Commentary

Why Should P-Values Be Abandoned in Scientific Research?

Hening Huang[1]

[1]. Independent researcher

It is widely acknowledged in the scientific community that the *p*-values generated by null hypothesis significance testing (NHST) procedures can be easily misunderstood, misinterpreted, and/or misused. This paper provides an in-depth analysis of the two-sample one-tailed *z*-test and the *p*-value it generates. We explore why *p*-values should be abandoned in scientific research.

Corresponding author: Hening Huang, heninghuang1@gmail.com

1. Introduction

P-values generated by null hypothesis significance testing (NHST) procedures have played an important role in scientific research for a long time. In recent decades, however, *p*-values and the related NHST paradigm have faced increasing criticism because many published scientific findings cannot be replicated by other researchers, leading to a concern known as the "reproducibility crisis," where the reliability of scientific findings is questioned. Halsey et al.[1] argued that a major cause of the lack of repeatability is the wide sample-to-sample variability in *p*-values. They explained “why *P* is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic.” To address the "reproducibility crisis," many authors strongly oppose the use of *p*-values and NHST and suggest abandoning them (e.g., Amrhein et al.[2]; McShane et al.[3]; Halsey[4]; Wasserstein & Lazar[5]; Wasserstein et al.[6]) and using estimation statistics (e.g., Claridge-Chang & Assam[7]; Berner & Amrhein[8]; Elkins et al.[9]; Huang[10]). Huang[11] argues that the *p*-value is not an appropriate probabilistic measure in scientific decision-making because it can be easily hacked through *N*-chasing; unless *p*-values and NHST are abandoned, the *p*-hacking problem caused by *N*-chasing cannot be solved. Halsey’s[4] states, “The reign of the *p*-value is over.” Recently, Trafimow et

doi.org/10.32388/3SYVEJ

al.[12] proposed using the gain-probability (G-P) analysis to replace NHST and *p*-values. The G-P analysis is essentially the same as the exceedance probability analysis[11]. However, some authors defend *p*-values and NHST (e.g., Lu & Belitskaya-Levy[13]; Verhulst[14]; Benjamini et al.[15]; Hand[16]; Lohse[17]; Chén et al.[18]). For example, Chén et al.[18] argue that *p*-values and NHST form a useful probabilistic decision-making system and that *p*-values will continue to play an important role in scientific research.

While many scientists and statisticians are still debating whether to completely abandon NHST and *p*-values or to continue using them, both sides of the debate generally acknowledge that *p*-values are often and easily misunderstood, misinterpreted, and/or misused. Common misconceptions about *p*-values include that the *p*-value measures the probability that the research hypothesis is true and that the *p*-value measures the probability that observed data are due to chance[18]. Goodman[19] discussed twelve *p*-value misconceptions raised from a two-group randomized experiment. Moreover, misinterpretation of *p*-values and NHST results even persists among people with substantial statistical education and working in statistics[20]. However, Goodman[19] stated,

> It is not the fault of researchers that the *P* value is difficult to interpret correctly. The man who introduced it as a formal research tool, the statistician and geneticist R.A. Fisher, could not explain exactly its inferential meaning. He proposed a rather informal system that could be *used*, but he never could describe straightforwardly what it meant from an inferential standpoint.

Mathematically, the *p*-value is defined as the tail probability calculated using a test statistic[18]. But the question is, what does the *p*-value (the tail probability) really mean in practical applications? This is a key question about the validity of using *p*-values in scientific research. We argue that the correct answer to this question will be common ground in the *p*-value debate. If we can figure out what *p*-values really mean, we can decide whether we should continue to use *p*-values or whether we should abandon them. In this way, the debate about *p*-values will end.

It is well known that the *p*-value generated by a NHST procedure is a random variable because it depends on the samples randomly drawn from the population of interest. However, it seems less known that there is a true (or theoretical) *p*-value and a theoretical distribution of the *p*-value (the usual *p*-value). Lazzeroni et al.[21] defined the "true population *p*-value" or π value as "the value of *p* when parameter estimates equal their unknown population values." They also introduced p-value confidence intervals for the true *p*-value. However, the Lazzeroni et al.[21] paper did not include mathematical details about the "true population *p*-value" and p-value confidence intervals.

This paper provides an in-depth analysis of the two-sample one-tailed *z*-test and the *p*-values it generates. In the following sections, Section 2 shows the true *p*-value of the two-sample one-tailed *z*-test and its meaning. Section 3 shows that the *p*-value (i.e., the usual *p*-value) is an estimate of the true *p*-value. Section 4 presents the theoretical distribution of the usual *p*-value. Section 5 presents a numerical example. Sections 6 and 7 present discussion and conclusion, respectively.

2. The true *p*-value of the two-sample one-tailed *z*-test and its meaning

Consider a controlled experiment with two groups of individuals: the treatment group (denoted by A) and the control group (denoted by B), which gives two independent samples (datasets for a measurable quantity *X*): $\{x_{A,1}, x_{A,2}, ..., x_{A,n}\}$ and $\{x_{B,1}, x_{B,2}, ..., x_{B,n}\}$, where *n* is the sample size. We assume that the two datasets are randomly sampled from two independent normal distributions $X_A \sim N(\mu_A, \sigma_A)$ and $X_B \sim N(\mu_B, \sigma_B)$, respectively. For simplicity and without loss of generality, we further assume that $\sigma_A = \sigma_B = \sigma$ and σ is known. Let $\overline{x}_A$ and $\overline{x}_B$ denote the calculated sample means from the two datasets, respectively. The observed (treatment) effect size is $\overline{x}_A - \overline{x}_B$, the difference of the two sample means.

The *usual* *z*-score for the two-sample equal-variance *z*-test is

$$z = \frac{\overline{x}_A - \overline{x}_B}{\sigma\sqrt{2/n}} = \sqrt{\frac{n}{2}}d, \tag{1}$$

where $d = \frac{\overline{x}_A - \overline{x}_B}{\sigma}$ is the standardized *sample* effect size, often referred to as Cohen's *d*.

Note that $\overline{x}_A$ is an unbiased estimate of the population mean μ_A and $\overline{x}_B$ is an unbiased estimate of the population mean μ_B. When μ_A and μ_B are known, we can write the *true* *z*-score as

$$z_{true} = \frac{\mu_A - \mu_B}{\sigma\sqrt{2/n}} = \sqrt{\frac{n}{2}}d_{true}, \tag{2}$$

where $d_{true} = \frac{\mu_A - \mu_B}{\sigma}$ is the standardized *population* effect size, or *true* effect size.

Assuming that $\mu_A > \mu_B$, $z_{true} > 0$. We can calculate the *true* *p*-value for the two-sample one-tailed *z*-test as

$$p_{true} = Pr\left(Z < -z_{true}\right) = \Phi\left(-z_{true}\right) = \Phi\left(-\sqrt{\frac{n}{2}}d_{true}\right), \tag{3}$$

where $\Phi(.)$ is the cumulative distribution function (CDF) of the standard normal distribution $Z \sim N(0,1)$, and Z is the standardized difference between the two sample means, which can be written as

$$Z = \frac{\left(\overline{X}_A - \overline{X}_B\right) - (\mu_A - \mu_B)}{\sigma\sqrt{2/n}} \sim N(0,1), \tag{4}$$

where $\overline{X}_A$ and $\overline{X}_B$ are the sample means (random variables) that are normally distributed as $\overline{X}_A \sim N(\mu_A, \frac{\sigma}{\sqrt{n}})$ and $\overline{X}_2 \sim N(\mu_B, \frac{\sigma}{\sqrt{n}})$, respectively. Note that $N(\mu_A, \frac{\sigma}{\sqrt{n}})$ is the theoretical sampling distribution of the sample mean $\overline{X}_A$ and $N(\mu_B, \frac{\sigma}{\sqrt{n}})$ is the theoretical sampling distribution of the sample mean $\overline{X}_B$.

Substituting the expressions for Z and z_{true} into Eq. (3), Eq. (3) can be rewritten as

$$p_{true} = Pr\left(\left[Z = \frac{\left(\overline{X}_A - \overline{X}_B\right) - (\mu_A - \mu_B)}{\sigma\sqrt{2/n}}\right] < \left[-z_{true} = -\frac{\mu_A - \mu_B}{\sigma\sqrt{2/n}}\right]\right), \tag{5}$$

which can be rewritten as

$$p_{true} = Pr\left(\overline{X}_A - \overline{X}_B < 0\right) = \Pr\left(\overline{X}_A < \overline{X}_B\right). \tag{6}$$

Therefore, the true p-value p_{true} is the *theoretical* probability that the sample mean $\overline{X}_A$ is smaller than the sample mean $\overline{X}_B$. It is important to note that the true p-value is a *deterministic quantity,* because it is calculated from the *theoretical* sampling distributions of $\overline{X}_A$ and $\overline{X}_B$ and is independent of sample data. As long as the population parameters of the parent (population) distributions of X_A and X_B are known, the true p-value can be calculated using Eq. (3) or Eq. (6) for any sample size without using any data.

3. The usual *p*-value is an estimate of the true *p*-value

Now consider the case where the population means μ_A and μ_B are unknown. In this case, the theoretical sampling distributions of the sample means are not available. But we can obtain the estimated sampling distributions of the sample means $\overline{X}_A^{'} \sim N(\overline{x}_A, \frac{\sigma}{\sqrt{n}})$ and $\overline{X}_B^{'} \sim N(\overline{x}_B, \frac{\sigma}{\sqrt{n}})$ using the datasets $\{x_{A,1}, x_{A,2}, \ldots, x_{A,n}\}$ and $\{x_{B,1}, x_{B,2}, \ldots, x_{B,n}\}$. Similar to Z, we can write the statistic $\widehat{Z}$ as,

$$\widehat{Z} = \frac{\left(\overline{X}_A^{'} - \overline{X}_B^{'}\right) - (\overline{x}_A - \overline{x}_B)}{\sigma\sqrt{2/n}} \sim N(0,1). \tag{7}$$

Assuming that $\overline{x}_A > \overline{x}_B$, $z > 0$. The *p* value (i.e., the usual *p*-value) of the two-sample one-tailed *z*-test can be calculated as

$$p = Pr\left(\widehat{Z} < -z\right) = \Phi(-z) = \Phi\left(-\sqrt{\frac{n}{2}}d\right). \tag{8}$$

Substituting Eq. (7) and Eq. (1) into Eq. (8), Eq. (8) can be rewritten as

$$p = Pr\left(\left[\widehat{Z} = \frac{\left(\overline{X}_A^{'} - \overline{X}_B^{'}\right) - (\overline{x}_A - \overline{x}_B)}{\sigma\sqrt{2/n}}\right] < \left[-z = -\frac{\overline{x}_A - \overline{x}_B}{\sigma\sqrt{2/n}}\right]\right), \tag{9}$$

which can be rewritten as

$$p = Pr\left(\overline{X}_A^{'} - \overline{X}_B^{'} < 0\right) = \Pr\left(\overline{X}_A^{'} < \overline{X}_B^{'}\right) \tag{10}$$

Therefore, the usual *p*-value is the *estimated* probability that the sample mean $\overline{X}_A^{'}$ is smaller than the sample mean $\overline{X}_B^{'}$. In other words, the usual *p*-value is an *estimate* of the *true* *p*-value p_{true}; it is a random variable that can be described by a probability distribution.

4. The theoretical distribution of the usual *p*-value

As shown in Eq. (8), the usual *p*-value generated by the two-sample one-tailed *z*-test is a function of the *z*-score (Eq. (1)), which is a random variable following the standard normal distribution. Thus, the distribution of the usual *p*-value must be related to the distribution of the *z*-score. For simplicity, we will use *p*-value as a shorthand for the usual *p*-value hereafter. The probability density function (PDF) of the *p*-value can be determined by

$$f(p) = \frac{dz^{'}}{dp}g(z') \tag{11}$$

where $f(p)$ is the PDF of the *p*-value, $g(z')$ is the PDF of the random variable $Z^{'}$, which follows the shifted standard normal distribution $Z^{'} \sim N(z_{true}$,1), i.e. $Z^{'} = Z + z_{true}$. The validity of Eq. (11) can be easily verified by Monte Carlo simulations.

Since z_{true} is a constant, $dz^{'} = dz$. Then, Eq. (11) can be rewritten as

$$f(p) = \frac{1}{\left(\frac{dp}{dz}\right)} g(z') \tag{12}$$

Note that the value of *p* corresponding to a value of *z* is the CDF of *Z*, i.e. $p(z) = \Phi(z)$. Thus,

$$\frac{dp}{dz} = \frac{d\Phi(z)}{dz} = g(z) \tag{13}$$

Then, Eq. (11) can be rewritten as

$$f(p) = \frac{g(z')}{g(z)} \tag{14}$$

In the special case where $z_{true} = 0$, i.e. $\mu_A - \mu_B = 0$, Eq. (14) reduces to

$$f(p) = 1. \tag{15}$$

Equation (15) suggests that, when the true effect size is zero (or the null hypothesis is true), the *p*-value is uniformly distributed between 0 and 1, regardless of the sample size involved.

Furthermore, we can use the theoretical distribution of the *p*-value and the relationship between the *z*-score and *p*-value to construct the coverage intervals for the *p*-value. We consider here a coverage probability of 90%. The 90% *z*-score coverage interval is centered on z_{true} and can be written as

$$(z_{true} - 1.64485, z_{true} + 1.64485) \tag{16}$$

It is easy to show that

$$\int_{p_{(z_{true}-1.64485)}}^{p_{(z_{true}+1.64485)}} f(p) = 0.9 \tag{17}$$

where $p_{(z_{true}-1.64485)}$ is the *p*-value corresponding to the *z*-score $(z_{true} - 1.64485)$, which is the lower bound of the 90% *z*-score interval, and $p_{(z_{true}+1.64485)}$ is the *p*-value corresponding to the *z*-score $(z_{true} + 1.64485)$, which is the upper bound of the 90% *z*-score interval. Therefore, the 90% coverage interval for the *p*-value is

$$(p_{(z_{true}-1.64485)}, p_{(z_{true}+1.64485)}) \tag{18}$$

The *p*-value coverage interval will not be centered around the true *p*-value p_{true} because, unless the effect size is zero, the *p*-value distribution is not symmetric about the true *p*-value p_{true}. This can be seen from the example below. Note that the *p*-value coverage interval is a probability interval with fixed bounds; it is not a confidence interval with random bounds.

5. Numerical example

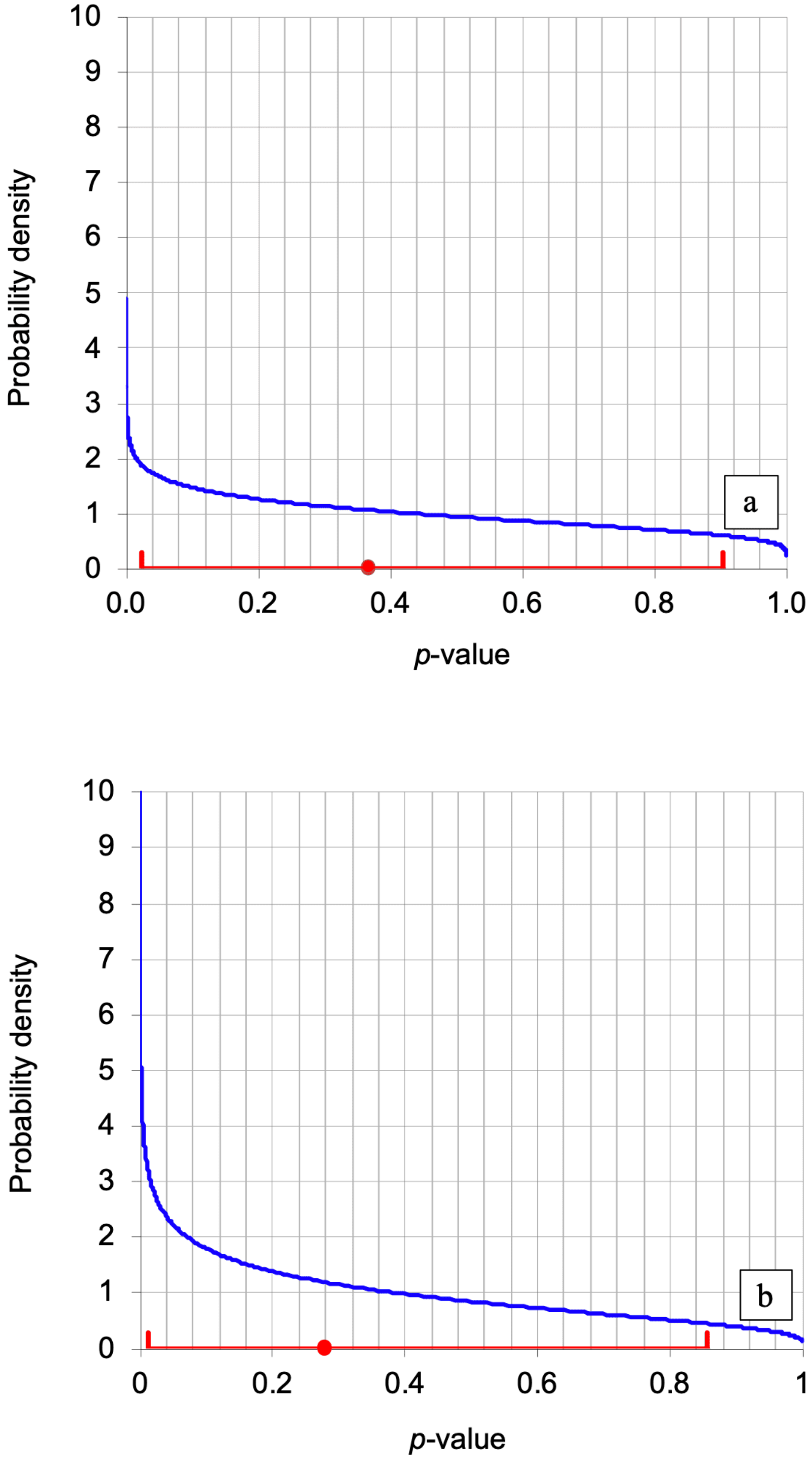
As a numerical example, we consider the two-sample one-tailed z-test on the samples (n=10, 30, 50, and 100) randomly drawn from two normal distributions of X_A and X_B with $\sigma_A = \sigma_B = 1$ and the true effect size $\mu_A - \mu_B = 0.15$. Results for the true p-value, the 90% p-value coverage interval, and the false positive effect rate are shown in Table 1. The false positive rate is defined as the cumulative probability of the p-values smaller than the critical p-value of 0.05. For this example, the true standardized effect size Cohen's d = 0.15, which is considered a "trivial effect" according to Cohen's effect size categories. Therefore, all z-tests should give p-value > 0.05, non-significant results, regardless of sample size.

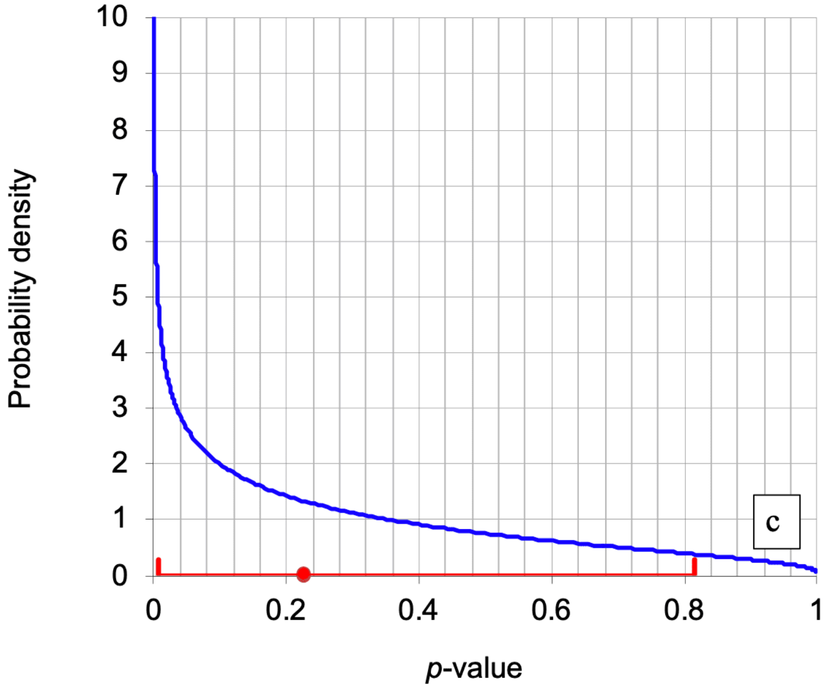
Sample size n	The true p-value	The 90% p-value coverage interval	False positive rate
10	0.369	(0.0238, 0.9048)	9.77%
30	0.281	(0.0130, 0.8563)	14.71%
50	0.227	(0.0083, 0.8146)	18.94%
100	0.144	(0.0034, 0.7205)	28.46%

Table 1. Results for the two-sample one-tailed z-tests on the samples randomly drawn from the two normal distributions of X_A and X_B ($\mu_A - \mu_B = 0.15$ and $\sigma_A = \sigma_B = 1$)

It can be seen from Table 1 that, as the sample size increases, the true p-value and the p-value coverage interval decrease. On the other hand, the false positive rate increases with the increase of the sample size.

Figure 1 shows the theoretical distributions of the p-value generated by the two-sample one-tailed z-tests on the samples (n=10, 30, 50, and 100) randomly drawn from the two normal distributions of X_A and X_B.


10
9
8
7
6
5
4
3
2
1
0
Probability density
0.0
0.2
0.4
0.6
0.8
1.0
p-value
a
10
9
8
7
6
5
4
3
2
1
0
Probability density
0
0.2
0.4
0.6
0.8
1
p-value
b


10
9
8
7
6
5
4
3
2
1
0
Probability density
c
0
0.2
0.4
0.6
0.8
1
p-value

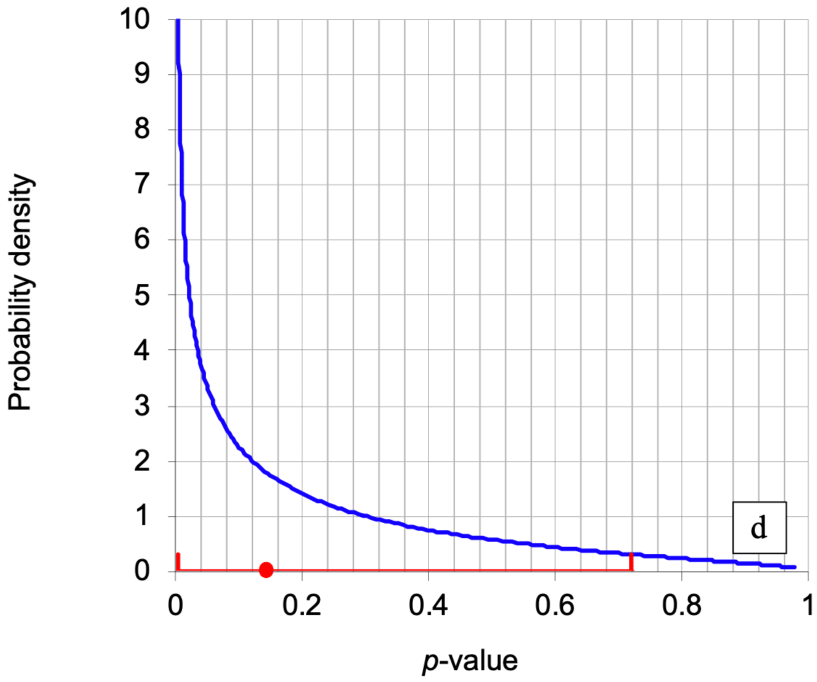


Figure 1. Theoretical distributions of the p-value generated by the two-sample one-tailed z-tests with the true effect size 0.15: (a) n=10, (b) n=30, (c) n=50, and (d) n=100. On each diagram, the true p-value is shown as a red dot. The 90% p-value coverage interval is shown between the two vertical bars, which corresponds to the 90% z-score coverage interval.

As can be seen from Figure 1, the distribution of the p-value varies substantially with sample size. The p-values are widely distributed between 0 and 1, indicating that there is great variation between the samples, regardless of sample size. Therefore, it is wrong to claim that the p-value reliably shows the degree of evidence against the null hypothesis[1].

6. Discussion

By decomposing the two-sample one-tailed z-test formulas, we have revealed that the true meaning of the p-value generated by it is the estimated probability that the sample mean ($\overline{X}_A$) of the treatment group is smaller than the sample mean ($\overline{X}_B$) of the control group. Then, the (1-p) value is the estimated probability that the sample mean ($\overline{X}_A$) is greater than the sample mean ($\overline{X}_B$). Assume that an experiment gives a p-value of 0.1. Then, the (1-p) value is 0.9. This means that there is a 90% chance (estimated) that the *mean score* of the treatment group will be higher than the *mean score* of the control group. Therefore, the two-sample one-tailed z-test actually compares the two groups at the "*sample mean level*" rather than at the "*individual level.*" In other words, the p-value is information about the difference between the two sampling distributions of the sample means $\overline{X}_A$ and $\overline{X}_B$. It is not information about the difference between the two population distributions of X_A and X_B.

It is important to note that, from a philosophical perspective, the fundamental principle of scientific inductive reasoning is that scientific claims must be based on statistical inference and domain knowledge about the *population* properties (i.e., population information) of the quantity under consideration (e.g., effect size)[11]. In many practical situations, however, we do not know population information (e.g., population parameters or population effect size), so we must use observed data to infer population information. Thus, statistical inference comes into play. For example, the population (or true) mean is the most important information about a population (or population distribution). In practical applications, since the population mean is usually unknown, the sample mean of the observed data is often used as an estimate of the population mean in scientific decision-making. In other words, the sample mean is the inferred population information (the population mean), so its use conforms to the fundamental principle of scientific inductive reasoning.

Although both the p-value and the sample mean are calculated using the observed data, their inferential implications are quite different. As we have shown in Section 3, the p-value is an estimate of the true p-value. However, the true p-value is *not* population information, so the p-value is *not* inferred population information. Therefore, the use of p-values violates the fundamental principle of scientific inductive reasoning. We argue that this is the fundamental philosophical reason why p-values should be abandoned in scientific research.

Furthermore, it is important to distinguish between two types of sample statistics: inferential and non-inferential. By definition, an inferential statistic is a sample statistic that can be used as an estimate of the corresponding population parameter. Examples of inferential statistics include the sample mean, sample standard deviation, and observed effect size. In contrast, a non-inferential statistic is a sample statistic that does not infer any population parameter or provide any information about the population (or population distribution). Instead, non-inferential statistics are merely characteristics or information about the sampling distribution. Examples of non-inferential statistics include standard errors, confidence intervals, and p-values. An important distinction between inferential and non-inferential statistics is whether a sample statistic depends on the sample size.

Inferential statistics are independent of sample size, although inferential statistics given by larger sample sizes should provide more precise and reliable inferences because large samples better represent the population. In contrast, non-inferential statistics are dependent on sample size; in fact, they are a function of sample size. Therefore, non-inferential statistics such as standard errors, confidence intervals, and *p*-values not only vary from sample to sample due to sampling error or noise, but also inherently decrease as sample size increases. However, standard errors and confidence intervals can be used as measures of uncertainty in an estimate of the population mean or observed effect size.

An appropriate probabilistic measure for scientific inductive reasoning is the exceedance probability: $\Pr(X_A > X_B)$[22]. The exceedance probability $\Pr(X_A > X_B)$ is inferred population information that does not depend on sample size. Therefore, unlike *p*-values, which can be easily hacked by *N*-chasing, the exceedance probability $\Pr(X_A > X_B)$ cannot be hacked by *N*-chasing. Furthermore, unlike *p*-values, which can be easily misunderstood or misinterpreted, the exceedance probability $\Pr(X_A > X_B)$ can be easily and clearly interpreted without causing confusion. Assume that an experiment gives a $\Pr(X_A > X_B)$ value of 0.7. This means that there is a 70% chance that a randomly picked person from the treatment group will score higher than a randomly picked person from the control group. In other words, unlike the *z*-test, which compares the two groups at the "*sample mean level*", the exceedance probability analysis compares the two groups at the "*individual level.*"

It should be mentioned that the concept of exceedance probability is essentially the same as the concept of gain-probability proposed by Trafimow et al.[23][12]. In addition, the meaning of $\Pr(X_A > X_B)$ is essentially the same as that of the common language effect size, the probability of superiority, or the area under the receiver operating characteristic[22]. Moreover, the concept of exceedance probability and its analysis have been applied to engineering fields such as environmental protection and water quality control (e.g.[24][25][26]). Detailed discussions about exceedance probability analysis can be found in Huang[22].

7. Conclusion

We have revealed that the true meaning of the *p*-value generated by the two-sample one-tailed z-test is the estimated probability that the mean score of the treatment group is smaller than the mean score of the control group. Accordingly, the (1-*p*) value is the estimated probability that the mean score of the treatment group is greater than the mean score of the control group. This interpretation of p-values does not involve the NHST setting and language and is therefore easy to understand even for people without statistical training.

The true p-value of the two-sample one-tailed z-test can be calculated when the population parameters are known. The usual p-value is an estimate of the true p-value. However, the true p-value is not population information, so the usual p-value is not inferred population information. Scientific inductive reasoning requires inferred population information. Therefore, it is wrong to use p-values for scientific inductive reasoning. This is the fundamental philosophical reason why p-values should be abandoned in scientific research.

Statements and Declarations

Conflicts of interest

No potential conflict of interest was reported by the author(s).

Funding

None.

References

1. [a, b] *Halsey L, Curran-Everett D, Vowler S, et al. (2015). "The fickle P value generates irreproducible results." Nat Methods, 12, 179–185. doi:10.1038/nmeth.3288.*
2. ^ *Amrhein V, Greenland S, McShane B. (2019). "Retire statistical significance." Nature 567, 305-307.*
3. ^ *McShane BB, Gal D, Gelman A, Robert C, Tackett JL. (2019). "Abandon Statistical Significance." The American Statistician, 73(sup1), 235–245. doi:10.1080/00031305.2018.1527253.*
4. [a, b] *Halsey LG. (2019). "The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum?" Biology Letters, 15(5), 20190174. doi:10.1098/rsbl.2019.0174.*
5. ^ *Wasserstein RL, Lazar NA. (2016). "The ASA's statement on p-values: context, process, and purpose." The American Statistician, 70, 129-133. doi:10.1080/00031305.2016.1154108.*
6. ^ *Wasserstein RL, Schirm AL, Lazar NA. (2019). "Moving to a World Beyond "p < 0.05."" The American Statistician, 73(sup1), 1–19. doi:10.1080/00031305.2019.1583913.*

7. ^Claridge-Chang A, Assam P. (2016). "Estimation statistics should replace significance testing." Nat Methods, 13, 108–109. doi:10.1038/nmeth.3729.
8. ^Berner D, Amrhein V. (2022). "Why and how we should join the shift from significance testing to estimation." J Evol Biol. 35(6), 777-787. doi: 10.1111/jeb.14009. PMID 35582935; PMCID PMC9322409.
9. ^Elkins MR, Pinto RZ, Verhagen A, Grygorowicz M, Söderlund A, Guemann M, Gómez-Conesa A, Blanton S, Brismée JM, Ardern C, Agarwal S, Jette A, Karstens S, Harms M, Verheyden G, Sheikh U. (2022). "Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors." Journal of physiotherapy, 68(1), 1–4. doi:10.1016/j.jphys.2021.12.001.
10. ^Huang H. (2023). "Statistics reform: practitioner’s perspective (preprint)." ResearchGate. https://www.researchgate.net/publication/373551061_Statistics_reform_practitioner's_perspective
11. a, b, cHuang H. (2024). "Comments on “The Roles, Challenges, and Merits of the p Value” by Chén et al." Basic and Applied Social Psychology, 1–7. doi:10.1080/01973533.2024.2442957.
12. a, bTrafimow D, Tong T, Wang T, Choy STB, Hu L, Chen X, Wang C, Wang Z. (2024). "Improving inferential analyses predata and postdata." Psychological methods, 10.1037/met0000697. Advance online publication. doi:10.1037/met0000697.
13. ^Lu Y, Belitskaya-Levy I. (2015). "The debate about p-values." Shanghai Arch Psychiatry. 27(6), 381-5. doi: 10.11919/j.issn.1002-0829.216027. PMID: 27199532; PMCID: PMC4858512.
14. ^Verhulst B. (2016). "In defense of p values." AANA J., 84(5), 305-308. PMID: 28366961 PMCID: PMC5375179
15. ^Benjamini Y, De Veaux R, Efron B, Evans S, Glickman M, Graubard BI, He X, Meng X-L, Reid N, Stigler SM, Vardeman SB, Wikle CK, Wright T, Young LJ, Kafadar K. (2021). "ASA President’s Task Force Statement on Statistical Significance and Replicability." Harvard Data Science Review, 3(3). doi:10.1162/99608f92.f0ad0287.
16. ^Hand DJ. (2022). "Trustworthiness of Statistical Inference." Journal of the Royal Statistical Society Series A: Statistics in Society, 185(1), 329–347. doi:10.1111/rssa.12752.
17. ^Lohse K. (2022). "In Defense of Hypothesis Testing: A Response to the Joint Editorial From the International Society of Physiotherapy Journal Editors on Statistical Inference Through Estimation." Physical Therapy, 102(11), 118. doi:10.1093/ptj/pzac118.
18. a, b, c, dChén OY, Bodelet JS, Saraiva RG, Phan H, Di J, Nagels G, Schwantje T, Cao H, Gou J, Reinen JM, Xiong B, Zhi B, Wang X, de Vos M. (2023). "The roles, challenges, and merits of the p value." Patterns (New York, N.Y.), 4(12), 100878. doi:10.1016/j.patter.2023.100878.

19. a, b*Goodman S. (2008). "A dirty dozen: twelve p-value misconceptions." Seminars in hematology, 45 (3), 135–140. doi:10.1053/j.seminhematol.2008.04.003.*

20. ^*Lytsy P, Hartman M, Pingel R. "Misinterpretations of P-values and statistical tests persists among researchers and professionals working with statistics and epidemiology." Ups J Med Sci. 2022 Aug 4;127. doi: 10.48101/ujms.v127.8760. PMID: 35991465; PMCID: PMC9383044.*

21. a, b*Lazzeroni LC, Lu Y, Belitskaya-Lévy I. (2016). "Solutions for quantifying P-value uncertainty and replication power." Nature methods, 13(2), 107–108. doi:10.1038/nmeth.3741.*

22. a, b, c*Huang H. (2022). "Exceedance probability analysis: a practical and effective alternative to t-tests." Journal of Probability and Statistical Science, 20(1), 80-97.*

23. ^*Trafimow D, Hyman MR, Kostyk A, Wang Z, Tong T, Wang T, Wang C. (2022). "Gain-probability diagrams in consumer research." International Journal of Market Research, 64(4), 470-483. doi:10.1177/14707853221085509.*

24. ^*Environment protection agency (EPA) (1991). Technical support document for water quality-based toxics control, Office of Water, Washington, DC, EPA/505/2-90-001*

25. ^*Di Toro DM. (1984). "Probability model of stream quality due to runoff." Journal of Environmental Engineering, ASCE, 110(3), 607-628.*

26. ^*Huang H, Fergen RE. (1995). "Probability-domain simulation - A new probabilistic method for water quality modeling." WEF Specialty Conference, "Toxic Substances in Water Environments: Assessment and Control" (Cincinnati, Ohio, May 14-17, 1995).*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.