Power and Negative Results

Edouard Machery*†

The use of power to infer null hypotheses from negative results has recently come under severe attack. In this article, I show that the power of a test can justify accepting the null hypothesis. This argument also gives us a new powerful reason for not treating *p*-values and power as measures of the strength of evidence.

1. Introduction. What rules should govern the inferences of null hypotheses from negative results? Following Cohen's influential work, psychologists sometimes rely on power considerations for this purpose. Recently, however, the use of power to infer null hypotheses from negative results has come under severe attack. In particular, Hoenig and Heisey (2001, 1) have discussed the "large literature advocating that power calculations be made whenever one performs a statistical test of a hypothesis and one obtains a statistically nonsignificant result," and they have concluded that "this approach, which appears in various forms, is fundamentally flawed." Finally, they have recommended replacing the use of power with other inferential practices such as equivalence testing. Hoenig and Heisey's argument has been influential, and it has led to some debate among applied statisticians, psychologists, and other scientists about the role of power for the inference of null hypotheses (e.g., Lenth 2007; Leventhal 2009).

In this article, I defend the use of power to infer null hypotheses from negative results: I show that Hoenig and Heisey's argument fails and that the power of a test can justify accepting the null hypothesis, as Cohen proposed. Furthermore, the rebuttal of Hoenig and Heisey's argument gives us a new powerful reason for not treating *p*-values and power as measures of the strength of evidence.

*To contact the author, please write to: Department of History and Philosophy of Science, University of Pittsburgh, 1017CL, Pittsburgh, PA 15260; e-mail: machery@pitt.edu.

†I would like to thank Greg Gandenberger for his comments on a previous version of this article.

Philosophy of Science, 79 (December 2012) pp. 808–820. 0031-8248/2012/7905-0023$10.00

https://doi.org/10.1086/667877 Published online by Cambridge University Press

In section 2, I briefly review the norm that, according to Cohen, should govern the inference of null hypotheses from negative results. In section 3, I describe Hoenig and Heisey's argument. In section 4, I rebut their argument. In section 5, I briefly respond to some possible objections.

2. Negative Results and Null Hypotheses.

2.1. For Memory. I call "alternative hypothesis" (H_A) the statistical hypothesis that is experimentally tested and that is accepted when the null hypothesis is rejected. The significance level (a.k.a. α level) is the largest value a p-value may have if the null hypothesis (H_0) is to be rejected. It is often set at .05 in experimental psychology. A negative result is any experimental result that does not permit the rejection of the null hypothesis because the p-value of the relevant statistic is larger than the significance level.

The power of a test is the probability of rejecting the null hypothesis if the null hypothesis is false and if a particular point alternative hypothesis is true.[1] Thus, 1 minus the power of a test is the probability that this test will commit a type II error when this point alternative hypothesis is true (failing to accept the alternative hypothesis when it is true—this probability is sometimes called β). The power of a test is influenced by the population effect size, the significance level, and the sample size. The larger the sample size, the higher the power of a test. The higher the significance level, the higher the power of a test. The larger the population effect size, the higher the power of a test.

Statisticians distinguish between observed power and prospective power. Power is said to be "observed" when one uses various descriptive statistics (e.g., the pooled sample variance or the difference between the sample means) to estimate the corresponding parameters of the populations from which the experimental samples are drawn. For instance, one could estimate the variances of the populations from which the samples are drawn by means of the variances of the samples themselves. One then computes the probability of rejecting the null hypothesis if samples of a particular size are drawn from these populations, assuming a particular significance level. By contrast, power is said to be "prospective" when one stipulates, instead of estimating, the values for the parameters of the populations from which the relevant samples are drawn. One then computes the probability of rejecting the null hypothesis that would obtain if the samples were drawn from populations with these parameter values. Prospective power is regularly used in planning experiments (but see Sedlmeier and Gigerenzer 1989).

[1] In psychology, this point alternative hypothesis almost always differs from the alternative hypothesis that is experimentally tested since the latter is typically a composite hypothesis (e.g., that the mean of a population is different from 0). Various considerations determine which point alternative hypothesis is selected for computing the power of a test.

https://doi.org/10.1086/667877 Published online by Cambridge University Press

2.2. Negative Results, Null Hypotheses, and Power. Jacob Cohen proposed that a null hypothesis can be inferred from a negative result if and only if the test in which a negative result was obtained has a high power—that is, if and only if the test was such that the null hypothesis would have been rejected if it were false. More precisely, he writes (1988, 16):

> What is really intended by the invalid affirmation of a null hypothesis is not that the population ES [effect size] is literally zero, but rather that it is negligible, or trivial. This proposition may be validly asserted under certain circumstances. . . . If the research is performed with this n [i.e., the sample size needed to have power $1-\beta$] and it results in nonsignificance, it is proper to conclude that the population ES is no more than i, i.e., that it is negligible; this conclusion can be offered as significant at the β level. . . . Thus in using the same logic as that with which we reject the null hypothesis with risk equal to α, the null hypothesis can be accepted in preference to that which holds that $\text{ES}=i$ with risk equal to β.

Thus, Cohen is proposing the following inferential procedure:

Power-Based Inferential Procedure

1. To test psychological theory T, a hypothesis (H_A) is derived (e.g., $\mu_1-\mu_2\neq 0$).
2. A null hypothesis (H_0) is formulated (e.g., $\mu_1-\mu_2=0$).
3. An experiment is designed that has the power $1-\beta$ to detect an effect of size i, where i is small.[2]
4. If the p-value of the statistic computed from the data is above the significance level and if the power of the experiment for an effect size i is above some threshold t (where t is high; e.g., $1-\beta>.95$), then one rejects H_A and one accepts an approximation of H_0 (e.g., $\mu_1-\mu_2$ is equal to 0 or nearly so).[3]

The justification of this procedure is straightforward. Cohen notes that there is a parallelism between null hypothesis significance testing and the procedure just described. In the case of null hypothesis significance testing, one decides to reject the null hypothesis and to accept the alternative hypothesis

[2] Thus, in Cohen's approach, the point alternative hypothesis used in computing power corresponds to the smallest discrepancy from the null hypothesis that is of interest.

[3] Thus, strictly speaking, when one follows this procedure, one accepts an approximation of H_0 rather than H_0 itself. For convenience, I will not make this distinction in what follows.

https://doi.org/10.1086/667877 Published online by Cambridge University Press

when the *p*-value is smaller than the significance level. By doing so, one takes the risk of committing a type I error, but this risk is small, and in the long run one will not make more than a small number of type I errors if one follows this procedure consistently. In the case of the power-based inferential procedure, one decides to accept the null hypothesis and to reject the alternative hypothesis when and only when the power of the test is larger than a high threshold (say, .95).[4] By doing so, one takes the risk of committing a type II error, but this risk is small, and in the long run one will not make more than a small number of type II errors if one follows this procedure consistently.[5] If one takes null hypothesis significance testing to be justified because of its property of controlling long-run error rates, then by parity of reasoning one should hold the power-based inferential procedure to be justified too.

2.3. Examples. Psychologists rarely publish negative results, and as a consequence they have few occasions to infer the null hypothesis on the basis of the power of a test. Even in those relatively rare articles that report negative results, psychologists too often satisfy themselves with an informal analysis. For instance, Hare and colleagues (2007) compared chimpanzees' and bonobos' willingness to collaborate to obtain some food (experiment 2). Participants (either two chimpanzees or two bonobos) were paired; each participant had to pull a rope at the same time as the other participant in order to bring food toward itself. When the food was dispersed in a way that allowed each participant to benefit from the product of its contribution to the collaborative effort (i.e., some food) on its own, chimpanzees and bonobos collaborated equally. Inferring the null hypothesis from a negative result, Hare and colleagues write that "in support of the emotional-reactivity hypothesis, there was no difference between the species' ability to cooperate spontaneously to obtain divisible-dispersed food" (620). Power is not reported either in the main article or in the online supplementary data. No alternative formal analysis is presented.

Some articles do appeal to power to infer the null hypothesis from a negative result, although the power-based inferential procedure is not always rigorously applied. In his report of failed replications of Boroditsky's (2001) famous study about the cultural diversity of the mental representa-

[4] This threshold is often set a .8, but it is hard to see how one could justify the asymmetry with the usual value of the significance level if power is to be used for inferential purposes.

[5] Here, committing a type II error means failing to reject the null hypothesis or an approximation thereof when both are false.

https://doi.org/10.1086/667877 Published online by Cambridge University Press

tion of time, Chen (2007) estimated the population effect size by means of the effect size found in the study and, on this basis, computed the sample size needed to reach a power of .80, assuming a significance level of .05. While taking power into account is commendable, this analysis does not follow the power-based inferential procedure very closely. In another report about Boroditsky (2001), January and Kako (2007) also mentioned power, but they too failed to follow the power-based inferential procedure closely.

3. Hoenig and Heisey's Argument. Following Hoenig and Heisey, I examine observed power and prospective power successively. I also assume, without loss of generality, that the negative result is obtained in a one-sample *t*-test (testing whether the mean of a [by assumption] normally distributed population with unknown variance differs from a particular value—here, zero).

3.1. Observed Power 1. There are many ways to compute the observed power of an experiment, depending on the number of relevant population parameters that are estimated by means of the sample statistics. One could first estimate all the relevant population parameters. For instance, for a *t*-test, one could estimate the mean and variance of the population by means of the mean and variance of the sample drawn from it.

The problem with this approach is that, so computed, observed power is uninformative since there is then a one-to-one relationship between *p*-values and power. As a result, knowing the power of an experiment does not add any information to knowing the *p*-value. If one would not infer the null hypothesis from a negative result on the basis of the *p*-value of the relevant statistic alone (and, for obvious reasons, one should not do this), one should not infer from the observed power, so computed, either.

3.2. Observed Power 2. Of course, one need not estimate all the relevant population parameters by means of the corresponding sample statistics. Instead, one could estimate some population parameters (e.g., the variance of the population from which the sample in a *t*-test is drawn) by means of the sample statistics and stipulate the values of the others (e.g., the mean of this population) in order to compute the power of the test. In this case, there is not a one-to-one relation between the power of the test and the obtained *p*-value, and the problem just discussed disappears.

However, so computed, observed power gives rise to the following paradox. Experiments E_1 and E_2 are two (one-sample) experiments with the following properties. The sample sizes are the same ($N_1 = N_2$); the means observed in E_1, m_1, and E_2, m_2, are equal ($m_1 = m_2$); and the standard deviation obtained in E_1, s_1, is smaller than the standard deviation obtained

https://doi.org/10.1086/667877 Published online by Cambridge University Press

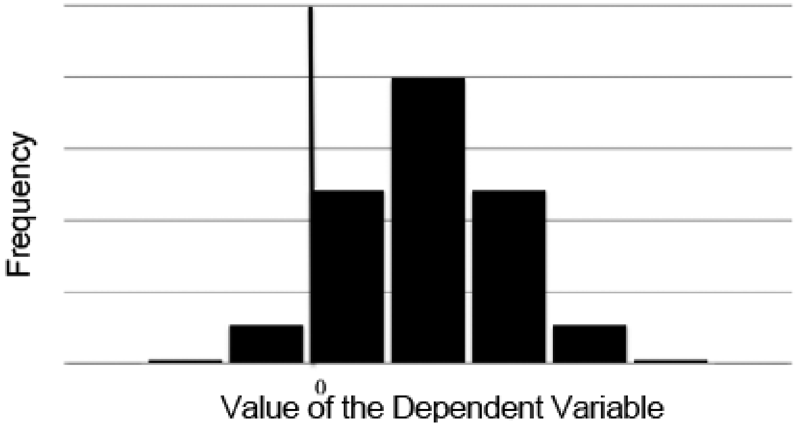

Figure 1. Histogram for E_1.

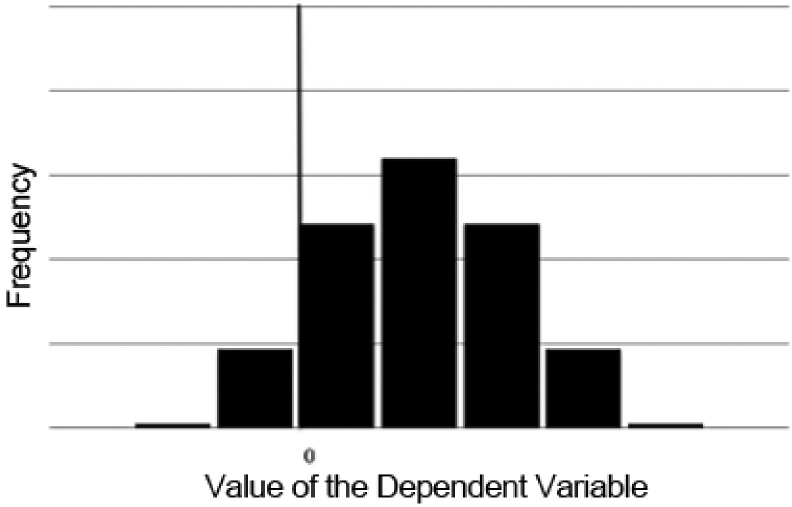

Figure 2. Histogram for E_2.

in E_2, s_2 ($s_1 < s_2$). Furthermore, both E_1 and E_2 yield nonsignificant findings, that is, negative results.

Now, because $t = m/(s/\sqrt{N})$ for a one-sample t-test, t is larger in E_1 than in E_2. Hence, the probability of obtaining a statistic of this size or a larger one if the null hypothesis is true is larger in E_1 than in E_2. That is, because $s_1 < s_2$, the p-value is lower in E_1 than in E_2 ($p_1 < p_2$). This point can be made intuitive as follows. Consider, for example, the two histograms in figures 1 and 2

https://doi.org/10.1086/667877 Published online by Cambridge University Press

corresponding to E_1 and E_2, respectively. The mean, mode, and median are the same in these two histograms, but the data are more spread out in E_2 than in E_1. Because the variance in E_2 is larger than the variance in E_1, the fact that the sample mean differs from 0 is less likely to be due to sampling in E_1 than in E_2. Thus, the sample in E_2 is more plausibly drawn from a population whose mean is 0 than is the sample in E_1.

Consider now the observed power of E_1 and E_2. Suppose that the null hypothesis is false and that the populations from which the two samples are drawn have the same mean, which is larger than 0 ($\mu_1 = \mu_2 > 0$). Variables s_1 and s_2 are used to estimate the standard deviations of these populations. For any stipulated population mean, E_1 is more likely to reject the null hypothesis than is E_2 since less of the mass is close to zero in the case of E_1. Figures 3 and 4 illustrate this point. Figure 3 represents the population from which the sample in E_1 is drawn, and figure 4, the population from which the sample in E_2 is drawn. Because a smaller part of the area under the curve is close to 0 in figure 3 than in figure 4, one is more likely to reject the null hypothesis by sampling randomly from the population represented in figure 3 than from the population represented in figure 4. That is, the power of E_1 is larger than the power of E_2.

We end up with a paradox. Because $p_1 < p_2$, E_1 provides more evidence against the null hypothesis and for the alternative hypothesis than E_2. Because the power of E_1 is larger than the power of E_2, E_1 also provides more evidence for the null hypothesis and against the alternative hypothesis than E_2.

3.3. Prospective Power. It may be tempting to suggest that the paradox results from the use of observed power rather than prospective power to infer the null hypothesis from a negative result. This would be a mistake, however: it is simple to derive a paradox similar to the one just presented when one uses prospective power instead of observed power.

Experiments E_1 and E_2 are two (one-sample) experiments such that $N_1 = N_2$, $m_1 = m_2$, and $s_1 < s_2$. Furthermore, both E_1 and E_2 yield negative results. To determine the power of E_1 and E_2, we stipulate that the two samples are drawn from the same population, whose mean is different from 0. Because the samples are drawn from an identical population and because the sample sizes are the same, the prospective power of E_1 and E_2 is the same. As a result, they provide the same amount of evidence for the null hypothesis and against the alternative hypothesis. However, because $s_1 < s_2$, the p-value is smaller in E_1 than in E_2, and E_1 provides more evidence against the null hypothesis and for the alternative hypothesis than E_2.

3.4. Conclusion. However power is computed, some pairs of experiments are such that, if it is appropriate to use power to infer the null hypothesis from a negative result, the first experiment seems to provide more

https://doi.org/10.1086/667877 Published online by Cambridge University Press

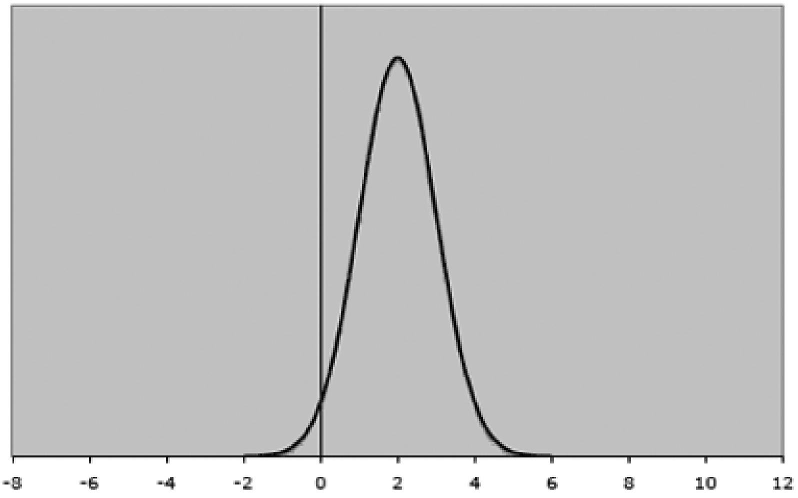

Figure 3. Population in E_1.

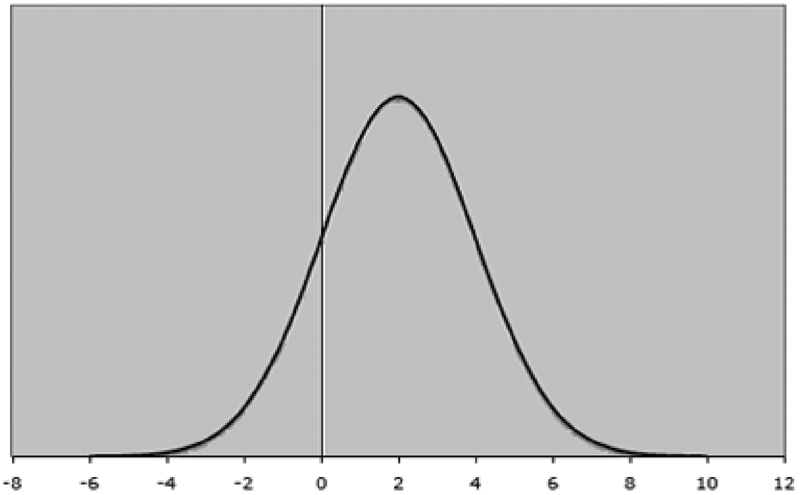

Figure 4. Population in E_2.

evidence both for and against the null hypothesis than the second experiment (or more evidence against the null hypothesis than the second experiment and the same amount of evidence for it as the second experiment). It seems natural to conclude that it is inappropriate to use power to infer the null hypothesis from a negative result. This is indeed the moral that Hoenig and

https://doi.org/10.1086/667877 Published online by Cambridge University Press

Heisey draw (2001, 5): “Power calculations tell us how well we might be able to characterize nature in the future given a particular state and statistical study design, but they cannot use information in the data to tell us about the likely states of nature. With traditional frequentist statistics, this is best achieved with confidence intervals, appropriate choices of null hypotheses, and equivalence testing.”

4. In Defense of Power-Based Inferences. In the remainder of this article, I defend the use of power to infer null hypotheses from negative results against the argument presented in section 3.

4.1. Two Interpretations of p-Values. The *p*-values can be understood in two different ways. First, one can assume that a *p*-value measures the strength of evidence against the null hypothesis. The smaller the *p*-value, the more evidence an experiment provides against the null hypothesis and for the alternative hypothesis. If the *p*-value in a first experiment is .005 and the *p*-value in a second experiment is .01, then the first experiment provides more evidence against the null hypothesis than the second. For Kempthorne and Folks (1971, 314), for instance, a *p*-value is “a way of producing a quantification of strength of evidence.”

Alternatively, one can assume that a *p*-value is a property of the data that allows for the application of a decision rule specifying when the null hypothesis is to be rejected and the alternative hypothesis accepted: reject the null hypothesis and accept the alternative hypothesis when and only when the probability of observing a statistic of this size or a more extreme one (e.g., a *t*-score) is lower than the significance level (say, .05). Under this interpretation, *p*-values are not taken to measure the strength of evidence against the null hypothesis. One does not treat the fact that the *p*-value in a first experiment is .005 and the *p*-value in a second experiment is .01 as meaning that the first experiment provides more evidence against the null hypothesis. Rather, one treats the two experiments identically: in both cases, one simply rejects the null hypothesis and accepts the alternative hypothesis.

4.2. Two Interpretations of Power. The two interpretations of *p*-values have counterparts in the case of power. First, one can assume that the power of an experiment measures the strength of evidence for the null hypothesis. Second, one can assume that the power of an experiment is a property of a test that allows for the application of a decision rule specifying when the null hypothesis is to be accepted and the alternative hypothesis rejected: accept the null hypothesis from a negative result when and only when power is above some threshold. Noticeably, Cohen assumes this second interpretation in the quotation given in section 2, and I have followed him when I

https://doi.org/10.1086/667877 Published online by Cambridge University Press

presented the rule governing the inference of null hypotheses from negative results.

4.3. Avoiding the Paradoxes. In the paradoxes described in section 3, p-values and power are treated as measures of evidence. One compares how much evidence for and against the null hypothesis E_1 and E_2 provide, and paradoxes emerge because E_1 provides more evidence both for and against the null hypothesis than does E_2 (or more evidence against the null hypothesis than does E_2 and the same amount of evidence for it as E_2).

However, if one endorses the alternative interpretations of p-values and power, then there is no experiment such that the antecedents of the rule about rejecting the null hypothesis on the basis of a low p-value and of the rule about accepting the null hypothesis on the basis of a high power are simultaneously satisfied. The former rule only applies when the p-value is below the significance level, while the second rule only applies when it is above the significance level (i.e., when a negative result occurs). As a result, in no experiment do these two rules result in contradictory decisions: accept and reject the null hypothesis. Thus, the paradoxes are avoided when p-values and power are treated not as measures of evidence but as properties of the data and test that allow for the application of decision rules.

One may argue that to circumvent the paradoxes discussed above it is sufficient to merely avoid interpreting power as a measure of evidence for the null hypothesis, while still interpreting p-values as measures of evidence. However, first, it is unclear what could justify the resulting asymmetry between power and p-values. Second, variants of the paradoxes discussed earlier that speak against this proposal are easy to formulate. Consider the case of observed power discussed in section 3.2: E_1 and E_2 are two (one-sample) experiments such that $N_1 = N_2$, $m_1 = m_2$, and $s_1 < s_2$. Furthermore, both E_1 and E_2 yield negative findings. As we have seen, in this case, $p_1 < p_2$, and the power of E_1 is larger than the power of E_2. It is possible that the power of E_1, but not the power of E_2, is above the threshold required for accepting the null hypothesis and rejecting the alternative hypothesis (e.g., .95). In this case, we would accept the null hypothesis in E_1 but not in E_2, even though E_1 provides more evidence against the null hypothesis than does E_2 (since p-values are interpreted as measures of evidence).

4.4. Moral. The moral that should be drawn from the paradoxes discussed in section 3 is not that, if one tests hypotheses by means of null hypothesis significance testing, one should not use power to infer a null hypothesis from a negative result, as one could have initially thought. What the paradoxes reveal is that one should not treat power and p-values as measures of evidence but rather as properties of the data and test that allow for the

https://doi.org/10.1086/667877 Published online by Cambridge University Press

application of decision rules for rejecting and accepting hypotheses. When one does so, one can use the power of an experiment to infer the null hypothesis from a negative result.

5. Objections and Responses.

5.1. The Proposed Solution to the Paradoxes Is Arbitrary. One could wonder why it is preferable to conclude from the paradoxes discussed in section 3 that p-values and power should not be treated as measures of evidence rather than to conclude that power should not be used to infer the null hypothesis from a negative result. One reason for embracing the first alternative is that interpreting p-values as measures of evidence gives rise to other problems. For instance, if p-values are measures of evidence, two experiments yielding identical p-values provide the same amount of evidence, even if their sample sizes are different (Cornfield's [1966] α postulate). This is inconsistent with treating the likelihood ratio as a measure of evidence since this ratio is influenced by the sample size (Royall 1986). It is worth noting briefly that, in contrast to other arguments against the interpretation of p-values as measures of evidence (e.g., Royall 1997), the argument advanced here does not appeal to, and thus is not dependent on, potentially controversial principles about the nature of evidence.

5.2. The Proposed Solution Leaves No Room for the Notion of Evidence. One could reject the proposed solution on the grounds that it leaves no room for measuring the strength of the evidence for statistical hypotheses. Since this short article is obviously not the place to discuss the role of the notion of evidence in statistical inference, I only note that philosophers should not view as patently absurd the idea that the acceptance of statistical hypotheses on the basis of statistics is not to be interpreted in evidential terms: after all, this interpretation was favored by Jerzy Neyman, whose methods are widely used throughout the sciences (Machery 2012).

5.3. The Proposed Solution Is in Tension with Some Common Practices and Norms in Psychology. Philosophers of science should plausibly hold scientists' practices as well as the norms they accept as prima facie correct. On this basis, one could object to the recommendation to avoid treating p-values and power as measures of evidence that it goes against both the usual practices of psychologists and the norms put forward by professional organizations in psychology such as the American Psychological Association (APA).

Psychologists often seem to hold that the smaller the p-value obtained in a test, the more confident one can be in rejecting the null hypothesis. Experimental research on psychologists' understanding of p-values supports this

https://doi.org/10.1086/667877 Published online by Cambridge University Press

claim: when psychologists are asked how confident they are in the rejection of the null hypothesis, their confidence decreases with increasing p-values (e.g., Poitevineau and Lecoutre 2001).[6] It is natural to interpret these findings as follows: psychologists view p-values as measures of the evidence against the null hypothesis, and their confidence in the rejection of the null hypothesis decreases with the diminishing available evidence against it.

Furthermore, some official norms in psychology seem to assume an evidential interpretation of p-values. In its influential report, the APA Task Force on Statistical Inference recommended that psychologists report the exact p-values instead of reporting whether p-values are below the significance level, as used to be common: “It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p value or, better still, a confidence interval” (Wilkinson and the Task Force on Statistical Inference 1999, 599). It is natural to conclude from this recommendation that the APA Task Force views p-values as measures of evidence. When they are so treated, the exact value of p matters: a p-value equal to, say, .03 provides more evidence against the null hypothesis than a p-value equal to, say, .04. By contrast, if p-values are just properties of the data that allow for the application of a rule for rejecting the null hypothesis, it does not matter whether a p-value is equal to .03 or .04, and reporting the exact value is of no particular use.

There is thus little doubt that the solution to the paradoxes proposed in this article is at odds with some common practices and norms in psychology. The disagreement with psychologists’ practices should not carry much normative weight, however, since their understanding of p-values and of the foundations of their statistical tests is often poor (e.g., Oakes 1986). The same is true of the disagreement with the norms in psychology. The norms put forward by professional organizations are not entirely consistent with one another (see Fidler [2002, 756] on the discrepancies about p-values between the report by the APA task force and the statistical section of the fifth edition of the *APA Publication Manual*). Furthermore, focusing only on the report of the APA task force, this report leaves psychologists without proper norms for inferring null hypotheses from negative results, a striking shortcoming. Finally, what norms of statistical inference psychology should adopt remains an area of active debate among psychologists and applied statisticians.

6. Conclusion. While the use of power to infer the null hypothesis from a negative result has recently been criticized, the proper solution to the paradoxes emerging from this use is to stop treating power and p-values as mea-

[6] This body of research is not without controversy. In particular, some studies, but not all, report a cliff effect at $p = .05$ (e.g., Rosenthal and Gaito 1963; but see Poitevineau and Lecoutre 2001).

https://doi.org/10.1086/667877 Published online by Cambridge University Press

sures of evidence. When one endorses this position, *p*-values and power are the basis of two symmetric rules for, respectively, rejecting and accepting the null hypothesis.[7]

REFERENCES

Boroditsky, Lera. 2001. "Does Language Shape Thought? Mandarin and English Speakers' Conceptions of Time." *Cognitive Psychology* 43:1–22.

Chen, Jenn-Yeu. 2007. "Do Chinese and English Speakers Think about Time Differently? Failure of Replicating Boroditsky (2001)." *Cognition* 104:427–36.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Cornfield, Jerome. 1966. "Sequential Trials, Sequential Analysis, and the Likelihood Principle." *American Statistician* 20:18–23.

Fidler, Fiona. 2002. "The Fifth Edition of the APA Publication Manual: Why Its Statistics Recommendations Are So Controversial." *Educational and Psychological Measurement* 62:749–70.

Hare, Brian, Alicia Melis, Patricia Woods, Sara Hastings, and Richard Wrangham. 2007. "Tolerance Allows Bonobos to Outperform Chimpanzees on a Cooperative Task." *Current Biology* 17: 619–23.

Hoenig, John M., and Dennis M. Heisey. 2001. "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." *Statistical Practice* 55:1–6.

January, David, and Edward Kako. 2007. "Re-evaluating the Evidence for Linguistic Relativity: Reply to Boroditsky (2001)." *Cognition* 104:417–26.

Kempthorne, Oscar, and Leroy Folks. 1971. *Probability, Statistics, and Data Analysis*. Ames: Iowa State University Press.

Lenth, R. V. 2007. "Statistical Power Calculations." *Journal of Animal Science* 85: E24–E29.

Leventhal, L. 2009. "Statistical Power Calculations: Comment." *Journal of Animal Science* 87: 1854–55.

Machery, Edouard. 2012. "Evidence and Cognition." Unpublished manuscript, University of Pittsburgh.

Oakes, Michael. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.

Poitevineau, Jacques, and Bruno Lecoutre. 2001. "Interpretation of Significance Levels by Psychological Researchers: The .05 Cliff Effect May Be Overstated." *Psychonomic Bulletin and Review* 8:847–50.

Rosenthal, Robert, and John Gaito. 1963. "The Interpretation of Levels of Significance by Psychological Researchers." *Journal of Psychology* 55:33–38.

Royall, Richard M. 1986. "The Effect of Sample Size on the Meaning of Significance Tests." *American Statistician* 40:313–15.

———. 1997. *Statistical Evidence: A Likelihood Paradigm.* New York: Chapman & Hall.

Sedlmeier, Peter, and Gerd Gigerenzer. 1989. "Do Studies of Statistical Power Have an Effect on the Power of Studies?" *Psychological Bulletin* 105:309–16.

Wilkinson, Leland, and the Task Force on Statistical Inference. 1999. "Statistical Methods in Psychology Journals: Guidelines and Explanations." *American Psychologist* 54:594–604.

[7] Or the rules are nearly symmetric: *p*-values are used to reject the null hypothesis and power to accept an approximation thereof.

https://doi.org/10.1086/667877 Published online by Cambridge University Press