EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 11(1)
January 2014: 4-10

One Swallow Doesn't Make a Summer: A Comment on Zacharias Maniadis, Fabio Tufano, and John List

Mitesh Kataria[1]

LINK TO ABSTRACT

In their article "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects," Zacharias Maniadis, Fabio Tufano, and John List—hereafter, MTL—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" (MTL 2014, 278). MTL's article is intended to provide "insights into the mechanics of proper inference" (ibid.). Although I agree with most of the conclusions in MTL (2014), in this comment I raise some important caveats.

Theory and analysis

MTL are interested in the "Post-Study Probability (*PSP*)," which is the probability that a research finding that is statistically significant is true (MTL 2014, 284). Their equation (1), reproduced here as my equation (1), gives a formula for *PSP*:

1. University of Gothenburg, 40530 Gothenburg, Sweden.

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

Interpretations of the expression's terms are as follows:

- $\alpha = P(\text{test wrong} \mid H_0)$, the probability that the test statistic rejects H_0 (i.e., erroneously favors H_1) when H_0 is true,
- $1 - \beta = P(\text{test correct} \mid H_1)$, the probability that a research hypothesis is found significant when it is true,
- $\pi = P(H_1)$, the unconditional probability that H_1 is true.[2]

Alternatively, we can write an expression for *PSP* 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

$$P(H_0 \mid D) = \frac{P(\text{test wrong} \mid H_0) \cdot P(H_0)}{P(\text{test wrong} \mid H_0) \cdot P(H_0) + P(\text{test correct} \mid H_1) \cdot P(H_1)} = 1 - PSP \tag{2}$$

Note the use of Bayes' theorem.[3] The approach represented in equation (2) is widely applied in medical and psychiatric diagnosis, where all of the terms in right-hand side of the equation are presumably known, including $P(H_0)$, which would be 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.

MTL 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 MTL (2014, 286) shows, for example, that if $P(H_0)$ is known and equals 0.99, and $P(\text{test wrong} \mid H_0) = 0.05$, and $P(\text{test correct} \mid H_1) = 0.80$, then Bayes' theorem allows us to calculate the conditional probability $P(H_0 \mid D) = \frac{(0.05) \cdot (0.99)}{(0.05) \cdot (0.99) + (0.80) \cdot (0.01)} = 0.86$, 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, given a statistically significant finding at the 5% level, that there is a true association. Moreover, this estimate is still far from the worst case that is presented. MTL calculate several *PSP*s under the assumption that the priors are in the interval $0.45 < P(H_0) < 0.99$. Based on the general impression from these calculations, MTL conclude that "it is

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

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

not unlikely that the *PSP* after the initial study is less than 0.5, as several plausible parameter combinations yield this result" (2014, 287). That is to say, the conjecture is that $P(H_0|D)$ is higher than 0.5. As mentioned, MTL (2014) 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 have problems.

It follows immediately from Bayes' theorem that $P(D|H_0) \neq P(H_0|D)$. About 20 years ago, in *American Psychologist*, Jacob Cohen (1994) raised this issue in the context of null hypothesis significance testing. Cohen made the point that there could be a chance as low as 40 percent that the statistically significant finding represented a true association even though $P(\text{test wrong}|H_0) = 0.05$, i.e., at a 5% significance level. In the same journal, Galen Baril and Timothy Cannon (1995) replied that, instead of using fabricated data to illustrate how different the probabilities can be, that is, that $P(D|H_0) \neq P(H_0|D)$, 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 null hypothesis significance testing as used "in the real world" but rather to demonstrate how wrong one can be when the logic of null hypothesis significance testing is violated. In light of the claims in MTL (2014), 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 $|d| > 0.2$ 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 to H_0 being true.[4] Baril and Cannon make use of an estimate from Joseph Rossi (1990) that the average statistical power for moderate effect sizes (i.e., $d > 0.2$) is 0.57. Finally, the conventional $P(\text{test wrong}|H_0) = 0.05$ is applied. Using Bayes' theorem, we now have: $P(H_0|D) = \frac{(0.05)\cdot(0.16)}{(0.05)\cdot(0.16)+(0.57)\cdot(0.84)} = 0.016$, that is, the *PSP* states that there is a 98.4 percent chance that the statistically significant finding will represent a true association. Such a statement would mean 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 parameter space different from that adopted in MTL

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 statistic offers a relative measure that facilitates sample size planning and power analysis.

(2014) is adopted. The example also shows that the Classical significance test can be even more conservative than realized. 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 be useful as a first approximation. Also note that even if we assume that the statistical power takes a considerably lower value of 0.20, the *PSP* then 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 hunting only for statistical significance, such assumption affirming a norm about how to apply classical statistics.[5]

Remember that MTL assumed priors in the range of $0.45 < P(H_0) < 0.99$ to calculate *PSP*, a range that is obviously far off from the neighborhood of $P(H_0) \approx 0.16$, and they show that there, even in the absence of other biases such as research competition and research misconduct, the Classical framework leads to an "excessive number of false positives" (2014, 278) compared to what is stated in the significance level.[6] But MTL's conclusion that we should embrace the Bayesian framework seems exaggerated. The conclusion is based on this selective empirical support that only considers $0.45 < P(H_0) < 0.99$ and excludes the neighborhood of $P(H_0) \approx 0.16$, a neighborhood that is appreciated to be a more realistic estimate and 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 postulated that it is a known, a postulation that is in line with the simulation in MTL (2014). But this should not go uncommented, because therein lies the real rub. Postulating that the unconditional probability is known facilitates assessment of the probability that a research hypothesis that is statistically significant is true. But this probability is feasible only in the Bayesian framework.

---

5. To understand the need of such norm, consider an economic experiment with a control and an experimental treatment. As soon as the experimental treatment has a non-zero percent of subjects that behave differently in the experimental treatment, retrieving a statistically significant result is only a matter of choosing the right sample size. A non-zero threshold, e.g., $|d| > 0.2$, adds a constraint on substantive significance. Choosing an appropriate threshold is of course a non-trivial task.

6. MTL's conclusion that 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 an "excessive number of false positives" is another way to claim that there is a positive difference between the conditional and reversed conditional probabilities. Importantly, there is no "excessive number of false positives" if we apply the standard definition in Classical statistics that the probability of false positives is the probability of error of the first kind.

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, that is, prevalence in the population, is considered to be known or available. For economic hypotheses, the unconditional probability $P(H_0)$ is hardly ever known. Bayesian statistics cope with this problem by assuming that the prior probability is a subjective belief, possibly mistaken, and subject to revisions.

This assumption, that the prior probability $P(H_0)$ is a possibly mistaken 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 necessarily estimate the unbiased $P(H_0 \mid D)$ in the Bayesian framework. Going back to the example of Baril and Cannon (1995), remember that the conditional probability was calculated to be $P(H_0 \mid D) = \frac{(0.05) \cdot (0.16)}{(0.05) \cdot (0.16) + (0.57) \cdot (0.84)} = 0.016$, and it was assumed that the unconditional probability is known and equals 0.160. Let us instead assume that the unconditional probability is unknown and that the subjective beliefs are that the prior corresponds to $P(H_0) = 0.99$. In this case, $P(H_0 \mid D) = \frac{(0.05) \cdot (0.99)}{(0.05) \cdot (0.99) + (0.57) \cdot (0.01)} = 0.897$. Hence, although $P(D \mid H_0) = 0.05$ is close to the correct benchmark of $P(H_0 \mid D) = 0.016$, the conditional probability based on subjective beliefs is considerably higher, namely at $P(H_0 \mid D) = 0.897$. The example demonstrates that it is easy to come up with counterexamples to MTL's (2014) simulation and thereby show that the Bayesian framework does not necessarily perform better than the Classical framework, and might even perform worse, in estimating $P(H_0 \mid D)$. 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 frameworks regarding prior beliefs about $P(H_0)$ 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 the belief. Although posterior $P(H_0 \mid D)$ undeniably has an appealing interpretation, it is only available through Bayes' theorem, which R. A. Fisher 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" (Fisher 1937, 7).

7. By incorporating subjective beliefs into the inference process, the risk of introducing errors or biases that would not otherwise be present is 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.

Although Fisher's position may be perceived as extreme, I mention it to place the difference between the Classical and Bayesian approach in an historical context.

Conclusions

Based on what is presented in Maniadis, Tufano, and List (2014), 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 by allowing for subjective beliefs, it is no longer trivial to decide whether the Classical or the Bayesian framework is better. MTL combined the assumption that the unconditional probability is known with a selective empirical setup that also favors the Bayesian framework by excluding many instances where the problems of the Classical approach are small. Such moves do, of course, make the simulation in MTL (2014) great for demonstrating the pitfalls of the Classical framework.

References

Baril, Galen L., and J. Timothy Cannon. 1995. What *Is* the Probability That Null Hypothesis Testing Is Meaningless? *American Psychologist* 50: 1098-1099.

Cohen, Jacob. 1994. The Earth Is Round ($p < .05$). *American Psychologist* 49: 997-1003.

Cohen, Jacob. 1995. The Earth Is Round ($p < .05$): Rejoinder. *American Psychologist* 50: 1103.

Fisher, R. A. 1937. *The Design of Experiments*, 2nd ed. London: Oliver & Boyd.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2014. One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *American Economic Review* 104(1): 277-290.

McCloskey, D. N. 1985. The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests. *American Economic Review* 75(2): 201-205.

8. While MTL (2014) 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 into the benchmark from which any observed deviations under the Classical approach are interpreted as bias.

Rossi, Joseph S. 1990. Statistical Power of Psychological Research: What Have We Gained in 20 Years? *Journal of Consulting and Clinical Psychology* 58(5): 646-656.

About the Author

Mitesh Kataria is senior researcher in the Strategic Interaction Group at the Max Planck Institute of Economics, Jena, Germany, and associate senior lecturer in the Department of Economics at the University of Gothenburg, Gothenburg, Sweden. His current research concerns are in applied (environmental and welfare) economics, empirical behavioral economics, and experimental economics. His email address is mitesh.kataria@economics.gu.se.

Maniadis, Tufano, and List's reply to this article
Go to archive of Comments section
Go to January 2014 issue