Too Good to be False: A Critique of Statistical Heaven

Chris H.J. Hartgerink1

1 Tilburg University, the Netherlands

WORD COUNT: XXXX

Author note

This paper is version controlled and all research files are publicly available at XXXX.

Abstract

\*placeholder\*

*Keywords:* p-value, nhst, 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; ) 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 more certain- and error free than they actually are. We regard that, given that a study is methodologically sound, there is only one error it can make: a false negative. In the current paper, this idea is outlined by reviewing hypothesis testing, 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 (Schmidt & Hunter, 1997, p. 42). 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 does not disprove a *scientific* hypothesis per se.[[1]](#footnote-2)

Additionally, 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 this cannot be inferred. 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 (i.e., *P[H0|D]*), 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 necessarily so in the *statistical*, as it contains uncertainty. Statistical hypothesis testing embodies the concept of scientific hypothesis testing, however, and because of this, results are regularly dichotomized in 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 (almost) 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 several unconscious fallacies and humans’ general lack of intuition for random sequences.

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). Unconsciously believing that sample data are representative of population data can lead to underestimating the variability that underlie the results, and subsequently overestimating the power of a study.

**Overconfidence.** Overconfidence is the overestimation of probabilities that one is correct and is defined as the discrepancy between degree of confidence and degree of accuracy. For psychology, such overconfidence with respect to power is clearly illustrated by Bakker (2014). She found a discrepancy ranging between .29 and .42 for power estimations and power calculations, where positive numbers indicate that researchers estimate power to be higher than it actually is in a study. In other words, systematic overestimation of power occurs, and thus also systematic underestimation of the false negative error rate. For statistical hypothesis testing, this leads to overconfidence of an effect that is found, to be assessed as true.

**Confirmation bias.** Confirmation bias causes an asymmetry in the evaluation of evidence: data that confirms a preconceived notion is more readily accepted as true. With respect to statistical hypothesis testing, this

**Hindsight bias.** Hindsight bias is often typified as the *knew-it-all-along* effect (Christensen-Szalanski & Willham, 1991; Fischhoff, 1975), where results seem obvious only after they occur. Considering results from statistical hypothesis testing, this might cause that

**Probabilistic results.** Statistical results are themselves subject to probabilities, and these probabilities cannot be assessed with the results from a single hypothesis test. This is similar to assessing the fairness of a coin based on a single flip. Subjectivists with respect to probabilities (), might still proclaim this is possible, but the fundaments of statistical hypothesis testing are rooted in the frequentist view of probabilities ().

**False positives versus false negatives**

False positive rates have become 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 the nature of 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 (e.g., 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 also be worried about human factors when inspecting α-control. The debate has since carried onward, discussing how to improve disclosing of research practices (LeBel et al., 2013), and what practices increase or decrease 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 (McDonald, 1997; Rossi, 1997; Steiger & Fouladi, 1997), hence, it is much more important to ask how large an effect is instead of whether there is an effect at all (). After all, as sample size increases, power increases, keeping other factors constant. Given that there is variable measurement error in all measurements (i.e., *X = T + E*; Novick, 1966), parameters of interest will never be exactly the same. Therefore, 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 research lines 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**

**(Journal) Effect distributions**

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 as these 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 null effect distribution, if this non-significance is not just an artifact from lack of power. Under a null distribution, *p*-values are distributed uniformly (Murdoch et al., 2008; Sackrowitz & Samuel-Cahn, 1999). Randomly sampling uniform *p-*values (XXXX times) simulates the null distribution. Subsequently, these generated *p-*values were used to recompute accompanying test-values, which were used to compute effect sizes under the null distribution (see Figure 1; grey line). Observed effects are plotted against this theoretical null distribution.

The difference between the theoretical null distribution and the observed distribution was tested with a one-sided Kolmogorov-Smirnov test, which clearly indicated a difference between the effect distributions. The Kolmogorov-Smirnov comparison tests whether two distributions differ from each other, and we use a one-sided test as the computed effect size metric (i.e., η2) has a state space of [0; 1]. Results indicated that the observed distribution significantly deviates from the null distribution; INSERTRESULTS. This shows that, considering all non-significant *p-*values, there is indication for an effect nonetheless. Of those effects, XX% can be considered small, XX% can be considered medium, and XX% can be considered large ().

However, all things considered, this only indicates that, across all papers, there is evidence for an effect in the non-significant results. This shows that the notion that non-significant results equal null effects is inappropriate, as there is a significant effect *hidden* in these results. To further develop our idea, we expand our research below on a paper level.

**Paper level results**

In the current section, an easy to use method is presented to test for presence of an effect across a set of statistical hypothesis tests, requiring only *p-*values. Subsequently, power simulations are presented and descriptive results are given for papers included in the XXXX dataset.

**Fisher’s inexact test.** Fisher’s test () was originally used to test for deviations from uniformity, which is also applicable for *p-*value distributions.As mentioned previously, if the null hypothesis is true, *p*-values are uniformly distributed — this makes the Fisher test an excellent test for inspecting the presence of an effect. However, as only non-significant results are of interest, and the test is based on *p-*values ranging from 0 to 1, the non-significant results need to be transformed to fit the same range. *P*-values are therefore transformed in accordance with the formula

which ensures similar properties for the selected, non-significant *p*-values. *P*-values smaller than α result in *p\**-values that are smaller than zero, which cannot be used in the subsequent tests. Larger than zero therefore indicates non-significant *p\**-values.

Two instances of this test are used: *F* and its complement *F—*. These are

and

where *k* is the number of non-significant *p-*values. Two tests are required, because the logarithmic transformation causes an asymmetry between large and small *p\**-values. Small *p\**-values result in large values after transformation (e.g., ) and large *p\**-values result in small values after transformation (e.g., ). The two tests are therefore also asymmetrical: *F* tests whether values close to α deviate from uniformity, whereas *F—* tests whether values close to 1 deviate from uniformity.

This test statistic is gamma distributed; *F* and *F—* ~ Γ(1, *k*).This gamma distribution is a consequence of the composition of the *F* and *F—* statistics. As the *p-*values are uniformly distributed under the null hypothesis, the natural logarithm of these *p*-values results in an exponential distribution (). Taking the sum over these exponentially distributed *p-*values results in a gamma distribution (), with rate parameter 1, and the shape parameter equaling the number of instances over which is summed (i.e., *k*; REF).

**Fisher test power.** To simulate the power of the Fisher test, the original test results in the dataset were bootstrapped. Sixteen hypothetical population effect sizes were imposed, from which non-significant results were simulated for all *t*, *F* and *r* test statistics in the original dataset. Per set of test statistics within a paper, 1000 iterations were run, where each iteration yielded a result for both Fisher tests per paper. Power was computed as the proportion of significant Fisher tests over the 1000 iterations.

For each test statistic, six steps were necessary to simulate the non-significant *p*-value. First, a critical value under the null distribution was needed. Based on the degrees of freedom in the dataset, this was easily calculated. Second, these same degrees of freedom were used to compute the non-centrality parameter, to determine the population distribution based on the imposed effect size. The non-centrality parameter (i.e., δ) was computed as

where

*N* was determined on the basis of the degrees of freedom (i.e., ; ). Third, the area under the curve of the population distribution was determined, where the result would yield a non-significant result (i.e., β). Fourth, a value was uniformly drawn between 0 and the β value that resulted from step three. Fifth, the accompanying test-value was computed, which was, sixth, used to compute the *p­*-value under the null distribution.

Simulation results indicate the power of both tests is highly sufficient, even when only few non-significant results are presented. Figure 2 visually summarizes the results. It is clear the power rapidly increases as a function of effect size

**Descriptive results.** Of the original test results in the dataset, Fisher tests showed substantial indication for underpowered results. Of the XXXX papers in the dataset, XXXX showed significant deviation from uniform *p-*values.

**Applying to a ‘failed’ research line**

**Discussion**

In recent years, concern has been voiced as to the state of psychological science and its practices — it has even been postulated that some findings are *too good to be true* (Francis, 2012). In the current paper, we shed light on the findings that are seemingly neglected: the non-significant results. We say these are *too good to be false*: these non-significant results still show signs of an effect, and should therefore not be neglected.

**Recommendations**

References

Footnotes

Table 1

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

*Note.* Columns indicate the true situation in the population, rows indicate the statistical conclusion based on sample data. The true positive rate is also called power, and the true negative rate is also called XXXX.

Table 2

*Note.* Alpha was assumed to be .05 to determine significant or not.

*Figure 1*

Observed effects versus simulated null effects.

*Figure 2*

Visual depiction of power simulations of Fisher tests. Thick line indicates the

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)