Too Good to be False: A Critique of Statistical Heaven

Chris H.J. Hartgerink1

Marcel A.L.M. Van Assen1

Jelte M. Wicherts1

1 Tilburg University, the Netherlands

WORD COUNT: XXXX words

Author note

Abstract

\*placeholder\*

*Keywords:* p-value, significance, underpowered, effect size, fisher

**Too Good to be False: A Critique of Statistical Heaven**

Statistical hypothesis testing[[1]](#footnote-2) is the crux of contemporary psychological science, and the validity of a study is regularly assessed by the results of individual hypothesis tests. There is much value in the concept of hypothesis testing, but the application of individual hypothesis tests is subject to unconscious psychological fallacies (i.e., law of small numbers; Tversky & Kahneman, 1971, confirmation bias; Nickerson, 1998, hindsight bias; Fischhoff, 1975, overconfidence; ) and tends to overlook the probabilistic nature of statistical results (Murdoch, Tsai, & Adcock, 2008). Additionally, the use of *only* the results of hypothesis tests has been severely critiqued (see Harlow, Mulaik, & Steiger, 1997). We are concerned these fallacies might cause a notion we call *statistical heaven*, where results from statistical hypotheses are interpreted as if error free. Before we outline how this notion arises, a short review of hypothesis testing is in place.

Scientific- and statistical hypothesis testing

One of the widely accepted tenets within the philosophy of science is that of falsification () via scientific hypothesis testing. Falsification requires all postulates to have the possibility of being disproven, to be considered scientific to begin with. Even though it is not *the* demarcation criterion (), it is one of the main drivers behind the hypothetico-deductive framework in the empirical sciences by use of scientific hypotheses tests. In this hypothetico-deductive framework researchers (1) observe, (2) form explanations and predictions, which are subsequently (3) tested and (4) results of these tests are fed back in to step 1 or 2.

Statistical hypothesis is a practical application of scientific hypothesis testing, but it is not *equal to* scientific hypothesis testing. In similar vein as predictions being deduced from hypotheses, which in themselves are deduced from theory, statistical hypothesis testing is deduced from the notion of scientific hypothesis testing. Just like a disconfirmed rejection cannot disprove an entire theory (only indicate some faults), a rejected statistical HT cannot disprove a scientific HT, only indicate there *might* be something wrong.

Statistical heaven

*Statistical heaven* is the notion that research findings in statistical hypothesis tests are unconsciously judged as more certain-, and error free than they actually are. In short, this means that researchers assess Type I and Type II error rates as being absent, implying that all significant results are true positives and all non-effects are true negatives. In the previous section, we indicated that true negatives are moot in the presence of high power and We do not intend this to sound asinine, and recognize that this discrepancy between the properties of hypothesis tests and practice is due to humans’ lack of intuition for random sequences and several unconscious fallacies.

The unconscious belief in the law of small numbers (Tversky & Kahneman, 1971) causes the notion that sample data are highly representative of population data, resulting in conclusions on the sample level that are readily generalized to the population level. However, population data are theoretical values that can never be directly measured — observed scores are a sum of the true score and the random measurement error (i.e., *X = T + E*; Novick, 1966). Although this belief becomes more justified as sample size increases, samples in psychological science are relatively small to the population that is being studied in most instances ().

Confirmation bias

Hindsight bias

Overconfidence

When using hypothesis tests as a tool to make inferences, there are four possible outcomes as a result of the absence/presence of an effect in the population, and absence/presence of an effect in the sample data (see Figure 1).

Although of much value in developing methods with calibrated specificity (i.e., Type I error equaling α), the null situation never arises in practical research settings as there is always some difference due to random noise, however minor. Subsequently, a difference due to random noise *can* be significant when the power is large enough. This is easily illustrated by simulating two samples (*nk* = 5,000,000), with equal mean and variance (*Mk* = 1, *sk =* .1), but with a miniscule amount of variation in random error (*e1 =* .002 and *e2* = .001 ~ *U* [0,*ek*]). This results in a 100% rejection rate of the null hypothesis of equal means over 1000 randomly generated sample pairs, whereas in truth the null hypothesis *is* true. In other words, the null can only be true if there is no measurement error at all — a thing we consider impossible. This is a hyperbolic example, meant to illustrate that the null canandwillbe determined to be false given enough power, purely on the basis of variation in measurement error.

Aligned with this, is the notion that our attention as a scientific enterprise, in general, should be on controlling the rate of false-negatives — not the rate of false-positives (Fiedler, Kutzner, & Krueger, 2012). Fiedler et al. (2012) argued this from a theory development perspective, positing that false-negatives cause forgoing of hypotheses and false-positives encourage further testing. Discovery rates of errors are proposed to be higher for false-positives due to such further testing, making false-negatives more costly. Our purely statistical argument reaches the same conclusion. However, we posit that null hypotheses are false to begin with given enough power, and therefore only false-negatives are a possibility, and hence, of interest.[[2]](#footnote-3)

The current paper is mainly concerned with the application of hypothesis tests and its subsequent conclusions with respect to the presence or absence of an effect. Although disputed by some (e.g., ), null hypotheses are theoretically *always* subject to rejection, given a large enough sample and/or precise measure (i.e., power). In other words, null hypotheses are moot, as they will always prove to be rejected in some instance.

The unconscious human fallacies and the probabilistic nature of statistical results are at interplay, resulting in a fallacy of its own accord. In the current paper, we explain the origins of this fallacy, which we call the fallacy of *statistical heaven*. Subsequently, we investigate how this fallacy of statistical heaven has affected research in psychological science, and propose a method to decrease itsThese unconscious fallacies make the interpretation of results arduous and can lead to seemingly paradoxical conclusions. Additionally, hypothesis testing has intricacies that are often misconceived (see Schmidt & Hunter, 1997), which adds . In the current paper, we propose a method that

One of the major problems is the dichotomization of results: either there is an effect, or there is not. In light of Popper’s () hypothetico-deductivism, such a dichotomy seemingly makes sense, as Popper’s demarcation criterion for science is falsification. Hypothesis testing embodies this demarcation criterion, and is oft seen as a way of attaining objectivity of results (). This could be true, were it not for the probabilistic nature of statistics.

This probabilistic nature makes the reasoning underpinning falsification inherently problematic, as Cohen () aptly illustrated. Take for example the following syllogism as taken from Cohen ():

“If a person is an American, then he is not a member of Congress. (WRONG!)

This person is a member of Congress.

Therefore, he is not an American.” (p. 23, Cohen, 1997)

This example shows correct logic, but an incorrect premise — the conclusion, given the premise, is correct however. If we subsequently correct the premise, and make it probabilistic, we attain the following:

“If a person is an American, then he is probably not a member of Congress (TRUE, RIGHT?)

This person is a member of Congress.

Therefore, he is probably not an American.” (p. 24, Cohen, 1997)

The premise is now correct, but the resulting conclusion does not make logical sense. In fact, this conclusion is formally the same as saying, if the data are unlikely under the null-hypothesis, the null is probably not true.

These explorations of the logic of hypothesis testing indicate that falsification can serve as a demarcation criterion for science, but that the probabilistic implementation of it as done by hypothesis testing does not hold. This provides a serious problem for the idea of *hypothesis testing as objective criterion.* But why does hypothesis testing prevail despite of these strong, logical counterarguments?

We propose the notion of *statistical heaven* to be the main driver behind this prevailing of hypothesis testing, and more importantly, the firm persistent belief in the value of hypothesis testing. Statistical heaven is the place where research makes no errors, and has a 100% chance of finding an effect if there actually is one (i.e., α = 0; 1-β = 1). This notion easily arises, as research has repeatedly indicated that humans lack randomness of mind, i.e., they have bad intuition for probabilities. Take for example the persistent belief in the law of small numbers (), the Monte Hall problem (), the conjunction fallacy () or simply the generation of random data (). Power and Type I error rates (or other results for that matter) remain probabilities, and therefore are subject to that same bad intuition. Since researchers fail to take into account these random elements to hypothesis testing, they resort to thinking their results indicate certainties: there is an effect, or there is none.

Seeing how this ignorance of the probabilistic elements of hypothesis testing reduces findings from probabilistic to deterministic, it is easy to see why hypothesis testing prevails. The probabilistic syllogism, as illustrated by Cohen (), was clearly false. However, if we remove the probabilistic element, the logic reduces to that which makes full sense. In other words: probabilistic hypothesis testing makes no sense, as it contains flawed logic, but deterministic hypothesis testing makes very much sense. Too bad hypothesis testing is never deterministic and will remain probabilistic, no matter how much we will it otherwise.

* Short history on the debate in psychological science
  + Differentiate between theoretical context (content that is being translated to hypotheses) and statistical context (the results of the hypotheses being tested)
  + Significance testing and philosophy of hypothetico-deductivism and objectivism
    - NHST as embodiment of the falsification criterion
    - Effect/no effect as pseudo-objectivism
* Why a more optimistic view can be taken of non-significant studies
  + Statistical heaven
    - How it arises
      * Confirmation bias (Rosenthal, 1964)
      * Belief in the law of small numbers (Tversky & Kahnemann)
    - Researchers are human and lack randomness of mind — statistics uses probability and if humans cannot see such probabilities properly (Wason task; Monte Hall problem) than they should not be allowed to filter the probabilistic resul
  + Why it is incorrect
* Talk about α and β error control (Fiedler 2012)
  + We focus on β error control and try to show that α error control is non-vital in practical research (very important in theoretical research on methods, i.e., specificity of tests)
    - More specific hypotheses
* Lack of power in psychological science throughout the last decades

Method

Results

Discussion

* Recommendations

1. Hypothesis testing will be referred to as HT in the remainder of the text. [↑](#footnote-ref-2)
2. Note that this superceding of false-negatives over false-positives works only under the assumption of proper research, i.e., lacking QRPs etc. [↑](#footnote-ref-3)