JENA ECONOMIC
RESEARCH PAPERS

# 2013 – 030

One Swallow Doesn´t Make a Summer – A Note

by

Mitesh Kataria

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

One Swallow Doesn't Make a Summer – A Note

Mitesh Kataria†

Abstract

Maniadis et al. (2013) present a theoretical framework that aims at providing insights into the mechanics of proper inference. They suggest that a decision about whether to call an experimental finding noteworthy, or deserving of great attention, should be based on the calculated post-study probability. Although I in large agree with most points in Maniadis et al. (2013), this note raises some important caveats.

Keywords: Bayes' theorem, Methodology

JEL classification: C11, C12, C80

†Max Planck Institute of Economics, Strategic Interaction Group, Kahlaische Strase 10, D-07745 Jena, Germany. Tel: +49 3641 686 632. E-mail: kataria@econ.mpg.de

Acknowledgments: I am grateful for helpful suggestions on earlier versions of this note from Ivan Soraperra and Vittoria Levati. The usual disclaimer applies.

Introduction

In a recent article, Maniadis et al. (2013) claim that their "…framework highlights that, at least in principle, the decision about whether to call a finding noteworthy, or deserving of great attention, should be based on the estimated probability that the finding represents a true association, which follows directly from the observed p-value, the power of the design, the prior probability of the hypothesis, and the tolerance for false positives." One of the authors' contributions "is to illustrate this basic point by means of a theoretical framework that provides insights into the mechanics of proper inference." Although I agree with most of the conclusions in Maniadis et al. (2013), this note raises some important caveats.

Theory and Analysis

The basic framework in Maniadis et al. (2013) consists in calculating the post-study probability (PSP), which is the probability that a research finding that is statistically significant, is true.

$$PSP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

where $\alpha = P(test\,wrong|H_0)$ and reads as the probability that the test statistics rejects H_0 (i.e. erroneously favors H_1) when H_0 is true, $1-\beta = P(test\,correct|H_1)$ and reads as the probability that a research hypothesis is found significant when it is true, and $\pi = P(H_1)$ is the unconditional probability that H_1 is true.[1] Alternatively, we can write it in terms of the probability that the null hypothesis H_0 is true given that the data D provides support for the alternative hypothesis H_1. The probability that a research finding that is statistically significant is false is

$$P(H_0|D) = \frac{P(test\,wrong|H_0)\cdot P(H_0)}{P(test\,wrong|H_0)\cdot P(H_0) + P(test\,correct|H_1)\cdot P(H_1)} = 1 - PSP$$

[1] Hence α denotes the probability of a type 1 error and β denotes the probability of type 2 error and $1-\beta$ denotes the power of the test. In repeated random sampling α and β are the long run frequencies of type 1 and type 2 errors.

Note the use of Bayes' theorem.[2] This approach is widely applied in medical and psychiatric diagnosis where all of the terms in right-hand side of the equation are assumingly known, including $P(H_0)$, which is the unconditional probability of the prevalence of a disease in the population. Calculating the PSP, therefore, is of great value and provides information on how likely it is that a patient who is given a positive diagnosis actually has a disease.

Maniadis et al. (2013) remind us that the probability of rejecting H_0 when H_0 is true (i.e. the probability of committing type 1 error) is not equal to the probability that the hypothesis H_0 is true when H_0 is rejected. Table 2 in Maniadis et al. (2013) shows, for example, that if $P(H_0)$ is known and equals 0.99, and $P(test\ wrong|H_0) = 0.05$, and $P(test\ correct|H_1) = 0.80$, then Bayes' theorem allows us to calculate the conditional probability $P(H_0|D) = \frac{(0.05)\cdot(0.99)}{(0.05)\cdot(0.99)+(0.80)\cdot(0.01)} = 0.86,$ which is the posterior probability that the null is true when the researcher rejects the null. Hence, the PSP states that there is only a 14 percent chance that the statistically significant finding at the 5% level will represent a true association. Moreover, this is still far from the worst case that is presented. Maniadis et al. (2013) calculates several PSP's under the assumption that the priors ranges between $0.45 < P(H_0) < 0.99$. Based on the general impression from these calculations, Maniadis et al. (2013) conclude that "…it is not unlikely that the *PSP* after the initial study is less than 0.5 as several plausible parameter combinations yield this result (presented by bold fonts)". [3] As mentioned, Maniadis et al. (2013) suggest that a decision about whether to call an experimental finding noteworthy, or deserving of great attention, should be based on the Bayesian post-study probability since the Classical procedure is shown to be problematic.

$P(D|H_0) \neq P(H_0|\text{D})$ follows immediately from Bayes' theorem. About 20 years ago, Cohen (1994) raised this issue in the context of Null Hypothesis Significance Testing (NHST) in one of the major psychology journals. The point was made that there could be a chance as low as 60 percent that the statistically significant finding will represent a true association when $P(test\ wrong|H_0) = 0.05$ i.e. at a 5% significance level. Baril and Cannon (1995) replied that instead of demonstrating that $P(D|H_0) \neq P(H_0|\text{D})$ with fabricated data to illustrate how different these probabilities can be, it would be more informative to estimate how large the gap between the conditional and reversed conditional probabilities is *likely* to be. In his reply Cohen (1995) made clear that his example was not intended to model NHST as used "in the

[2] A more sophisticated approach would require the specification of a prior distribution and not only the prior probability.

[3] This is to say, the conjecture is that $P(H_0|D)$ is higher than 0.5.

real world" but rather to demonstrate how wrong one can be when the logic of NHST is violated. In light of the claims in Maniadis et al. (2013), there is a need to revisit the results in Baril and Cannon (1995).

The starting point in Baril and Cannon (1995) is that statistical power cannot be sufficiently good to detect all effect sizes. Assuming that the effect sizes follow a standard normal distribution centered at zero and that scientists' only detect and consider effect sizes $\pm\ 0.2d$ as relevant (d is what is known as Cohen's effect size, i.e., it is the difference between means divided by the standard deviation), approximately 16 percent could be considered as equivalent of H_0 being true.[4] Based on Rossi (1990), the average statistical power for moderate effect sizes (i.e. $d > 0.2$) is appreciated to be 0.57. Finally, the conventional $P(test\,wrong|H_0) = 0.05$ is applied. Using Bayes' theorem, we now have: $P(H_0|D) = \frac{(0.05)\cdot(0.16)}{(0.05)\cdot(0.16)+(0.57)\cdot(0.84)} = 0.016,$ i.e. the PSP states that there is a 98.4 percent chance that the statistically significant finding will represent a true association. This means that the probability of H_0 being true given a significant test is 0.016, which is not very different from 0.05 which is, in turn, the probability of a significant test given that H_0 is true. Clearly, $0.016\ \neq 0.05$, but still the conditional and reversed conditional probabilities are shown to be not very different once a different parameter space is adopted than in Maniadis et al. (2013). Although it is possible that estimates (e.g. statistical power) are different in economic experiments compared to psychological experiments, using estimates from a related field can still serve as a first informative approximation. Also note that even if we assume that the statistical power takes a considerable lower value of 0.20, the PSP equals 0.95 which means that there is a 95 percent chance that the statistically significant finding will represent a true association. More crucial to our results is that we assumed that scientists are willing to consider economic significance instead of only hunting statistical significance, which partly is a normative statement on how to apply classical statistics.

Remember that Maniadis et al. (2013) assumed priors in the range of $0.45 < P(H_0) < 0.99$ to calculate PSP, which is obviously far off from the neighborhood of $P(H_0)$ ≈0.16, and they show that in the absence of other biases, such as research competition and research misconduct, the Classical framework still leads to an "excessive number of false positives"

[4] The point that economists *should* consider economic significance together with statistical significance is raised by McCloskey (1985). In case absolute substantive significance is hard to corroborate, Cohen's *d* statistics offers a relative measure that facilitates sample size planning and power analysis.

compared to what is stated in the significance level.[5] The conclusion in Maniadis et al. (2013) that we *should* embrace the Bayesian framework seems exaggerated and is based on selective empirical support that only considers 0.45 < $P(H_0) < 0.99$ and excludes support in the neighborhood of $P(H_0) \approx 0.16$ which is appreciated to be a more realistic estimate that would change their main result.

At this point we have not even taken into account that the prior could be biased but instead we have treated it as a known which is in line with the simulation in Maniadis et al. (2013). But this should not go uncommented because therein lies the real rub. Knowing the unconditional probability facilitates assessment of the probability that a research hypothesis that is statistically significant is true but it is only feasible in the Bayesian framework. Note, however, that the context in medical diagnosis where PSP is often calculated is different from hypothesis testing in economics. In medicine, the aim is to find the conditional probability that an individual patient who is given a positive diagnosis actually has the disease and the unconditional probability, i.e., prevalence in the population, is considered attainable. For economic hypotheses, the unconditional probability $P(H_0)$ is hardly ever known. Bayesian statistics cope with this problem assuming that the prior probability is a subjective belief that is subject to revisions.

The assumption that the prior probability $P(H_0)$ is a subjective belief, facilitates a move from the Classical to a Bayesian framework, even when the prior is unknown. What is worth emphasizing is that based on a single experiment and using prior beliefs we do not estimate the unbiased $P(H_0|D)$ in the Bayesian framework, i.e. both approaches can lead to erroneous conclusions[6]. Going back to the example of Baril and Cannon (1995), remember that the conditional probability was calculated to $P(H_0|D) = \frac{(0.05)\cdot(0.16)}{(0.05)\cdot(0.16)+(0.57)\cdot(0.84)} = 0.016$ and it was assumed that the unconditional probability is known and equals 0.16. Let us instead assume that the unconditional probability is unknown and that the subjective beliefs are that the prior

[5] The conclusion that the Classical statistics leads to an "excessive number of false positives" is reached under the definition that the benchmark probability of false positives is the probability that H_0 is true when H_0 is rejected. The significance level in Classical statistics on the other hand measures the probability to reject H_0 when H_0 is true (i.e. error of the first kind). Hence the claim that Classical statistics leads to "excessive number of false positives" is another way to claim that there is a positive difference between the conditional and reversed conditional probabilities.

[6] Neyman-Pearson error probabilities have a long-run repeated sampling interpretation. E.g. probability of making type 1 error can only be controlled in repeated sampling, opposed to a single experiment.

corresponds to $P(H_0) = 0.99$. In this case, $P(H_0|D) = \frac{(0.05)\cdot(0.99)}{(0.05)\cdot(0.99)+(0.57)\cdot(0.01)} = 0.897$. Hence, while $P(D|H_0) = 0.05$ is close to the correct benchmark of $P(H_0|D) = 0.016$, the conditional probability based on subjective beliefs is considerably higher, namely at $P(H_0|D) = 0.897$. The example demonstrate that it is easy to come up with counterexamples to Maniadis et al.'s (2013) simulation and show that the Bayesian framework does not necessarily perform better than the Classical framework but might even perform worse in terms of estimating the $P(H_0|D)$. In the example above the PSP calculation underestimates the probability that a statistically significant research finding is true.[7]

The conceptual difference between the Classical and Bayesian framework based on prior beliefs of $P(H_0)$ also deserves to be mentioned. In Classical statistics a probability is the long-run relative frequency, while in the Bayesian framework a probability is the degree of beliefs. While posterior $P(H_0|D)$ undeniably has an appealing interpretation, it is only available through Bayes' theorem which Fisher (1937) rejected with the motivation that it requires one to: "…regard mathematical probability not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes" (p.6). Although Fisher's position may be perceived as extreme, we mention it to place the difference between the Classical and Bayesian approach in a historical context.

Conclusions

Based on what is presented in Maniadis et al. (2013), the conclusion that only a Bayesian analysis provides "proper inference" seems exaggerated. The assumption that the unconditional probability $P(H_0)$ is known[8] implies that the Bayesian approach can only be better but never worse than the Classical approach in their simulation. Once we relax this assumption (i.e. allow for subjective beliefs) it is no longer trivial to decide whether the Classical or the Bayesian framework is better. This they combined with a selective empirical setup that also favors the Bayesian framework by excluding many instances where the bias in the Classical approach is small. This makes the simulation in Maniadis et al. (2013) great in

[7] By incorporating subjective beliefs into the inference process, the risk of introducing errors or biases that would not otherwise be present is, of course, inevitable. On the other hand, the Bayesian approach is particularly useful when one has strong prior knowledge of a situation and wants to summarize the accumulated evidence.

[8] While Maniadis et al. (2013) make use of different values of $P(H_0)$ to calculate the difference between conditional and reversed conditional probabilities, in each calculation it is assumed that $P(H_0)$ is known (unbiased) which makes the Bayesian approach the benchmark and the Classical approach biased.

demonstrating the pitfalls of Classical framework, however, their conclusion about “proper inference” is questionable. Finally, also note that the simulation in Maniadis et al. (2013) focuses on the gap between conditional and reversed conditional probabilities but ignores all other arguments that could be perceived important in a comparison between the Classical and Bayesian framework.

References

Baril, G. L., and Cannon, J. T. 1995. “What *is* the probability that null hypothesis testing is meaningless?,” *American Psychologist, 50,* 1098-1099.

Cohen, J. (1994). “The earth is round (p < .05),” *American Psychologist, 49,* 997-1003.

Cohen, J. (1995). “The earth is round (p < .05): Rejoinder,” *American Psychologist, 50,* 1103.

Fisher, R. A. (1937). The design of experiments. London: Oliver & Boyd.

Maniadis, Z., Tufano, F., and List, J. A.*:* “One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects,” *Amer. Econ. Rev.,* Forthcoming.

McCloskey, D. N. "The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests," *Amer. Econ. Rev.*, May 1985, 75 (2), 201-05.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Counseling and Clinical Psychology, 58,* 646-656.