BMC Medical Research Methodology

RESEARCH

Open Access

A logical analysis of null hypothesis significance testing using popular terminology

Richard McNulty*

Abstract

Background: Null Hypothesis Significance Testing (NHST) has been well criticised over the years yet remains a pillar of statistical inference. Although NHST is well described in terms of statistical models, most textbooks for non-statisticians present the null and alternative hypotheses (H_0 and H_A, respectively) in terms of differences between groups such as ($\mu_1 = \mu_2$) and ($\mu_1 \neq \mu_2$) and H_A is often stated to be the research hypothesis. Here we use propositional calculus to analyse the internal logic of NHST when couched in this popular terminology. The testable H_0 is determined by analysing the scope and limits of the *P*-value and the test statistic's probability distribution curve.

Results: We propose a minimum axiom set NHST in which it is taken as axiomatic that H_0 is rejected if *P*-value< α. Using the common scenario of the comparison of the means of two sample groups as an example, the testable H_0 is {($\mu_1 = \mu_2$) and [($\bar{x}_1 \neq \bar{x}_2$) due to chance alone]}. The H_0 and H_A pair should be exhaustive to avoid false dichotomies. This entails that H_A is ¬{($\mu_1 = \mu_2$) and [($\bar{x}_1 \neq \bar{x}_2$) due to chance alone]}, rather than the research hypothesis (H_T). To see the relationship between H_A and H_T, H_A can be rewritten as the disjunction H_A: ({($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to chance alone]} ∨ {($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to ($\mu_1 \neq \mu_2$) alone]} ∨ **{($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to ($\mu_1 \neq \mu_2$) alone]}**). This reveals that H_T (the last disjunct in bold) is just one possibility within H_A. It is only by adding premises to NHST that H_T or other conclusions can be reached.

Conclusions: Using this popular terminology for NHST, analysis shows that the definitions of H_0 and H_A differ from those found in textbooks. In this framework, achieving a statistically significant result only justifies the broad conclusion that the results are not due to chance alone, not that the research hypothesis is true. More transparency is needed concerning the premises added to NHST to rig particular conclusions such as H_T. There are also ramifications for the interpretation of Type I and II errors, as well as power, which do not specifically refer to H_T as claimed by texts.

Keywords: Logic, Null hypothesis significance test, Hypothesis testing, Statistical inference, Statistical significance, Type I error, Type II error, Power

Background

Null Hypothesis Significance Testing (NHST[1]) and the Confidence Interval (CI) or estimation method are the pillars of statistical inference [1–5]. NHST is perhaps the more common of the two for the analysis of research questions [6]. In NHST a null hypothesis (H_0) is rejected

*Correspondence: richard.mcnulty@health.nsw.gov.au

Emergency Department, Blacktown Mount Druitt Hospitals, Blacktown Rd, Blacktown, Sydney, NSW 2148, Australia

[1] "NHST" is probably the most widely used abbreviation for the various names applied to hypothesis and significance tests 1. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods* 2000; 5: 241–301. 2000/08/11. DOI: https://doi.org/10.1037/1082-989x.5.2.241.

in favour of an alternative hypothesis (H_A) only if the *P*-value, *P* (observed data or more extreme | H_0), falls below a pre-specified α-level. The latter is the maximum probability we are prepared to tolerate of erroneously rejecting H_0. If the *P*-value is less than α, then this is called a statistically significant result and H_0 can be rejected. Some familiarity with NHST will be assumed in this paper. NHST is a combination of two different statistical theories: R. A. Fisher's *P*-value significance test, and the Neyman-Pearson technique of hypothesis testing. The two groups never intended to unite the theories, with well-known antagonisms existing between them [7]. However, NHST gained traction perhaps due to its appeal as a mechanical decision tool. Parallel to its popularity is the detailed, sharp criticism it has received from several quarters. Problems raised include: the misinterpretation of the *P*-value as $P(H_0 \mid \text{observed data})$ rather than *P* (observed data or more extreme | H_0); the artificial dichotomous nature of statistical significance; and the conflation of statistical significance with clinical importance [8]. In fact, *P*-values have even been temporarily banned from some journals [9]. More recently, the correct level of statistical significance (*P*-value or α cut-off) has again been debated [10]. However, rather than cover old ground, we will here present a new logical analysis of a popular version of NHST presented in textbooks. NHST is perhaps best explained in terms of statistical models [11]. However, in most popular textbooks for non-statisticians, NHST is frequently presented in terms of the difference between population or sample groups and framed in reference to the research hypothesis. The need for an in-depth focus on the logic of NHST when couched in these terms can be seen from the following summary.

Starting with H_0, there are various definitions offered. H_0 is the hypothesis of no difference or association between groups [1, 5, 12–27]. Using population means (μ) as an example, this is H_0: $\mu_1 = \mu_2$, meaning there is no difference in the population [2, 28–33]. In addition, there is the idea that H_0 is the opposite/reverse/complement/negation of the test/experimental/study/research hypothesis [1, 3, 6, 25, 27, 28]. In clinical studies, this segues to the stronger claim that the absence of a difference is due to a lack of treatment effect [3, 5, 6, 13, 20, 21, 28, 31, 34–36]. In contrast to the idea of "no difference" is the anticipation that chance or random variation will produce a difference between the sample means [37]. Some texts unite the two ideas about the presence and absence of difference into one H_0 which states there is no difference in the population and the difference in the sample groups is due to chance [2, 38–41]. Although a symbol exists for the mean of the sample group ($\bar{x}$), there was no example of this more complex version of H_0 translated into symbols in any text sampled. In fact, some texts mention this more complex H_0 only to quickly drop the idea and revert to H_0: $\mu_1 = \mu_2$ anyway [27, 42].

Moving on to the definition of H_A, we find similar themes phrased in a contrary fashion. H_A is the hypothesis that there is a difference or association between the groups [12, 13, 22, 23, 32]. Some specify that the groups are the populations such that H_A: $\mu_1 \neq \mu_2$ [2, 4, 24]. This type of difference is described as statistically significant [26] or real [2, 17, 18, 42, 43]. H_A is elsewhere proposed to be: the experimental/ research/ study hypothesis [3, 5, 6, 28, 36, 43]; or the hypothesis that there is a treatment effect [1, 6, 20, 33, 34, 39]; or the contradictory or complementary hypothesis to H_0 [14, 34, 35, 42]. There are attempts to unite claims about the population and sample groups, namely that the difference in the sample groups is due to the difference in the population [42]. Again, in the texts sampled, the latter hypothesis was never translated into symbols or further pursued.

Another area of disagreement, apart from the content of H_A, is the strength of the conclusion when rejecting H_0. Some claim we accept H_A as true [1, 5, 16, 20, 23] or real [18]. There are also softer versions that state H_A is just "supported" or is "probably true" [6, 19]. Alternatively, conclusions can be framed in terms of the test hypothesis being true [2, 15, 16, 20, 27, 29, 33–35, 43, 44], or more tentatively, that we gain confidence or support for the test hypothesis [6, 25, 28, 31, 41, 42]. More bewildering still are claims suggesting there are multiple other hypotheses or explanations! [1, 12, 16, 21, 34, 35, 40]

The interpretation of the phrase "statistically significant" [2, 5, 21, 34, 39, 40, 42], often abbreviated to just "significant" [21, 25, 27, 28, 30, 33–35], ranges from the claim that the data are not due to chance [24, 45] to the weaker claim that the data are unlikely to be due to chance [2, 18, 40].

In NHST, H_0 and H_A are presented as a hypothesis pair. A commonly presented pair is H_0: $\mu_1 = \mu_2$ and H_A: $\mu_1 \neq \mu_2$. This hypothesis pair is mutually exclusive and exhaustive which some texts explicitly state are desirable characteristics [1, 19, 46]. Elsewhere, however, H_0 and H_A are frequently presented as a non-exhaustive, false dichotomy between the test hypothesis and the hypothesis that the results are due to chance [3, 6, 16, 18, 19, 24, 25, 27, 34, 38, 40, 41, 44].

From the above we see that this family of interpretations of NHST provides no consensus on many aspects. This poses a challenge to interpreting NHST when expressed in this fashion. From within the framework of this popular terminology, the purpose of the present paper is to

1/ define H_0, H_A, power and type I and II errors,
2/ define the minimum axiom set for NHST and
3/ make transparent which assumptions are needed to conclude the research hypothesis is true.

Methods

Here we assume the common terminology of expressing NHST in terms of differences between populations or sample groups and in reference to the research hypothesis. The scope and limits of the *P*-value, the test statistic and its probability distribution curve (PDC) will be used to arbitrate on the correct form of H_0 and H_A within this framework. Propositional calculus will be employed to analyse NHST. We also acknowledge multi-factorial hypotheses. For example, we can hypothesise that the difference between two sample groups is due to bias, chance or an intervention. These hypotheses are independent which entails that they can act in combination to produce the results. To disambiguate between single- or multi-factorial hypotheses, the term "alone" will be used to refer to the former. For example, "$(\bar{x}_1 \neq \bar{x}_2)$ due to chance alone" means chance is the only factor involved in the sample group difference, as opposed to chance acting in concert with other factors to produce the results.

Results

For consistent vocabulary throughout this paper, we will use as our example the common scenario of comparing the means of two sample groups. The appropriate test statistic for this is the *t*-statistic which has its relevant PDC. We will commence by stating the minimum axiom set needed for a NHST to function. To this end, we accept as axiomatic that if $P(\text{observed data or more extreme} \mid H_0) < \alpha$, then reject H_0 and accept H_A.

The testable H_0

In the introduction we saw that H_0 had various definitions including H_0: $\mu_1 = \mu_2$ or the "opposite" of the research hypothesis. Understandably, these are H_0's that we would like to test, but that does not guarantee that these candidates are testable. Here we propose a new approach: the decision concerning which is the correct H_0 should be determined by the scope and limits of the actual technique that will be used to reject H_0. In our example, the decision to reject H_0 is based on the *P*-value of the *t*-statistic read off from its PDC. The PDC yields the probability of finding the observed *t*-statistic value (or more extreme) due to chance alone when there is no difference in the population means. In symbols, (something which never appeared in the texts mentioned in the introduction), the PDC gives us

$$P\big(\text{observed } t-\text{statistic value or more extreme} \mid \{(\mu_1 = \mu_2) \text{ and } [(\bar{x}_1 \neq \bar{x}_2) \text{ due to chance alone}]\}\big).$$

Given that the definition of the *P*-value is

$$P(\text{observed } t-\text{statistic value or more extreme} \mid H_0),$$

we can now see that the H_0 which the *P*-value and PDC can actually test must be

$$H_0 : \{(\mu_1 = \mu_2) \text{ and } [(\bar{x}_1 \neq \bar{x}_2) \text{ due to chance alone}\}.$$

In other words, it is the hypothesis that the finding in the sample groups is due to chance or random variation alone and does not reflect a difference in the underlying population.

Rejecting $(\mu_1 = \mu_2)$

Textbooks often claim that we can use NHST to reject $(\mu_1 = \mu_2)$. However, this is not logically possible with the minimum axiom set NHST. To demonstrate this, we will need to transform $(\mu_1 = \mu_2)$ to a logically equivalent proposition and use propositional calculus. The proposition $(\mu_1 = \mu_2)$ is a proposition about the equality of the population means, but states nothing about the sample group means $(\bar{x})$. Using a truth table (Table 1), we can rewrite $(\mu_1 = \mu_2)$ in a logically equivalent way such that the sample group means do appear in the proposition but without any claim being made about them.[2] Note that $P(\bar{x}_1 = \bar{x}_2) = 0$, so any proposition containing $(\bar{x}_1 = \bar{x}_2)$ can be eliminated from the analysis.

From Table 1, $(\mu_1 = \mu_2) \equiv$

$$\big(\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ due to chance alone}]\} \vee \{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ not due to chance alone}]\}\big). \quad (1)$$

Logical equivalence is established because whenever $(\mu_1 = \mu_2)$ is true, 1 is true too, and whenever $(\mu_1 = \mu_2)$ is false, 1 is also false. This transformation now allows us to see why eliminating the testable H_0 does not logically imply the elimination of $(\mu_1 = \mu_2)$. Let the first

[2] Truth tables analyse the truth of complex propositions based on giving truth values of true (T) or false (F) to its elemental components. When propositions are subject to logical analysis here, we shall use the symbols of propositional calculus: "$\wedge$" for "and"; "$\vee$" for "or"; and "$\neg$" for "not" used to express negation. "$\neg$X" means "It is not the case that X." "$\equiv$" means "is equivalent to" such that "X $\equiv$ Y" means "proposition X is equivalent to proposition Y."

Table 1 Truth table for $(\mu_1 = \mu_2)$ and its logical equivalent

$\mu_1 = \mu_2$	$(\bar{x}_1 \neq \bar{x}_2)$ due to chance alone	$(\bar{x}_1 \neq \bar{x}_2)$ not due to chance alone	$\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2)$ due to chance alone]}	$\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2)$ not due to chance alone]}	$(\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2)$ due to chance alone]} $\vee$ $\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2)$ not due to chance alone]})
T	T	F	T	F	T
T	F	T	F	T	T
F	T	F	F	F	F
F	F	T	F	F	F

disjunct of 1 be called C, and the second disjunct E. Thus, 1 becomes the disjunction C v E. We recognise C as the testable H_0. The PDC can assess C, and so it may be possible to reject C depending on the *P*-value. However, the PDC cannot assess E. So even if we do reject C, we cannot reject E, and therefore we cannot reject the whole proposition C v E. Since 1 is logically equivalent to $(\mu_1 = \mu_2)$, we see that we cannot reject $(\mu_1 = \mu_2)$ using the minimum axiom set NHST. In other words, $(\mu_1 = \mu_2)$ is not rejected when we reject the testable H_0: $\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2)$ due to chance alone]}. To reject $(\mu_1 = \mu_2)$, a further premise will need to be added, namely $\neg\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2)$ not due to chance alone]}.

The real H_A

We take it as axiomatic that H_0 and H_A are mutually exclusive: the hypotheses should not overlap in the sample space. An issue identified in the introduction was whether the hypothesis pair should also be exhaustive. There are serious consequences when the pair are made into a false dichotomy. An obvious criticism is that other possibilities are simply ignored. Furthermore, it opens a Pandora's box of candidates for H_A. Frequently the research or test hypothesis (here H_T) is proposed as H_A. This is the proposition that there is a difference in the population due to the study intervention or treatment and the finding in the sample groups is due to this difference alone. In symbols

$$H_T : \{(\mu_1 \neq \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ due to } (\mu_1 \neq \mu_2) \text{ alone}]\}.$$

However, if false dichotomies are allowed, what is to prevent other hypotheses being proposed as H_A? Such as the hypothesis that bias or confounding produced the results, or some other hypothesis, or even combinations of hypotheses given that they are all independent propositions. In a false dichotomy the selection of H_A is subject to prejudice.

The above problems are avoided by forming an exhaustive hypothesis pair. To avoid logical errors of negation, it is critical to note that H_A must be the negation of the entire proposition represented by H_0, not just a negation of part of H_0. So H_A must be $\neg H_0$ and the real H_A: $\neg\{(\mu_1 = \mu_2)$ and $[(\bar{x}_1 \neq \bar{x}_2)$ due to chance alone]}. Therefore, the only justifiable exhaustive hypothesis pair is

$$H_0 : \big\{(\mu_1 = \mu_2) \text{ and } [(\bar{x}_1 \neq \bar{x}_2) \text{ due to chance alone}]\big\},$$

$$H_A : \neg\big\{(\mu_1 = \mu_2) \text{ and } [(\bar{x}_1 \neq \bar{x}_2) \text{ due to chance alone}]\big\}.$$

The relationship between H_A and H_T

H_A is a more complex proposition than H_T. Once again, we can transform H_A into a logically equivalent proposition which has H_T as a component. Let H_A be represented by ¬(G ∧ J), where G is "$\mu_1 = \mu_2$," and J is "$(\bar{x}_1 \neq \bar{x}_2)$ due to chance alone." The truth table for ¬(G ∧ J) is shown in Table 2.

Table 2 shows that ¬(G ∧ J) is true (bold **T** in last column) when G and ¬J are true (the second row), or ¬G and J are true (the third row), or ¬G and ¬J are true (the last row). This allows us to formulate a disjunction logically equivalent to ¬(G ∧ J). Thus $\neg(G \wedge J) \equiv (G \wedge \neg J) \vee (\neg G \wedge J) \vee (\neg G \wedge \neg J)$. Now $\neg J \equiv \{(\bar{x}_1 = \bar{x}_2) \vee [(\bar{x}_1 \neq \bar{x}_2)$ not due to chance alone]}. However, as stated previously, we can eliminate $(\bar{x}_1 = \bar{x}_2)$ making $\neg J \equiv [(\bar{x}_1 \neq \bar{x}_2)$ not due to chance alone]. Substituting back, $H_A \equiv$

$$(\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ not due to chance alone}]\} \vee \{(\mu_1 \neq \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ due to chance alone}]\} \vee \{(\mu_1 \neq \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ not due to chance alone}]\}).$$

Furthermore, the second disjunct is a contradiction and can be eliminated giving

$$\begin{aligned} H_A : &\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ not due to chance alone}]\} \\ &\vee \{(\mu_1 \neq \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ not due to chance alone}]\}. \end{aligned} \quad (2)$$

Where does H_T lie in 2? H_T is contained within the last disjunct of 2, $\{(\mu_1 \neq \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2)$ not due to chance alone]}. The latter disjunct expresses the proposition that there is a difference found in the population and also that the sample group difference is not due to

Table 2 Truth table for ¬(G ∧ J)

G	¬G	J	¬J	G ∧ J	¬(G ∧ J)
T	F	T	F	T	F
T	F	F	**T**	F	**T**
F	**T**	**T**	F	F	**T**
F	**T**	F	**T**	F	**T**

Table 3 Adding premises to NHST to conclude H_T. Comparison of group means is used as an example. H_T (in bold) is defined in the text

Perform NHST: if *P*-value ≥ α, then fail to reject H_0. If *P*-value < α, H_0 is rejected and conclude H_A: ¬{($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to chance alone]}			
Further steps	**Aim to conclude H_T**	**Assume "there is no bias"**	**Aim to conclude H_B**
Additional premises	(1) ¬{($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to chance alone]} (2) ¬{($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to ($\mu_1 \neq \mu_2$) alone nor chance alone]}	(1) ¬{($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to bias]}	(1) ¬ {($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to bias alone nor chance alone]} (2) ($\mu_1 = \mu_2$)
Reasoning	H_A: ¬{($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to chance alone]} ≡ H_A: ({($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to chance alone]} ∨ {($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to ($\mu_1 \neq \mu_2$) alone nor chance alone]} ∨ **{($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to ($\mu_1 \neq \mu_2$) alone]}**). Use 2 steps of disjunction elimination with (1) and (2)	H_A: ¬{($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to chance alone]} ≡ H_A: ({($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to bias nor chance alone]} ∨ {($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to bias]} ∨ **{($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to ($\mu_1 \neq \mu_2$) alone]}** ∨ {($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to ($\mu_1 \neq \mu_2$) alone]}). Use disjunction elimination with (1)	H_A: ¬{($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to chance alone]} ≡ H_A: ({($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to bias alone]} ∨ {($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to bias alone nor chance alone]} ∨ ($\mu_1 \neq \mu_2$)). Use 2 steps of disjunction elimination with (1) and (2)
Conclusion	Therefore **{($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to ($\mu_1 \neq \mu_2$) alone]}, i.e., H_T**	Therefore ({($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to bias nor chance alone]} ∨ {($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) not due to ($\mu_1 \neq \mu_2$) alone]} ∨ **{($\mu_1 \neq \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to ($\mu_1 \neq \mu_2$) alone]}**)	Therefore {($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to bias alone]}, i.e., H_B

chance alone, but instead is due to some other alternative. The other alternatives include the test intervention or bias or some other unknown or even a combination of these given that the alternatives are independent hypotheses. Taking this into account we can rewrite 2 such that $H_A \equiv$

Some texts claim that all that is needed to conclude H_T when H_0 is rejected is the assumption that there is no bias [35, 47]. However, Table 3 illustrates exactly which premises are needed in order to conclude H_T. Apart from assuming no bias, it is also necessary to assume

$$\{(\mu_1 = \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ not due to chance alone}]\} \vee \{(\mu_1 \neq \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ not due to } (\mu_1 \neq \mu_2) \text{ alone}]\} \vee \{(\boldsymbol{\mu_1 \neq \mu_2}) \wedge [(\bar{\boldsymbol{x}}_1 \neq \bar{\boldsymbol{x}}_2) \textbf{ due to } (\boldsymbol{\mu_1 \neq \mu_2}) \textbf{ alone}]\}. \quad (3)$$

The last disjunct of 3 is H_T (in bold), indicating that H_T is just one sub-hypothesis of H_A.

Finally, the answer to the question "What do we accept when we reject H_0?" is: we accept the real H_A or its logical equivalent (3). Therefore, a statistically significant finding, expressed in these common terms, should be interpreted as meaning that the data is not due to chance alone. Statistical significance is not a licence to accept H_T.

there are no combination hypotheses in which chance plays a role. A corollary is that if NHST could lead us to conclude H_T of its own accord, no further premises would be required. What would the conclusion be if indeed we only assumed that there was no bias? The middle column of Table 3 shows the conclusion. In a model which stipulates that the possible causes of the sample group difference are chance, bias or the intervention (or combinations thereof), the conclusion would be

$$\boldsymbol{\{(\mu_1 \neq \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \textbf{ due to } (\mu_1 \neq \mu_2) \textbf{ alone}]\}} \vee \{[(\mu_1 \neq \mu_2) \wedge [(\bar{x}_1 \neq \bar{x}_2) \text{ due to } (\mu_1 \neq \mu_2) \text{ and chance}]\}.$$

The effect of further premises on the minimum axiom set NHST

It is only by adding premises to NHST that we can conclude anything other than the real H_A. The danger with this strategy is that of partially assuming what is being proved. Table 3 presents examples of premises that if added to NHST would rig different conclusions.

The first disjunct in bold is H_T, showing that the conclusion is more complex than H_T alone. The last column demonstrates that a different package of additional premises can be tailored to reach a different conclusion such as the hypothesis that bias produced the results, here represented as H_B: {($\mu_1 = \mu_2$) ∧ [($\bar{x}_1 \neq \bar{x}_2$) due to bias alone]}. Similar to arithmetic, the process in Table 3 is commutative. The same results are achieved if we were to make the assumptions first and then do the NHST or vice versa —— the order does not matter.

Application to other statistical problems

So far we have focused on the comparison of sample group means. However, with appropriate changes in vocabulary we can define the real H_0 and H_A for other scenarios —— mutatis mutandis, as they say. As illustrations, H_0 and H_A in general form, for the comparison of sample group proportions, and for correlation are presented in Table 4.

Failure to reject H_0

What are we to conclude if we fail to reject H_0? The axiom of NHST states that we reject H_0 if P-value < α. This does not logically imply that if P-value ≥ α we must accept H_0 —— the axiom and the claim about accepting H_0 are logically distinct ideas. So if P-value ≥ α, we should merely state we have failed to reject H_0 rather than we accept H_0.

Power (1-β), type I (α) and type II (β) errors

Textbooks which express NHST in terms of the research hypothesis also tend to carry this over to descriptions of Type I and II errors, as well as power calculations. However, this is fraught with error as can be seen when we apply the real definitions of H_0 and H_A. Type I error is the probability of eliminating H_0, and accepting H_A, when in fact H_0 is true. Using the real definitions of H_0 and H_A gives us type I error:

$$P\left(\text{rejecting } \left\{\left(\mu_1 = \mu_2\right) \text{ and } \left[\left(\bar{x}_1 \neq \bar{x}_2\right) \text{ due to chance alone}\right]\right\} \middle| \left\{\left(\mu_1 = \mu_2\right) \text{ and } \left[\left(\bar{x}_1 \neq \bar{x}_2\right) \text{ due to chance alone}\right]\right\}\right).$$

Importantly, type I error is not the probability of accepting H_T when H_0 is true. Since H_A is a disjunction, there are multiple propositions that can make it true, with H_T being just one of these. So $P(H_A) > P(H_T)$ and P(mistakenly accepting H_T) > P(mistakenly accepting H_A). The conflation of H_T with H_A results in underestimating the probability of mistakenly accepting H_T.

Similarly for type II error which is the probability of not rejecting H_0, and not accepting H_A, when H_0 is false and should have been rejected. Namely,

$$P\left(\text{not rejecting } \left\{\left(\mu_1 = \mu_2\right) \text{ and } \left[\left(\bar{x}_1 \neq \bar{x}_2\right) \text{ due to chance alone}\right]\right\} \middle| \text{it is not the case that } \left\{\left(\mu_1 = \mu_2\right) \text{ and } \left[\left(\bar{x}_1 \neq \bar{x}_2\right) \text{ due to chance alone}\right]\right\}\right).$$

Type II error is not the probability of not accepting H_T when H_0 is false. A low probability of not accepting H_A does not logically imply a low probability of not accepting H_T. P(not accepting H_T) > P(not accepting H_A) because more propositions need to be rejected in order to accept H_T. The conflation of H_T with H_A results in underestimating the probability of not accepting H_T when H_0 is false.

Power (1- β) refers to the probability of rejecting H_0 and accepting H_A given H_0 is false. Specifically, power is

$$P\left(\text{rejecting } \left\{\left(\mu_1 = \mu_2\right) \text{ and } \left[\left(\bar{x}_1 \neq \bar{x}_2\right) \text{ due to chance alone}\right]\right\} \middle| \text{it is not the case that } \left\{\left(\mu_1 = \mu_2\right) \text{ and } \left[\left(\bar{x}_1 \neq \bar{x}_2\right) \text{ due to chance alone}\right]\right\}\right).$$

However, it does not refer to P(accepting $H_T \mid H_T$). The power to conclude H_T < the power to conclude H_A. The conflation of H_T with H_A results in overestimating the power to conclude H_T because H_T is just one part of H_A.

Discussion

NHST has been well described in terms of statistical models. However, it is also commonly presented in terms of group comparisons and with reference to the research hypothesis. Despite this being a popular interpretation, there is currently no standardised approach. The variation in definitions of H_0 and H_A, how they should be paired and conclusions that can be drawn by eliminating H_0 motivated this new logical analysis. Looking at the conditions of the P-value we can see that there can be only one testable H_0. Presenting H_0 and H_A as a false dichotomy is common but unjustifiable. Combining these two ideas entails that H_A is $\neg H_0$. Texts should acknowledge this and also make transparent any premises added in order to reach a conclusion other than $\neg H_0$ when H_0 is rejected.

It may be thought that using the estimation or CI method can avoid the problems of expressing NHST in these terms. However, this is not true if the estimation method is used as a de facto NHST. The estimation method can be used as a NHST because the CI is mathematically related to the α-level and the P-value such that if the CI does not cross zero (or 1 for ratios), we can claim statistical significance. In the context of using CI as a NHST, the conclusions of the present paper are relevant. Consequently, when using the CI method, the correct interpretation of statistical significance would be to accept the real H_A and not claim that H_T is true. Of course, there are other appealing features of the CI method and the present discussion is limited only to its use as a significance test.

A limitation of the present paper is that we have not questioned the axiom of NHST that we reject H_0 if the P-value < α. An analysis of this axiom deserves a paper in its own right which discusses inductive logic and defines the conditions under which the axiom is reliable. The issue in the present paper has been solely that if we are to use NHST as it is commonly presented it should at

Table 4 H_0 and H_A for common scenarios. H_A has also been transformed into its logical equivalent to identify H_T (in bold)

Scenario	General Form	Comparing proportions (Chi-squared test)	Correlation
H_0 **and** H_A	H_0: there is no finding in the population and the finding in the sample group is due to chance alone H_A: it is not the case that H_0, therefore H_A: it is not the case that (there is no finding in the population and the finding in the sample group is due to chance alone) ≡ [(there is no finding in the population and the finding in the sample group is not due to chance alone) or (there is a finding in the population and the finding in the sample group is not due to the population finding alone) or **(there is a finding in the population and the finding in the sample group is due to the population finding alone)**]	H_0: $(\hat{p}_1 = \hat{p}_2) \wedge [(p_1 \neq p_2)$ due to chance alone] H_A: $\neg H_0$, therefore H_A: $\neg\{(\hat{p}_1 = \hat{p}_2) \wedge [(p_1 \neq p_2)$ due to chance alone]$\} \equiv$ H_A: $(\{(\hat{p}_1 = \hat{p}_2) \wedge [(p_1 \neq p_2)$ not due to chance alone]$\} \vee \{(\hat{p}_1 \neq \hat{p}_2) \wedge [(p_1 \neq p_2)$ not due to $(\hat{p}_1 \neq \hat{p}_2)$ alone]$\} \vee$ **$\{(\hat{p}_1 \neq \hat{p}_2) \wedge [(p_1 \neq p_2)$ due to $(\hat{p}_1 \neq \hat{p}_2)$ alone]$\}$**)	H_0: $(\rho = 0) \wedge (r \neq 0$ due to chance alone) H_A: $\neg H_0$, therefore H_A: $\neg[(\rho = 0) \wedge (r \neq 0$ due to chance alone)$] \equiv$ H_A: $([(\rho = 0) \wedge (r \neq 0$ not due to chance alone)$] \vee \{(\rho \neq 0) \wedge [r \neq 0$ not due to $(\rho \neq 0)$ alone]$\} \vee$ **$\{(\rho \neq 0) \wedge [r \neq 0$ due to $(\rho \neq 0)$ alone]$\}$**)

least be with justifiable definitions of H_0 and H_A, transparent assumptions and valid deductions from the given premises.

Conclusions

NHST is commonly expressed in terms of differences between groups and with reference to the research hypothesis. Within this framework, logical analysis reveals that the minimum axiom set NHST (for comparing sample means) is as follows:

H_0: {($\mu_1 = \mu_2$) and [($\bar{x}_1 \neq \bar{x}_2$) due to chance alone]},
H_A: ¬{($\mu_1 = \mu_2$) and [($\bar{x}_1 \neq \bar{x}_2$) due to chance alone]}.
If P-value ≥ α, then fail to reject H_0.
If P-value < α, reject H_0 and conclude H_A.

At best, it can be concluded that if H_0 is rejected, the data were not due to chance alone. Texts should also be transparent about which assumptions have been added to rig a conclusion such as H_T. Care should also be exerted to avoid misinterpreting type I and II errors, as well as power, in terms of the research hypothesis.

Acknowledgements
The anonymous reviewers are thanked for many useful comments.

List of abbreviations and symbols
α: alpha-level. The pre-specified acceptable ceiling on the type I error. The threshold which defines the critical region of the PDC, or the threshold below which the *P*-value has to fall in order to reject H_0.
β: type II error. The probability of not rejecting H_0 when H_0 is false.
H_A: the alternative hypothesis to H_0 which is accepted only when H_0 is rejected.
H_B: the hypothesis that bias is solely responsible for the research finding.
H_0: the null hypothesis. In NHST, it is only rejected when *P*-value < α.
H_T: the test or research hypothesis. Sometimes cited as the candidate for H_A. For example, the hypothesis that a drug is the cause of a difference between two sample groups, or there is an association between two variables.
μ: mu. The mean of the population.
NHST: null hypothesis significance test/testing. It will be used here as an umbrella term referring to both "test" or "testing" which will be clear from the context.
P-value: *P*(observed data (or more extreme)) | H_0).
PDC: probability distribution curve of the test statistic.
p: the sample proportion.
p̂: the population proportion.
ρ (rho): population Pearson correlation coefficient.
r: sample group Pearson correlation coefficient.
$\bar{x}$: the mean of the sample group.
∧: and, used to express conjunction.
∨: or, used to express disjunction.
¬: not, used to express negation. "It is not the case that.."
≡: logical equivalence. E.g., "X ≡ Y" means proposition X is logically equivalent to proposition Y.

Author's contributions
RM is sole author. The author(s) read and approved the final manuscript.

Funding
N/a

Availability of data and materials
All data generated or analysed during this study are included in this published article.

Declarations

Ethics approval and consent to participate
N/a

Consent for publication
N/a

Competing interests
The authors declare no competing interests.

Received: 9 March 2022 Accepted: 20 July 2022
Published online: 19 September 2022

References

1. Daniel WW. Biostatistics : a foundation for analysis in the health sciences. 9th ed. Hoboken: Wiley; 2009.
2. Munro BH, Page EB. Statistical methods for health care research, vol. xi. 2nd ed. Philadelphia: Lippincott; 1993. p. 403.
3. Gallin JI, Ognibene FP, Johnson LL. Principles and practice of clinical research, vol. xvii. 4th ed. London: Academic Press; 2018. p. 80.
4. Mann PS, Lacke CJ. Introductory statistics, vol. xx. 7th ed. Hoboken: Wiley; 2010. p. 116.
5. Sullivan LM. Essentials of biostatistics in public health, vol. xii. 3rd ed. Burlington: Jones & Bartlett Learning; 2018. p. 376.
6. Field AP. Discovering statistics using IBM SPSS statistics : and sex and drugs and rock 'n' roll, vol. xxxvi. 4th ed. Los Angeles: Sage; 2013. p. 915.
7. Salsburg D. The lady tasting tea : how statistics revolutionized science in the twentieth century, vol. xi. New York: W.H. Freeman; 2001. p. 340.
8. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods. 2000;5:241–301. https://doi.org/10.1037/1082-989x.5.2.241.
9. Trafimow D, Marks M. Editorial. Basic Appl Soc Psychol. 2015;37:1–2. https://doi.org/10.1080/01973533.2015.1012991.
10. Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. JAMA. 2018;319:1429–30. https://doi.org/10.1001/jama.2018.1536.
11. Lehmann EL, Romano JP. Testing statistical hypotheses, vol. xiv. 3rd ed. New York: Springer; 2005. p. 784.
12. Stewart A. Basic statistics and epidemiology : a practical guide, vol. iv. 3rd ed. Oxford: Radcliffe Pub; 2010. p. 200.
13. Everitt B. Medical statistics from A to Z : a guide for clinicians and medical students, vol. vi. 2nd ed. Cambridge: Cambridge University Press; 2006. p. 249.
14. Gerstman BB. Basic biostatistics : statistics for public health practice, vol. xv. 2nd ed. Burlington: Jones & Bartlett Learning; 2015. p. 644.
15. Hickson M. Research handbook for health care professionals, vol. xiv. Chichester, U.K: Wiley-Blackwell; 2008. p. 184.
16. Katz MH. Study design and statistical analysis : a practical guide for clinicians. Cambridge: Cambridge University Press; 2006. p. 188.
17. Katz DL, Jekel JF. Jekel's epidemiology, biostatistics, preventive medicine, and public health, vol. xiii. 4th ed. Philadelphia, London: Saunders; 2014. p. 405.
18. O'Brien PMS, Broughton-Pipkin F. Introduction to research methodology for specialists and trainees. 3rd ed. Cambridge, New York: Cambridge University Press; 2017.
19. Townend J. Practical statistics for environmental and biological scientists, vol. x. Chichester; New York: Wiley; 2002. p. 276.
20. Bland M. An introduction to medical statistics, vol. xviii. 4th ed. Oxford: Oxford University Press; 2015. p. 427.
21. Wang D, Bakhai A. Clinical trials : a practical guide to design, analysis, and reporting, vol. xiii. London: Remedica; 2006. p. 480.
22. Guluma K, Wilson MP, Hayden S. Doing research in emergency and acute care : making order out of chaos. Chichester, West Sussex; Hoboken: Wiley; 2015.

23. Hulley SB. Designing clinical research. 4th ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins; 2013.
24. Peat JK, Barton B. Medical statistics : a guide to SPSS, data analysis, and critical appraisal. 2nd ed. Chichester, West Sussex ; Hoboken: John Wiley & Sons Inc.; 2014.
25. Harris M, Taylor G. Medical statistics made easy 3, vol. xii. 3rd ed. Banbury: Scion; 2014. p. 116.
26. Hofmann AH. Scientific writing and communication. Papers, proposals, and presentations. 3rd ed. New York: Oxford University Press; 2017.
27. Campbell MJ, Walters SJ, Machin D. Medical statistics : a textbook for the health sciences, vol. xii. 4th ed. Chichester, Hoboken: Wiley; 2007. p. 331.
28. Hill T, Lewicki P. Statistics : methods and applications : a comprehensive reference for science, industry, and data mining, vol. xvi. Tulsa: StatSoft; 2006. p. 832.
29. Riegelman RK. Studying a study and testing a test : how to read the medical evidence, vol. vii. 5th ed. Philadelphia: Lippincott Williams & Wilkins; 2005. p. 403.
30. Rees DG. Essential statistics, vol. xiii. 2nd ed. London, New York: Chapman and Hall; 1989. p. 258.
31. Kuzma JW, Bohnenblust SE. Basic statistics for the health sciences, vol. xvii. 4th ed. Mountain View: Mayfield Pub. Co; 2001. p. 364.
32. Peat JK, Barton B, Elliott EJ. Statistics workbook for evidence-based healthcare, vol. viii. Malden: Blackwell; 2008. p. 182.
33. Altman DG. Practical statistics for medical research, vol. xii. Boca Raton: Chapman & Hall/CRC; 1999. p. 611.
34. Myles PGT. Statistical methods for Anaesthesia and intensive care. Edinburgh: Butterworth-Heinemann; 2000.
35. Rosner B. Fundamentals of biostatistics, vol. xix. 8th ed. Boston: Cengage Learning; 2016. p. 927.
36. Petrie A, Sabin C. Medical statistics at a glance. 3rd ed. Chichester, Hoboken: Wiley-Blackwell; 2009. p. 180.
37. Campbell MJ, Swinscow TDV. Statistics at square one, vol. iv. 11th ed. Chichester, Hoboken: Wiley-Blackwell/BMJ Books; 2009. p. 188.
38. Argyrous G. Statistics for social and Health Research. Great Britain: Sage Publications; 2000.
39. McCaig C, Dahlberg L. Practical research and evaluation : a start-to-finish guide for practitioners, vol. p.viii. London: SAGE; 2010. p. 263.
40. Daly LE, Bourke GJ, Bourke GJ. Interpretation and uses of medical statistics, vol. xiii. 5th ed. Oxford: Blackwell Science; 2000. p. 568.
41. Kirkwood BR, Sterne JAC, Kirkwood BR. Essential medical statistics, vol. x. 2nd ed. Malden: Blackwell Science; 2003. p. 501.
42. Le CT, Eberly LE. Introductory biostatistics, vol. xvii. 2nd ed. Hoboken, New Jersey: Wiley; 2016. p. 591.
43. McKenzie S. Vital statistics: an introduction to health science statistics. Chatswood: Churchill Livingstone.
44. Glantz SA. Primer of biostatistics. 7th ed. New York: McGraw-Hill Medical Pub. p. 2002.
45. Gosall NaG G. The doctor's guide to critical appraisal. 4th ed. UK: Pastest.
46. Glover T, Mitchell K. An introduction to biostatistics, vol. x. 3rd ed. Long Grove: McGraw-Hill; 2016. p. 487.
47. Hill AB. Principles of medical statistics. 12th ed. New York: Oxford University Press; 1989.

Publisher's Note