Too Good to be False: A Critique of Statistical Heaven

Chris H.J. Hartgerink1

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 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; , fundamental attribution error; ) 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; Wilkinson & APA Task Force on Statistical Inference, 1999). We are concerned these fallacies might cause a notion we call *statistical heaven*, where results from statistical hypotheses are interpreted as if free from sampling error. We regard that, given that a study is methodologically sound, there is only one error it can make: a false negative. We outline this idea by first reviewing hypothesis testing, subsequently showing plausible psychological precedence for the notion of statistical heaven, and lastly introducing a method to indicate that null-effects suffer from a lack of power.

Scientific- and statistical hypothesis testing

One of the widely accepted tenets within the philosophy of science is that of falsification of scientific hypotheses (Popper, 1959). 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 (Maxwell, 1972), it is one of the main drivers behind the hypothetico-deductive framework in the empirical sciences where researchers (1) observe, (2) form explanations and predictions, which are subsequently (3) tested and (4) results are subsequently used to form new predictions.

Statistical hypothesis testing 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, the notion of statistical hypothesis testing is deduced from the notion of scientific hypothesis testing. And, just like a disconfirmed prediction does not disprove an entire theory, as it is a specification of the theory, a rejected *statistical* hypothesis test is not equal to a *scientific* hypothesis test.[[1]](#footnote-2)

The formal logic of falsification in *scientific* hypothesis testing does not fully transfer to *statistical* hypothesis testing. Scientific hypothesis testing is as follows (framed in null hypotheses/H0 and alternative hypotheses/H1; example adapted from Cohen, 1994):

*If H0, then not H1.*

*We observe H1.*

*Hence, not H0.*

The premise, the observation and the conclusion logically follow each other and therefore form a consistent reasoning. However, if we introduce the probabilistic elements that are also introduced in *statistical* hypothesis testing, the logic becomes more ambiguous:

*If H0, then probably not H1.*

*We observe H1.*

*Hence, probably not H0.*

The premise, and the observation are correct in this situation, but the conclusion is not. This conclusion states something about the probability of H0, but based on a statistical hypothesis test alone we cannot infer this. After all, the resulting *p*-value only indicates the probability of the sample data, *given* the null hypothesis (i.e., *P[D|H0]*). To come to a conclusion about the probability of H0, Bayes’ theorem[[2]](#footnote-3) would have to be used and a prior would be needed on the probability of the null hypothesis. Additionally, whereas the conclusion is dichotomous in the *scientific* hypothesis testing situation, it is not so in the *statistical*, as it is uncertain. Statistical hypothesis testing embodies the concept of scientific hypothesis testing, however, and because of this results are dichotomized in regular practice.

Such dichotomization of the probabilistic elements in statistical hypothesis testing give rise to the possibility of errors in the inferences that are made.

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 unconsciously assess Type I and Type II error rates as null, implying that all significant results are true positives and all non-effects are true negatives. 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’ general lack of intuition for random sequences and several unconscious fallacies.

Law of small numbers. 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 is more justified as sample size increases, samples in psychological science are often relatively small relative to the population that is being studied (i.e., *Nmedian* = 40; Marszalek, Barber, Kohlhart, & Holmes, 2011).

**Confirmation bias.**

**Hindsight bias.**

**Overconfidence.** Overconfidence leads to overestimation of probabilities that one is correct and is defined as the discrepancy between confidence and accuracy. It is most clearly illustrated by the finding that 80% of people think they belong to the 50% smartest people (Dunning & Kruger), which clearly is impossible. Overconfidence is more

**Fundamental attribution error.**

**False positives versus false negatives**

False positive rates have been widely debated throughout the last decade (Ioannidis, 2005; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), and some have called for more focal attention on false negative rates (Fiedler, Kutzner, & Krueger, 2012).

If we consider the nature of the false positives debate (i.e., α-control), and the nature of the false negatives debate (i.e., β-control), it can be determined that these debates are not parallel, but substantively different. These debates overlap, as α and β are negatively correlated (i.e., higher α criterion leads to more power), but the debates have seemingly not been connected — only separated (Fiedler et al., 2012).

The debate on false positives has focused on proper methodology. Recent research has indicated that undisclosed research practices inflate false positive rates (Simmons et al., 2011), and that such practices have become widespread (John et al., 2012). Typically, α-control was important when developing new statistical methods (i.e., calibration), but this research has shown that we should be worried more about human factors in α-control. The debate has since then carried onward, discussing how to improve disclosing of research practices (LeBel et al., 2013), and has stimulated further research to show what practices increase false positive rates (e.g., Murayama, Pekrun, & Fiedler, 2013).

The debate on false negatives has focused on sensitivity of data and tools to find effects (i.e., power in studies). Although not as lively as the debate on false positives, the false negatives debate is just as important, as we will see shortly. Fiedler et al. (2012) rightly argued that false negatives *are* costly, as finding ‘nothing’ causes research lines to be stopped, thinking there is no effect present whatsoever. In practice, there is *always* a statistical effect present () — it is much more important to ask how large this effect is instead of whether there is an effect. After all, as sample size increases, power increases, keeping other factors constant. Given that there is variable measurement error in all measurements (Novick, 1966), parameters of interest will never be exactly the same. Hence, any difference will be significant given a large enough sample. Considering that the power in a typical psychological experiment has been estimated at 35% (Bakker, Van Dijk, & Wicherts, 2012), this means there is a false negative rate of 65%. If possibly valuable research lines are being stopped at a 65% rate, due to being underpowered, we are neglecting valuable information and wasting valuable resources — especially if these are not published (van Assen, van Aert, Nuijten, & Wicherts, 2014).

The common ground between these two debates is substantial, as studies barely failing to achieve significant results can *lead to* research practices that inflate false positive rates.

This indicates that a possibility to (partially) solve the inflated false positive rate and inflated false negative rate is in increasing the power of statistical hypothesis tests.

As a science, we can then stop asking whether there is an effect or not, but ask ourselves how large the effect is and with what precision we can estimate it.

**(Statistically) Non-significant psychological effects**

**Data summary**

**Observed effects**

We included all XXXX non-significant *t*, *F*,and *r* values (assuming α = .05 for all results), from which η2 effect sizes were computed (equal to *R*2 in interpretation). These test statistics were selected because they are, technically, all *F*-values. A squared *t-*value equals an *F-*value, and an *r-*valuecan be computed into a *t-*value, and subsequently an *F*-value. The dataset also includes χ2-values, but these are not readily computed into an η2 effect size, whereas the other included test values (i.e., *Z-*values, Wald-values) do not allow for computing exact effect sizes.

If statistical null results are interpreted as null effects, the observed non-significant effects should clearly follow the theoretically predicted effect null distribution, if this non-significance is not just an artifact from not enough power. Under a null distribution, *p*-values are distributed uniformly ().

Fisher’s inexact test

, where and

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

Method

Results

Discussion

* Recommendations

References

Footnotes

Table 1

|  |  |  |
| --- | --- | --- |
|  | H0 | H1 |
| ‘H0’ | 1-α | β |
|  | *True negative* | *Type II* |
| ‘H1’ | α | 1-β |
|  | *Type I* | *True positive* |

*Note.* Columns indicate the true situation in the population, rows indicate the statistical conclusion based on sample data.

*Figure 1*

Observed effects versus simulated null effects.

1. This becomes especially confusing due to R.A. Fisher, who equated statistical hypothesis testing with scientific hypothesis testing. [↑](#footnote-ref-2)
2. [↑](#footnote-ref-3)