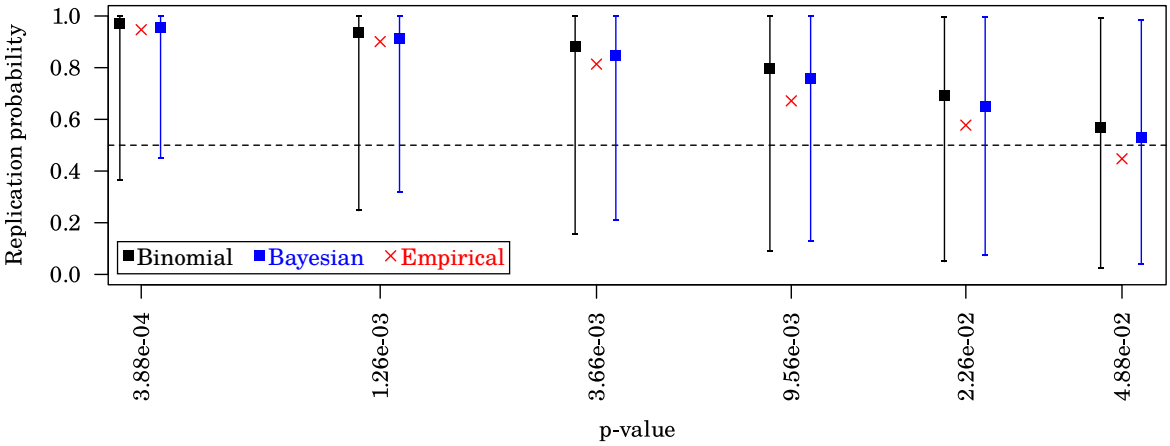

Figure 11: Replication probability as a function of the binomial p-value for experiments with $N = 44$ datasets. Empirical replication probabilities $\mathcal{F}_{wa}$ are marked by red $\times$. The point estimate of the replication probability under the binomial model, $\widehat{\mathcal{P}}_b$, is marked by the solid black square. The point estimate of the replication probability under the Bayesian model, $\widehat{\mathcal{P}}_B$, is marked by the solid blue square. The p-values (from largest to smallest) result from the experiments in groups 2, 3, 4, 7, 8, and 9 (cf. Table 2, Appendix D). Vertical lines are approximate 95% prediction intervals. The p-value axis is in log-scale.

Significant results with p-values just below the common threshold of 0.05 replicate only with chances of around 50%. If a replication probability of more than 0.95 is desired, then the p-value should be no larger than 0.00274.

3.1.2 Study #2: Comparison of NB and SVM over 44 datasets

In study #2, we considered $a = 10$ similar groups of experiments as before, except that we increased the number of datasets per experiment to $N = 44$ and changed the sizes of the datasets as shown in Table 2, Appendix D. The remaining procedure was essentially the same as that for study #1 (Algorithm 2, Appendix A).

Figure 11 shows the replication probability under the binomial and the Bayesian model. Figure 12 shows the replication probability under the Wilcoxon model. Here, for $p = 0.0016$ (green), the empirical replication probability is $\mathcal{F}_{w,a=7} = 0.94394$. The estimated replication probability is $\widehat{\mathcal{P}}_{w,a=7} = 0.96742$. The mean absolute deviance between the empirical and analytical probabilities is 0.0162 (range of 0.00022 to 0.04494).

3.2 Comparison of Two Classifiers in Cross-validation

3.2.1 Study #3: Two identical models, but one has access to an oracle

In study #3, the learning set consists of 1000 cases of two classes (500 positive and 500 negative cases) and 20 numeric features. The feature values of the negative cases were

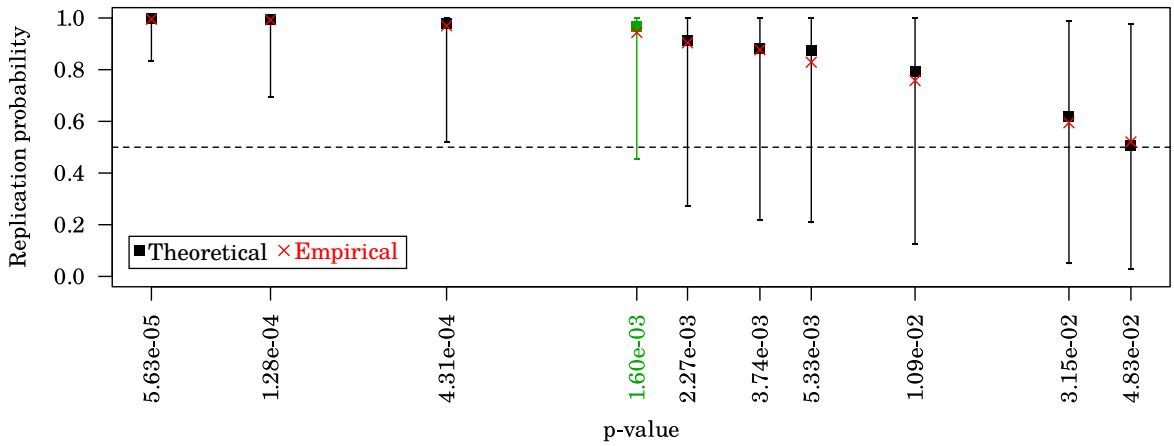

Figure 12: Replication probability as a function of the Wilcoxon p-value for experiments with $N = 44$ datasets. The empirical replication probability is marked by the red $\times$. The point estimate of the replication probability under the Wilcoxon model, $\widehat{\mathcal{P}}_w$, is marked by the solid square. Vertical lines indicate approximate 95%-prediction intervals. The p-value axis is in log-scale.

randomly sampled from $\mathcal{N}(0, 1)$, and the values of the positive cases were randomly sampled from $\mathcal{N}(0.3, 1)$.

We compared two support vector machines, SVM_o (classifier A) and SVM (classifier B), in 10-fold stratified cross-validation. First, SVM_o is an identical copy of SVM. Then, SVM_o was given access to an oracle that revealed the true class labels of a small percentage q of the validation cases, thereby giving an advantage to SVM_o over SVM. This procedure was repeated 1000 times, each time with a newly sampled learning set from the same distribution. Then, we slightly incremented the percentage of class labels that were revealed to SVM_o and repeated the experiments. These percentages were 1%, 2%, 3%, and so on, up to 20%. For example, for $q = 3$, the oracle revealed the true class labels of 3% of the cases in each of the 10 validation sets. Since each validation set had exactly 100 cases, the class labels of three randomly selected cases were revealed per validation set for $q = 3$. Loosely speaking, q acts like a "tuner knob" that controls the significance of the difference in performance: by increasing q, SVM_o is expected to perform better than SVM, so the difference becomes "more significant."

In Study #3, there are $q = 1...20$ groups, each consisting of 1000 experiments. Algorithm 3, Appendix B, describes the experiments in detail.

For $q = 1$ and $q = 2$, the oracle had no significant effect on the performance of SVM_o. In the first experiment with $q = 3$, we observed a variance-corrected t-statistic of $T_{31} = 2.308$, with an associated p-value of $p_{31} = 0.046$. Over 1000 repetitions, each time with newly sampled datasets, the mean t-statistic was $\bar{T}_3 = 2.493$, with the corresponding p-value of 0.0343. Among the 1000 p-values, only 496 were smaller than 0.05, which means that the empirical replication probability for the experiment with $q = 3$ is

$$\mathcal{F}_q = \frac{496 - 1}{1000 - 1} \approx 0.495$$

where we use the subscript q to indicate the experimental group. Using the mean t-value of 2.493 as the point estimate for the effect caused by the oracle, the point estimate for the replication probability for the experiment from group $q = 3$ is $\widehat{\mathcal{P}}_q = 0.604$.

Figure 13 shows the point estimates, $\widehat{\mathcal{P}}_t$ (Equation 13), of the replication probabilities for $q \in \{3, 4, ..., 20\}$, with the estimated upper and lower limits.

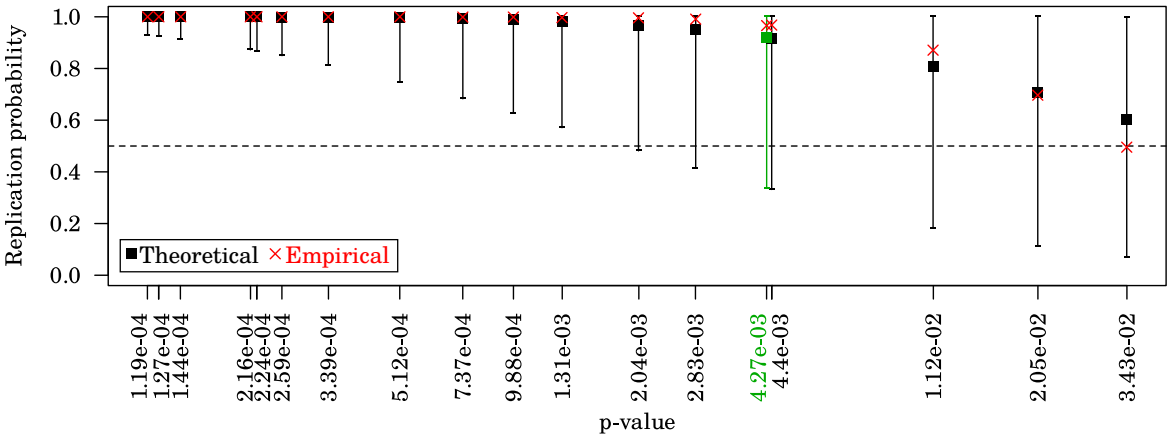

Figure 13: Replication probability of experiments from group $q \in \{3, 4, ..., 20\}$ with 95%-prediction intervals for the comparison of SVM and SVM_o in 10-fold stratified cross-validation. Empirical replication probabilities $\mathcal{F}_q$ are marked by red $\times$. Estimated replication probabilities, $\widehat{\mathcal{P}}_t$, are marked by black squares. The p-value axis is in log-scale.

For $q = 4$, the p-value is 0.0205. The point-estimated replication probability is $\widehat{\mathcal{P}}_{t,q=4} = 0.7054$ and relatively close the empirical replication probability of $\mathcal{F}_q = 0.6967$. For smaller p-values, the point-estimated replication probabilities are slightly lower than the empirical probabilities. If $q = 7\%$ of the class labels are revealed by the oracle ($p = 0.00427$, highlighted in green in Figure 13), then the empirical replication probability is $\mathcal{F}_{q=7} = 0.96597$, and the estimated replication probability is $\widehat{\mathcal{P}}_{q=7} = 0.91996$.

3.2.2 Study #4: Two different models and an increasing feature space

In study #4, we compared a support vector machine (SVM, classifier A) with random forests (RF, classifier B) in 10-fold stratified cross-validation. The learning set consists of 1000 cases of two classes (500 positive and 500 negative cases) and a variable number of numeric features, ranging from 4 to 30. The feature values of the negative cases were randomly sampled from $\mathcal{N}(0, 1)$, and the values of the positive cases were randomly sampled from $\mathcal{N}(0.3, 1)$. First, the number of features was $d = 4$. We repeated the 10-fold cross-validation 1000 times and obtained 1000 p-values for the difference in performance between SVM and RF. These experiments are in group $d = 4$.

Next, the number of features was incremented by 1, and the process repeated, to create the groups of experiments with $d = 5, 6, ..., 20$. Algorithm 4, Appendix C, describes the experiments.

For $d < 11$, there was no difference in performance. Figure 14 shows the point estimates with prediction intervals and the empirical replication probabilities for experiments with $d \in \{11, 12, ..., 30\}$.

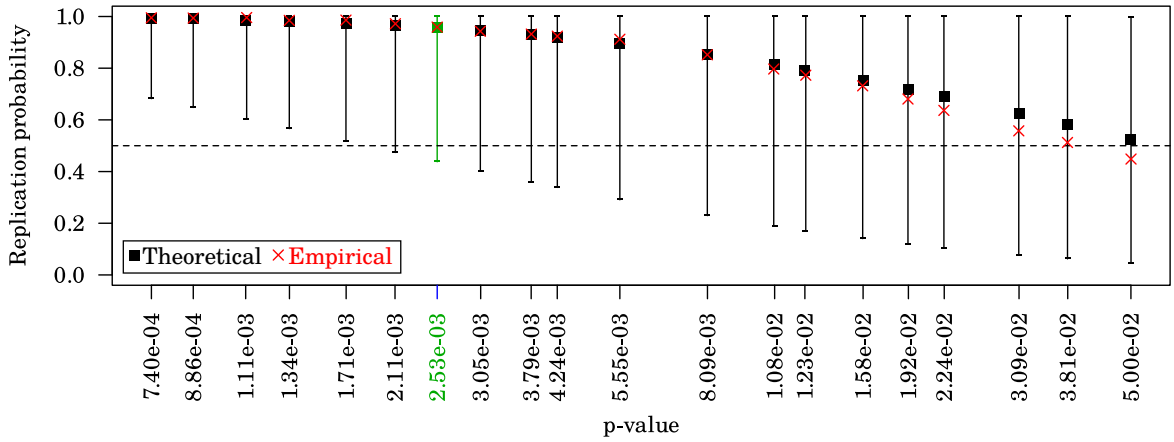

Figure 14: Point estimates (black squares) of the replication probability (Equation 13) for experiments with $d \in \{11, 12, ..., 30\}$, with 95%-prediction intervals (Equation 14), for the comparison of SVM and RF in 10-fold stratified cross-validation. Empirical replication probabilities $\mathcal{F}_d$ are marked by red $\times$. For $p = 0.00253$ (highlighted in green), the empirical replication probability is $\mathcal{F}_{d=24} = 0.95896$ and the estimated replication probability is $\widehat{\mathcal{P}}_{d=24} = 0.95612$. The p-value axis is in log-scale.

Support vector machines are known to perform well for high-dimensional datasets, and we therefore expected that more predictive features would give an advantage to SVM. This was indeed the case: for increasingly larger d, we observed larger and larger mean t-values. So here, the number of features, d, acts like a "tuner knob" controlling the significance.

For $d = 11$, the mean t-value was 2.2619, with a p-value of 0.050 (see the largest p-value in Figure 14). The point-estimated replication probability is $\widehat{\mathcal{P}}_{d=11} = 0.5235$. Among 1000 replications, a significant p-value was observed 449 times; hence, the empirical replication probability is only $\mathcal{F}_{d=11} = (449-1) \,/\, (1000-1) = 0.4484$. Here, the theoretical replication probability overestimates the empirical replication probability.

Of particular interest is the replication probability of 0.95. Our theoretical analysis has revealed that for a p-value smaller than 0.00281, the replication probability is larger than 0.95 (Figure 4). In the experiments of study #4, the closest observed p-value is 0.00253 (highlighted in green in Figure 14), for $d = 24$. For this p-value, the estimated replication probability is $\widehat{\mathcal{P}}_{d=24} = 0.95612$, and the empirical replication probability is $\mathcal{F}_{d=24} = 0.95896$. Here, the empirical and estimated results are in relatively good agreement.

4. Discussion

The p-value gives an answer to the following question: "Assuming that the null hypothesis is true, and given the experimenter's testing and stopping intentions, what are the chances of obtaining a result as extreme as the actually observed result, or an even more extreme result?" By contrast, the replication probability gives an answer to a different question: "What are the chances that a statistically significant result will replicate?" We speculate that this is a fundamental question that many researchers have in mind when they employ a significance test. To address this question, we considered two of the most commonly used designs in supervised learning, that is, the comparison of two classifiers over multiple datasets and the comparison of two classifiers in k-fold cross-validation. For these experimental designs, we derived formulas for the replication probability and evaluated them empirically in four studies, as summarized in Table 1.

Table 1: Summary of the four studies. In studies #1 and #2, the p-values are based on the Wilcoxon test. In studies #2 and #3, the p-values are based on the variance-corrected resampled t-test. In each study, the empirical replication probability closest to 0.95 is shown. $\widehat{\mathcal{P}}$ is the estimated replication probability, $\mathcal{F}$ is the empirical replication probability.

#	Study	Experimental group	p-value	$\widehat{\mathcal{P}}$	$\mathcal{F}$	$\widehat{\mathcal{P}} - \mathcal{F}$
1	NB vs. SVM, 20 datasets	$a = 7$	0.00274	0.96645	0.95696	0.00949
2	NB vs. SVM, 44 datasets	$a = 7$	0.00160	0.96742	0.94394	0.02348
3	SVM vs. SVM_o, 10-fold CV	$q = 7$	0.00427	0.91996	0.96597	−0.04601
4	SVM vs. RF, 10-fold CV	$d = 24$	0.00253	0.95612	0.95896	−0.00284
		Average	0.00279	0.95249	0.95646	−0.00397

Taken together, the four studies suggest that for a p-value between 0.00160 and 0.00427, the empirical replication probability $\mathcal{F}$ is around 0.95 and close to the corresponding estimated replication probability, $\widehat{\mathcal{P}}$. We assume that this replication probability is of particular interest in practice, as it provides a relatively strong reassurance that the results will replicate. These results suggest that a p-value should preferably be around 0.003 to bestow a relatively high trust in the replicability of the experiments. This value is close to the significance threshold of 0.005 that was previously proposed for claims of new discoveries (Benjamin et al., 2018). Both the theoretical and empirical analyses have shown that p-values just below the common threshold of 0.05 are insufficient to raise high hopes in the replicability of the experiments. Although we focused on the replicability of classification benchmark experiments, we believe that this finding is generalizable to other applications of NHST and might explain, at least in part, why many studies fail to be replicable (Gundersen et al., 2022).

Unless the p-values are extremely small, the prediction intervals for the replication probability are very wide, indicating that there is a high uncertainty. This is due to the fact that the effect size cannot be estimated with a high precision based on a single experiment. Because of the lack of precision, Miller (2009) advises against reporting the replication

probability. The width of the prediction intervals and the uncertainty are clearly a concern; currently, we do not see how to solve this problem when using only the data from a single study.

A different approach was adopted by Killeen (2005) who proposed an expression for the replication probability, p_{rep}, that does not involve the estimate of the population effect size, δ. The proposed p_{rep} is the average of the replication probability over *all* values of δ, weighted by the likelihood that each value would have given the observed effect size d. This is an interesting idea, but the expression for the replication probability then becomes more complicated. Also, the weighted averaging has been criticized as lacking sufficient theoretical underpinning (Cumming, 2005, 2006) and as being a quasi-power coefficient rather than a proper probability (Maraun and Gabriel, 2010).

Our studies have several limitations. The controlled experiments reflect an ideal scenario, since the only source of random variation is the random sampling in the replications. In practice, we would expect further sources of variation; for example, when a researcher tries to replicate the experiments of another researcher, it is unlikely that the follow-up experiment would be carried out in really exactly the same way as the initial experiment. For example, when classifiers are compared over multiple datasets, different researchers might select the benchmark datasets based on different criteria. Also, seemingly minor implementation issues, such as the seeding of the random number generator, can also influence the results. Many machine learning models, such as neural networks, use randomization processes (Zhuang et al., 2022). Deep learning algorithms use nondeterminism to improve the training efficiency, which means that different training runs lead to different models with different accuracies (Pham et al., 2020).

When we calculate the replication probability, we make the assumption that the initial experiment is a representative example from the pool of all similar experiments. For example, consider again the randomly selected experiment from group 3 of study #1 (Figure 7, highlighted). The mean success rate for this pool of experiments is $\bar{\hat{\theta}} = 0.752$, or about 15 successes in 20 trials, with a binomial p-value of 0.0414. But what if, just by chance, we got lucky in our initial experiment and observed 18 successes? In 90 out of 1000 replications in group 3, we observed 18 successes indeed. Now the observed success rate of $\hat{\theta} = \frac{18}{20} = 0.90$, with a p-value of 0.0004, would drastically overestimate the empirical replication probability. When we calculate the replication probability, we therefore need to assume that were are dealing with a typical, representative experiment from the pool of similar experiments.

In this work, we were primarily interested in the question of estimating the probability of replicating a significant result. A related question is: "What is the probability that the replication experiment is significant, although the initial experiment is not?" It is indeed instructive to consider the replication probability of non-significant results. For example, in study #3, when the oracle revealed 2% of the validation cases to SVM_{o}, the mean t-value from 1000 repetitions was 1.9364, with a p-value of 0.0848. Assuming the common significance threshold of 0.05, this is a non-significant result. But in 1000 replications, exactly 200 experiments turned out to be significant (each with $p < 0.05$). Hence, the probability that the follow-up experiment *is* significant is $\frac{200}{999} \approx 0.20$. So despite the non-significant result from the initial study, there is an about 20% chance that the replication will give a significant result.

How should the findings of the present study be used by the machine learning community? The major insight is that barely significant p-values do not imply high chances of replicability. There are indeed many problems with significance testing; for an overview, see for example (Berrar, 2022). There have even been suggestions of abandoning NHST altogether (Amrhein and Greenland, 2018; Berrar and Dubitzky, 2019); however, given that NHST is so widely used, this seems unrealistic. If replication is of primary interest—and we argue that it should be—then authors of scientific articles reporting new discoveries should therefore aim for much smaller p-values, i.e., around 0.003, not just below 0.05. In particle physics, claims of important novel discoveries commonly need to meet the so-called 5σ-criterion, which is equivalent to a p-value of $3 \cdot 10^{-7}$ (Lyons, 2013). In their guidelines for authors, journal editors and conference organizers should place greater emphasis on replication and not on statistical significance.

5. Conclusions

The replication probability, together with its prediction interval, can help with the interpretation of the p-value, as it directly answers a central question:"Now that we have a significant finding, what are the chances that it will replicate?" However, the replication probability should not replace the p-value, but instead complement it, as this probability can be estimated with reasonably high precision only for extremely small p-values. We showed both theoretically and empirically that a p-value just below the common significance threshold of 0.05 is clearly insufficient to warrant a high confidence that a significant result will replicate. For a reasonably high chance of replicability of around 95%, the significance threshold should be around 0.003. Unless the p-value is extremely small, there is a high uncertainty about the replication probability, as reflected by the width of the prediction intervals.

Code and Data Availability

The R code is available at the project website at `https://osf.io/7vqfn/`.

Acknowledgments

We thank the reviewers very much for their detailed and constructive comments on this manuscript. Funding in support of this work: TASMAL award from the Open Societal Challenges, The Open University, UK.

Appendix A

Algorithm 2: Comparison of NB and SVM over multiple datasets

$N \leftarrow 20$ // Initialize number of benchmark datasets. 20 in study #1, 44 in #2.
$A \leftarrow \{190, 240, 290..., 1290\}$ // Initialize start sizes of datasets (shown for study #1).
$k_{\max} \leftarrow 1000$ // Initialize maximum number of replications.
for a in A **do**
  $s_{ba} \leftarrow 0$ // number of significant results under binomial model in experiment with a.
  $s_{wa} \leftarrow 0$ // number of significant results under Wilcoxon model in experiment with a.
  for k from 1 to $k_{\max}$ **do**
    for i from 1 to N **do**
      $\mathbf{T}_i \leftarrow [\,],\ \mathbf{W}_i \leftarrow [\,]$ // Initialize training and test set.
      $\mathbf{t}^+ \leftarrow [\,],\ \mathbf{w}^+ \leftarrow [\,]$ // Row vectors of positive cases.
      $\mathbf{t}^- \leftarrow [\,],\ \mathbf{w}^- \leftarrow [\,]$ // Row vectors of negative cases.
      $n_i \leftarrow a + 10 \cdot i$ // Number of cases in training and test set.
      $d_i \leftarrow i + 10$ // Number of features or dimension of $\mathbf{t}$ and $\mathbf{w}$.
      $\mu_i \leftarrow 0.3 + \frac{1}{200}(N + 1 - i)$
      $\sigma_i \leftarrow 1 + \frac{1}{50}(N + 1 - i)$
      for j from 1 to $\frac{1}{2}n_i$ // Half of the cases are positive.
      do
        sample each of the d_i elements in $\mathbf{t}_j^+$ from $\mathcal{N}(\mu_i, \sigma_i^2)$
        sample each the d_i elements in $\mathbf{w}_j^+$ from $\mathcal{N}(\mu_i, \sigma_i^2)$
        append $\mathbf{t}_j^+$ to $\mathbf{T}_i$
        append $\mathbf{w}_j^+$ to $\mathbf{W}_i$
      for j from $(\frac{1}{2}n_i + 1)$ to n_i // Half of the cases are negative.
      do
        sample each of the d_i elements in $\mathbf{t}_j^-$ from $\mathcal{N}(0, 1)$
        sample each of the d_i elements in $\mathbf{w}_j^-$ from $\mathcal{N}(0, 1)$
        append $\mathbf{t}_j^-$ to $\mathbf{T}_i$
        append $\mathbf{w}_j^-$ to $\mathbf{W}_i$
      train NB and SVM on $\mathbf{T}_i$
      test NB and SVM on $\mathbf{W}_i$
    compare performance of NB and SVM over N test sets
    calculate success rate $\hat{\theta}_{ak}$
    calculate binomial p-value $p_{b,ak}$
    if $p_{b,ak} < 0.05$ *and* $\hat{\theta}_{ak} \geq \frac{15}{20}$ **then**
      $s_{ba} \leftarrow s_{ba} + 1$
    calculate $Z_{w,ak}$ // Equation 10
    calculate Wilcoxon p-value $p_{w,ak}$
    if $p_{w,ak} < 0.05$ *and* $sign\{Z_{w,ak}\} = +$ **then**
      $s_{wa} \leftarrow s_{wa} + 1$
  $\bar{\hat{\theta}}_a \leftarrow \frac{1}{k_{\max}} \sum \hat{\theta}_{ak}$ // mean success rate for experiment with a.
  $\mathcal{F}_{ba} \leftarrow (s_{ba} - 1)\ /\ (k_{\max} - 1)$ // empirical replication probability under binomial model for experiment with a.
  $\bar{Z}_{wa} \leftarrow \frac{1}{k_{\max}} \sum Z_{w,ak}$ // mean Z_w for experiment with a.
  $\mathcal{F}_{wa} \leftarrow (s_{wa} - 1)\ /\ (k_{\max} - 1)$ // empirical replication probability under Wilcoxon model for experiment with a.
  calculate the estimated replication probabilities (and their limits) for the experiment with a according to Equation 5, Equation 8, and Equation 11.

Appendix B

Algorithm 3: Comparison of SVM and SVM_o in cross-validation.

$Q \leftarrow \{1, 2, 3..., 20\}$ // percentages of revealed class labels.
$k_{\max} \leftarrow 1000$ // maximum number of replications.
$n \leftarrow 1000$ // number of cases in learning set.
$d \leftarrow 20$ // number of features in learning set.
$f_{\max} \leftarrow 10$ // maximum number of cross-validation folds.
for q in Q **do**
  $s_q \leftarrow 0$ // number of significant results in experiment with q.
  for k from 1 to $k_{\max}$ **do**
    $\mathbf{L}_k \leftarrow [\,]$ // learning set.
    $\mathbf{l}^+ \leftarrow [\,]$ // row vector of positive cases.
    $\mathbf{l}^- \leftarrow [\,]$ // row vectors of negative cases.
    for j from 1 to $\frac{1}{2}n$ // half of the cases are positive.
    do
      sample each of the d_i elements in $\mathbf{l}_j^+$ from $\mathcal{N}(0.3, 1)$
      append $\mathbf{l}_j^+$ to $\mathbf{L}_k$
    for j from $(\frac{1}{2}n + 1)$ to n // Half of the cases are negative.
    do
      sample each of the d_i elements in $\mathbf{l}_j^-$ from $\mathcal{N}(0, 1)$
      append $\mathbf{l}_j^-$ to $\mathbf{L}_k$
    for i from 1 to $f_{\max}$ **do**
      stratified random sampling to generate $f_{\max}$ validation sets $\mathbf{V}_1$ to $\mathbf{V}_{f_{\max}}$
      generate training set $\mathbf{T}_i \leftarrow \mathbf{L}_k \setminus \mathbf{V}_i$
      train SVM on $\mathbf{T}_i$
      $\text{SVM}_\text{o} \leftarrow$ SVM // SVM_o is identical to SVM.
      apply SVM to $\mathbf{V}_i$ // predict the validation cases with SVM.
      calculate accuracy of SVM of the i^{th} validation set, $acc(\text{SVM}, \mathbf{V}_i)$
      apply SVM_o to $\mathbf{V}_i$ // predict the validation cases with SVM_o.
      randomly select q% of cases from $\mathbf{V}_i$ and replace the predicted class labels from SVM_o by the real class labels. // access oracle.
      calculate accuracy of SVM_o of the i^{th} validation set, $acc(\text{SVM}_\text{o}, \mathbf{V}_i)$
      $\delta_i \leftarrow acc(\text{SVM}_\text{o}, \mathbf{V}_i) - acc(\text{SVM}, \mathbf{V}_i)$ // difference in accuracy.
    $\bar{\delta}_{qk} \leftarrow \frac{1}{f_{\max}} \sum \delta_i$ // mean difference in accuracy.
    $T_{qk} \leftarrow \bar{\delta}_{qk} / \sqrt{(1/f + |\mathbf{V}|/|\mathbf{T}|)\ var(\delta)}$ // t-value.
    derive p-value p_{qk} from T_{qk} with $f - 1$ degrees of freedom.
    if $p_{qk} < 0.05$ *and* $sign\{T_{qk}\} = +$ **then**
      $s_q \leftarrow s_q + 1$
  $\mathcal{F}_q \leftarrow (s_q - 1)\ /\ (k_{\max} - 1)$ // empirical repl. prob. for experiment with q.
  $\bar{T}_q \leftarrow \frac{1}{k_{\max}} \sum T_{qk}$ // mean t-value for experiment with q.
  calculate the point estimate (and limits) of the replication probability $\widehat{\mathcal{P}}_q$ for experiment with q according to Equation 13.

Appendix C

Algorithm 4: Comparison of SVM and RF in cross-validation.

$D \leftarrow \{4, 5, ..., 30\}$ // number of predictive features in learning set.
$k_{\max} \leftarrow 1000$ // maximum number of replications.
$n \leftarrow 1000$ // number of cases in learning set.
$f_{\max} \leftarrow 10$ // maximum number of cross-validation folds.
for d in D **do**
  $s_d \leftarrow 0$ // number of significant results in experiment with d.
  for k from 1 to $k_{\max}$ **do**
    $\mathbf{L}_k \leftarrow [\,]$ // learning set.
    $\mathbf{l}^+ \leftarrow [\,]$ // row vector of positive cases.
    $\mathbf{l}^- \leftarrow [\,]$ // row vectors of negative cases.
    for j from 1 to $\frac{1}{2}n$ // half of the cases are positive.
    do
      sample each of the d_i elements in $\mathbf{l}_j^+$ from $\mathcal{N}(0.3, 1)$
      append $\mathbf{l}_j^+$ to $\mathbf{L}_k$
    for j from $(\frac{1}{2}n + 1)$ to n // Half of the cases are negative.
    do
      sample each of the d_i elements in $\mathbf{l}_j^-$ from $\mathcal{N}(0, 1)$
      append $\mathbf{l}_j^-$ to $\mathbf{L}_k$
    for i from 1 to $f_{\max}$ **do**
      stratified random sampling to generate f validation sets $\mathbf{V}_1$ to $\mathbf{V}_f$
      generate training set $\mathbf{T}_i = \mathbf{L}_k \setminus \mathbf{V}_i$
      train SVM on $\mathbf{T}_i$
      train RF on $\mathbf{T}_i$
      apply SVM to $\mathbf{V}_i$ // predict the validation cases with SVM.
      apply RF to $\mathbf{V}_i$ // predict the validation cases with RF.
      calculate accuracy of SVM for the i^{th} validation set, $acc(\text{SVM}, \mathbf{V}_i)$
      calculate accuracy of RF for the i^{th} validation set, $acc(\text{RF}, \mathbf{V}_i)$
      $\delta_i = acc(\text{SVM}, \mathbf{V}_i) - acc(\text{RF}, \mathbf{V}_i)$ // difference in accuracy.
    $\bar{\delta}_{dk} = \frac{1}{f_{\max}} \sum \delta_i$ // mean difference in accuracy.
    $T_{dk} \leftarrow \bar{\delta}_{dk} / \sqrt{(1/f + |\mathbf{V}|/|\mathbf{T}|)\ var(\delta)}$ // variance-corrected t-value.
    derive p-value p_{dk} from T_{dk} with $f - 1$ degrees of freedom.
    if $p_{dk} < 0.05$ *and* $sign\{T_{dk}\} = +$ **then**
      $s_d \leftarrow s_d + 1$
  $\mathcal{F}_d \leftarrow (s_d - 1)\ /\ (k_{\max} - 1)$ // empirical rep. prob. for experiment with d.
  $\bar{T}_d \leftarrow \frac{1}{k_{\max}} \sum T_{dk}$ // mean t-value for experiment with d.
  calculate the point estimate (and limits) of the replication probability $\widehat{\mathcal{P}}_d$ for experiment with d according to Equation 13.

Appendix D

Table 2: Number of cases per dataset in study #2, comparing NB and SVM on 44 benchmark datasets (rows). Each column represents one experimental group. For example, the third dataset in the second group contains 130 cases.

	1	2	3	4	5	6	7	8	9	10
1	80	110	150	200	210	250	270	370	470	570
2	90	120	160	210	220	260	280	380	480	580
3	100	130	170	220	230	270	290	390	490	590
4	110	140	180	230	240	280	300	400	500	600
5	120	150	190	240	250	290	310	410	510	610
6	130	160	200	250	260	300	320	420	520	620
7	140	170	210	260	270	310	330	430	530	630
8	150	180	220	270	280	320	340	440	540	640
9	160	190	230	280	290	330	350	450	550	650
10	170	200	240	290	300	340	360	460	560	660
11	180	210	250	300	310	350	370	470	570	670
12	190	220	260	310	320	360	380	480	580	680
13	200	230	270	320	330	370	390	490	590	690
14	210	240	280	330	340	380	400	500	600	700
15	220	250	290	340	350	390	410	510	610	710
16	230	260	300	350	360	400	420	520	620	720
17	240	270	310	360	370	410	430	530	630	730
18	250	280	320	370	380	420	440	540	640	740
19	260	290	330	380	390	430	450	550	650	750
20	270	300	340	390	400	440	460	560	660	760
21	280	310	350	400	410	450	470	570	670	770
22	290	320	360	410	420	460	480	580	680	780
23	300	330	370	420	430	470	490	590	690	790
24	310	340	380	430	440	480	500	600	700	800
25	320	350	390	440	450	490	510	610	710	810
26	330	360	400	450	460	500	520	620	720	820
27	340	370	410	460	470	510	530	630	730	830
28	350	380	420	470	480	520	540	640	740	840
29	360	390	430	480	490	530	550	650	750	850
30	370	400	440	490	500	540	560	660	760	860
31	380	410	450	500	510	550	570	670	770	870
32	390	420	460	510	520	560	580	680	780	880
33	400	430	470	520	530	570	590	690	790	890
34	410	440	480	530	540	580	600	700	800	900
35	420	450	490	540	550	590	610	710	810	910
36	430	460	500	550	560	600	620	720	820	920
37	440	470	510	560	570	610	630	730	830	930
38	450	480	520	570	580	620	640	740	840	940
39	460	490	530	580	590	630	650	750	850	950
40	470	500	540	590	600	640	660	760	860	960
41	480	510	550	600	610	650	670	770	870	970
42	490	520	560	610	620	660	680	780	880	980
43	500	530	570	620	630	670	690	790	890	990
44	510	540	580	630	640	680	700	800	900	1000

References

V. Amrhein and S. Greenland. Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, 2(4):4, 2018.

Association for Computing Machinery. Artifact review and badging version 1.1 – August 24, 2020, 2020. URL `https://www.acm.org/publications/policies/artifact-review-and-badging-current`. accessed 4 January 2024.

M. Baker and D. Penny. Is there a reproducibility crisis? *Nature*, 533:452–454, 2016.

M.J. Bayarri and J.O. Berger. *P* values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142, 2000.

A. Benavoli, G. Corani, and F. Mangili. Should we really use post-hoc tests based on mean-ranks? *Journal of Machine Learning Research*, 17(5):1–10, 2016.

A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017.

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. 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. Munafó, 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, 2018.

J.O. Berger and M. Delampady. Testing precise hypotheses. *Statistical Science*, 2(3):317–352, 1987.

D. Berrar. *Confidence curves*: an alternative to null hypothesis significance testing for the comparison of classifiers. *Machine Learning*, 106(6):911–949, 2017.

D. Berrar. Using *p*-values for the comparison of classifiers: Pitfalls and alternatives. *Data Mining & Knowledge Discovery*, 36:1102–1139, 2022.

D. Berrar and W. Dubitzky. Should significance testing be abandoned in machine learning? *International Journal of Data Science and Analytics*, 7(4):247–257, 2019.

X. Bouthillier, C. Laurent, and P. Vincent. Unreproducible research is reproducible. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 725–734. PMLR, 2019.

R.P. Carver. The case against statistical significance testing. *Harvard Educational Review*, 48(3):378–399, 1978.

C.J. Clopper and E.S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.

A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8):70–79, 2020.

D. Colquhoun. The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, 4:171085, 2017.

G. Cumming. Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, 16(12):1002–1004, 2005.

G. Cumming. Understanding replication: confidence intervals, p values, and what's likely to happen next. In *7th International Conference on Teaching Statistics,*, pages 1–6, 2006.

G. Cumming. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3(4):286–300, 2008.

G. Cumming. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge, Taylor & Francis Group, New York/London, 2012.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

C. Drummond. Replicability is not reproducibility: Nor is it good science. *Proceedings of Evaluation Methods for Machine Learning Workshop at the 26th ICML, Montreal, Canada*, pages 1–6, 2009.

C. Drummond and N. Japkowicz. Warning: Statistical benchmarking is addictive. Kicking the habit in machine learning. *Journal of Experimental and Theoretical Artificial Intelligence*, 2:67–80, 2010.

D.A. Eisner. Reproducibility of science: Fraud, impact factors and carelessness. *Journal of Molecular and Cellular Cardiology*, 114:364–368, 2018.

R.C. Fraley and M.J. Marks. The null hypothesis significance testing debate and its implications for personality research. In R.W. Robins, R.C. Fraley, and R.F. Krueger, editors, *Handbook of Research Methods in Personality Psychology*, pages 149–169. Guilford, New York, 2007.

E.W. Gibson. The role of p-values in judging the strenght of evidence and realistic replication expectations. *Statistics in Biopharmaceutical Research*, 13(1):6–18, 2020.

G. Gigerenzer. Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218, 2018.

S. Goodman. A comment on replication, p-values and evidence. *Statistics in Medicine*, 11: 875–879, 1992.

S. Goodman. Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999.

S. Goodman. A dirty dozen: Twelve P-value misconceptions. *Seminars in Hematology*, 45 (3):135–140, 2008.

S. Greenland, S.J. Senn, K.J. Rothman, J.B. Carlin, C. Poole, S.N. Goodman, and D.G. Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350, 2016.

O.E. Gundersen. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A*, 379(2197):20200210, 2021.

O.E. Gundersen, K. Coakley, and C. Kirkpatrick. Sources of irreproducibility in machine learning: A review. pages 1–8, 2022. URL `https://arxiv.org/abs/2204.07610v1`. accessed 30 January 2024.

R.L. Hagen. In praise of the null hypothesis significance test. *American Psychologist*, 52 (1):15–23, 1997.

D. Harrington, R.B. D'Agostino, C. Gatsonis, J.W. Hogan, D. J. Hunter, S.-L.T. Normand, J.M. Drazen, and M.B. Hamel. New guidelines for statistical reporting in the journal. *New England Journal of Medicine*, 381(3):285–286, 2019.

P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pages 327–3214. AAAI Press, 2018.

R. Hubbard and M.J. Bayarri. P values are not error probabilities. *Technical Report University of Valencia; Accessed 8 February 2021*, 2003. URL `http://www.uv.es/sestio/TechRep/tr14-03.pdf`. accessed 30 January 2024.

M. Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018.

J.P.A. Ioannidis. The proposal to lower p value thresholds to .005. *JAMA*, 319(14):1429–1430, 2018.

P.R. Killeen. An alternative to null-hypothesis significance tests. *Psychological Science*, 16 (5):345–352, 2005.

R.A. Klein, M. Vianello, and F. Hasselman, et al. Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

J.K. Kruschke. Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018.

J. Lai, F. Fidler, and G. Cumming. Subjective p intervals—researchers underestimate the variability of p values over replication. *Methodology*, 8(2):51–62, 2012.

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

P. Langley. Machine learning as an experimental science. *Machine Learning*, 3:5–8, 1988.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

M.A. Lones. How to avoid machine learning pitfalls: a guide for academic researchers, 2024. URL `https://arxiv.org/abs/2108.02497`. accessed 30 January 2024.

L. Lyons. Discovering the significance of 5σ. *arXiv e-prints*, page arXiv:1310.1284, 2013. doi: 10.48550/arXiv.1310.1284.

M. Maraun and S. Gabriel. Killeen's (2005) p_{rep} coefficient: Logical and mathematical problems. *Psychological Methods*, 15(2):182–191, 2010.

B.B. McShane, D. Gal, A. Gelman, C. Robert, and J.L. Tackett. Abandon statistical significance. *The American Statistician*, 73(sup1: Statistical Inference in the 21st Century: A World Beyond $p < 0.05$):235–245, 2019.

G. Melis, C. Dyer, and P. Blunsom. On the state of the art of evaluation in neural language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada*, pages 285–286. OpenReview.net, 2018.

D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F¿ Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2022. URL `https://CRAN.R-project.org/package=e1071`. R package version 1.7-11.

J. Miller. What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, 15(4):617–640, 2009.

S.A. Mulaik, N.S. Raju, and Harshman R.A. There is a time and a place for significance testing. In L.L. Harlow, S.A. Mulaik, and J.H. Steiger, editors, *What if there were no significance tests?* Routledge Classic Editions, 2016.

C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52: 239–281, 2003.

B.A. Nosek, C.R. Ebersole, A.C. DeHaven, and D.T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences of the USA*, 115(11):2600–2606, 2018.

R. Nuzzo. Statistical errors. *Nature*, 506:150–152, 2014.

M.L. Oakes. *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. Wiley, New York, 1986.

H.V. Pham, S. Qian, J. Wang, T. Lutellier, J. Rosenthal, L. Tan, Y. Yu, and N. Nagappan. Problems and opportunities in training deep learning software systems: An analysis of variance. In *35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 771–783, 2020.

J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(1):1–20, 2021.

H.E. Plesser. Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11:article 76, 2018.

H. Sackrowitz and E. Samuel-Cahn. p values as random variables–expected p values. *The American Statistician*, 53(4):326–331, 1999.

F.L. Schmidt and J.E. Hunter. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik, and J.H. Steiger, editors, *What if there were no significance tests?*, pages 35–60. Routledge, 2016.

D. Sculley, J. Snoek, A. Wiltschko, and A. Rahimi. Winner's curse? On pace, progress, and empirical rigor. Proceedings of 6th International Conference on Learning Representations, Workshop Track, pages 1–4, 2018.

H. Semmelrock, S. Kopeinik, D. Theiler, T. Ross-Hellauer, and D. Kowald. Reproducibility in machine learning-driven research. *arXiv e-prints*, page arXiv:2307.10320, 2023. doi: 10.48550/arXiv.2307.10320.

S. Senn. A comment on replication, p-values and evidence. *Statistics in Medicine*, 21: 2437–2444, 2002.

D.J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. 4th edition, Chapman and Hall, CRC, 2007.

S. Sonnenburg, ML. Braun, CS. Ong, S. Bengio, L. Bottou, G. Holmes, Y. LeCun, K-R. Müller, F. Pereira, CE. Rasmussen, G. Rätsch, B. Schölkopf, A. Smola, P. Vincent, J. Weston, and RC. Williamson. The need for open source software in machine learning. *Journal of Machine Learning Research*, 8:2443–2466, 2007.

A. Stang, C. Poole, and O. Kuss. The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology*, 25:225–230, 2010.

K. Stapor, P. Ksieniewicz, S. García, and M. Woźniak. How to design the fair experimental classifier evaluation. *Applied Soft Computing*, 104:107219, 2021.

J. Vanschoren, J.N. van Rijn, B. Bischl, and L. Torgo. OpenML networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

J. Wainer. A Bayesian Bradley-Terry model to compare multiple ML algorithms on multiple data sets. *Journal of Machine Learning Research*, 24:1–34, 2023.

R.L. Wasserstein, A.L. Schirm, and N.A. Lazar. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1: Statistical Inference in the 21st Century: A World Beyond $p < 0.05$):1–19, 2019.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

D. Zhuang, X. Zhang, S. Song, and S. Hooker. Randomness in neural network training: Characterizing the impact of tooling. In *Proceedings of the 5th MLSys Conference, Santa Clara, CA, USA*, pages 316–336, 2022.