ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2021, Vol. 25, No. 1
ISSN 1507-3866; e-ISSN 2449-9994

PHILOSOPHICAL OR EMPIRICAL INCOMMENSURABILITY OF FREQUENTIST VERSUS BAYESIAN THINKING

David Trafimow
New Mexico State University, Las Cruces, United States
e-mail: dtrafimo@nmsu.ed

ORCID: 0000-0002-0788-8044

© 2021 David Trafimow
This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/

Quote as: Trafimow, D. (2021). Philosophical or empirical incommensurability of frequentist versus Bayesian thinking. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, *25*(1).

DOI: 10.15611/eada.2021.1.02
JEL Classification: C1

Abstract: Frequentists and Bayesians disagree about the soundness of performing calculations based, in an important part, on prior information. The disagreement goes back to a basic philosophical disagreement about how to conceptualize the meaning of probability. As frequentists and Bayesians use the term differently, there is a basic philosophical incommensurability. However, this philosophical incommensurability need not imply an empirical incommensurability. It is possible for there to be, simultaneously, philosophical incommensurability and empirical commensurability. This possibility implies consequences that this article discusses.

Keywords: incommensurability, commensurability, a priori procedure, estimation, hypothesis testing, Bayes factors.

1. Introduction

There is a long-standing argument between frequentists and Bayesians about whether it is valid or invalid to use prior information. The argument is foundational because Bayesian statistics feature prior information whereas frequentist statistics do not. Frequentist statistics are superior to Bayesian statistics because frequentists do not make assumptions about prior information that may not be true. However, Bayesian statistics are superior to frequentist statistics because it is possible to draw stronger conclusions. For example, frequentist statistics do not provide the probability that a hypothesis is true whereas Bayesian statistics – if one buys into the validity of prior probability distributions – do provide this, Or, in terms of intervals, a frequentist confidence interval does not provide the probability that the population parameter

of interest is within the interval. However, a Bayesian credible interval does provide that probability, again if the researcher 'buys into' the validity of prior probability distributions.

Hence, the debate between frequentists and Bayesians comes down, largely, to the issue of whether it is permissible to use prior information to draw conclusions. The debate can be summarized in the form of two pointed questions; one from a frequentist to a Bayesian and the other from a Bayesian to a frequentist.

- Pointed question from a frequentist to a Bayesian: "How can you justify making assumptions about the prior distribution when you don't know, or even have a good idea, about the nature of that prior distribution?"
- Pointed question from a Bayesian to a frequentist: "How can you justify not using all of the information that is available to you to draw the strongest conclusions possible?"

Both pointed questions lead to a third question, "Is it justifiable to use prior information?" This is the present topic.

2. What is probability?

Most frequentists consider probability to be a relative frequency or propensity (see Popper, 1983, for a philosophically-based review). For instance, if there are black and white balls in an urn, and one is randomly selecting balls, with replacement, then the relative frequency definition of probability implies that the probability of choosing a white ball is the frequency of white balls in the urn, divided by the total number of balls in the urn. To take another instance, imagine a coin-flipping machine. An aficionado of the propensity conception of probability would argue that the probability of heads depends on the propensity of the coin-machine system to produce heads. To say that the probability of heads is 0.50 is to say that it is a characteristic of the coin-machine system that it produces heads at that rate.

In contrast, although there are objective Bayesian approaches to the notion of probability (see Gillies, 2000 for a review), most Bayesians consider probability to be a state of belief. Thus two people, who have different information, could rationally disagree about probabilities. The tension between probability as a relative frequency or propensity, versus probability as a state of belief, has existed for centuries (see Stigler, 1990, for a review). In fact, some writers have dramatized the difference by using subscripts: probability_1 and probability_2 for frequentist versus Bayesian probability, respectively.

Gillies (2000) performed an important service by pointing out the many philosophical dilemmas in which Bayesians can find themselves. The harder dilemmas are associated with objective Bayesian approaches, though even subjective Bayesian approaches are not without problems (also see Trafimow, 2006). There is no point in reiterating the dilemmas, but it seems worthwhile to summarize. Objective Bayesian approaches tend to be problematic because (a) it is easy to set up scenarios

where Bayesian probabilities give different answers than probabilities based on relative frequencies or propensities, (b) different objective Bayesian methods result in different answers, or (c) it often is not clear how to be objective in the first place. Subjective Bayesian probabilities suffer from the problem that one has to admit that different people can assign different probabilities with both of them being 'rational' or 'right'. Frequentists naturally find this sort of admission uncomfortable at best, and downright nonsensical at worst.

Thus we have reached an impasse. One either accepts that the definition of probability allows for the assignment of different values by different people or one does not accept it. There is a philosophical incommensurability between frequentists and Bayesians due to their different probability conceptions; they are not speaking the same language. Unsurprisingly, this incommensurability complicates the process of discussing the two approaches.

Is there a way out of the impasse that allows a reasonable discussion to ensue?

3. Philosophically trivial Bayes

Suppose for a while, that we accept a frequentist conception of probability as being most philosophically palatable.[1] As the previous section indicates, this seems to leave no room for Bayesian thinking, but there are two reasons why this conclusion is premature. The first reason is being considered now, and the second, and more important one, is reserved for later.

The first reason is philosophically trivial and can be exemplified with very simple Bayesian cases. The underlying premise is that Bayes theorem is a proper theorem in the probability calculus, even for frequentists, provided that the researcher has frequentist probabilities to instantiate into the equation. There are many undergraduate homework assignments where the textbook or teacher provides the relevant probabilities for undergraduates to instantiate into Bayes theorem, and the answers are objectively correct. This level of thinking has been used in research too, for instance, when there is high quality epidemiological data on which frequentist probabilities can be based. Trafimow and Trafimow (2016) provided such an example in the case of slipped disks (herniated nucleus pulposus). The available orthopedic literature included epidemiological work that provided the base rate for slipped disks (0.03 or less according to Andersson, 1991; Borenstein, Wiesel, and Boden, 1995; Frymoyer, 1988; Lawrence et al., 2008), the conditional probability of a positive MRI result given that the person has a slipped disk (0.804 according to Boos et al., 1995), and the probability of a positive MRI result given that the person

[1] According to the so-called Bayesian "principal principle", even a Bayesian should use relative frequency information to assign probabilities if that information is available. This could be considered a tacit admission that probabilities assigned on the basis of relative frequency information are superior to probabilities assigned based on degrees of belief.

does not have a slipped disk (0.30 according to Jensen et al., 1994; Marshall et al., 1977). Instantiating these values into Bayes theorem results in the following:

$$P(S|+) = \frac{P(S)P(+|S)}{P(S)P(+|S)+P(+|\neg S)[1-P(S)]} = \frac{(0.03)(0.30)}{(0.03)(0.30)+(0.804)[1-(0.03)]} = 0.0114;$$

where

- $P(S|+)$ is the probability of a slipped disk given a positive MRI result,
- $P(S)$ is the base rate probability of a slipped disk (0.03),
- $P(+|S)$ is the probability of a positive MRI given a slipped disk (0.30), and
- $P(+|\neg S)$ is the probability of a positive MRI given not a slipped disk (0.804).

Thus, at the philosophically trivial level, where there is good relative frequency information on which to base probability designations, we see that Bayes theorem can be used in a way that both frequentists and Bayesians would accept. Naturally, this avoids the nontrivial issue of what to do when quality relative frequency or propensity information is not available.

4. Philosophically nontrivial Bayes

Frequentists and Bayesians suffer an incommensurability with respect to basic philosophy; their probability meanings do not match, and thus they are speaking different languages. However, the Philosophically Trivial Bayes section nevertheless shows that a basic philosophical incommensurability need not lead to an empirical incommensurability; as the slipped disk example shows, there exist empirical cases where frequentists and Bayesians would agree. That philosophical incommensurability does not necessarily entail empirical incommensurability suggests an important possibility. Perhaps Bayesian approaches can be useful to frequentists, and perhaps not, depending on the quality of the prior information that the researcher happens to have. Hence we have arrived at the nontrivial Bayesian issue about how to think when there is a lack of frequentist prior probability information, i.e. from a frequentist point of view, the prior information is of lesser quality than in the slipped disk example.

Let us reconsider the slipped disk example, but this time supposing a lack of relevant epidemiological data, so there is no way to have a frequentist estimate of the base rate probability of slipped disks. What value does one instantiate into Bayes theorem? Laplace's Principle of Indifference could be argued to support that the researcher should use a base rate probability of 0.50. Instantiating that value into Bayes theorem results in a posterior value, for the probability of a slipped disk given a positive MRI result, namely 0.728. This is very different than the value of 0.0114 and dramatizes how easily a Bayesian can go wrong from a frequentist perspective. Undoubtedly, there are more sophisticated Bayesian approaches. For example, an alternative way to use Laplace's Principle of Indifference would be to assume a prior uniform distribution, rather than using a single number, or a non-Laplacian Bayesian

might use some other distribution (normal, beta, gamma, Cauchy, etc.). These more sophisticated approaches might result in numbers that are less 'off', but there would remain a substantial discrepancy between these hypothetical results and the published ones. As there are also many possible ways to approach the issue of which prior distribution to use, and different ways to choose parameters to use even if there were agreement about the prior distribution itself; Bayesians can disagree with each other, and do disagree with each other, about which prior distributions and parameters to use for different problems. This is not the only issue, as empirically based numbers were available for the true and false positive rates pertaining to slipped disks and MRIs, but there is no guarantee that these will be generally available.

The contrast between using Bayes when there are empirically-based (or frequency-based) values to instantiate into the theorem, versus when there is a lack of such values, suggests four extreme but illustrative categories of possibilities.

4.1. Four extreme possibilities

The four extreme possibilities are as follows:

1. There is high-quality prior probability information and low-quality posterior probability information.
2. There is low-quality prior probability information and high-quality posterior probability information.
3. There is high-quality prior and posterior probability information.
4. There is low-quality prior and posterior probability information.

Let us consider the implications of each. When there is high-quality prior probability information and low-quality posterior information, it makes sense to be an empirical Bayesian, even if one is a philosophical frequentist. This is because the empirical Bayesian will obtain a more accurate frequentist posterior probability than the empirical frequentist who eschews prior probability information.

Let us consider the example of white and black balls in an urn, under sampling with replacement. Imagine that there are two phases to the study. In Phase I, 1,000,000 balls are randomly sampled whereas only 10 balls are randomly sampled in Phase II. As there is random sampling in both phases, the obvious thing to do is consider the total sample to be 1,000,000 + 10 = 1,000,010 cases. The estimated probability of choosing a white ball would be the frequency of white balls sampled across the two phases, divided by 1,000,010. However, it is illustrative not to combine the two samples and see where it leads. Not combining the two samples results in separate, and almost certainly different, Phase I and Phase II probabilities. As the Phase I probability is based on a much larger sample size than the Phase II probability, it is obvious that the Phase I probability should dominate the final estimate. In this example, Bayesians and frequentists would agree that it would be a blatant mistake to ignore the Phase I probability.

Reversing the Phase I and Phase II sample sizes dramatically changes the conclusion in the previous paragraph. In this case, the Phase I probability could be dropped with very little harm, and it is obvious that the Phase II probability should dominate.

Cases three and four are relatively uninteresting, therefore they can be addressed quickly. When all the information is of high-quality, it does not matter what the researcher will do, as the probability estimate obtained will be good regardless of whether it is based on the Phase I probability, the Phase II probability, or both. When there is only low-quality information, the obtained probability estimate likely will be poor, regardless of whether it is based on the Phase I probability, the Phase II probability, or both.

4.2. Increasing the realism of the example

The previous example is unrealistic because it ignores the fact that there often is a qualitative difference between the nature of the prior and posterior information. The lack of realism was useful because it dramatized extreme conditions where frequentists and Bayesians would agree. Let us increase the realism of the example by introducing a qualitative contrast remaining with the black and white balls in an urn, but this time changing the nature of the Phase I information. Now, suppose that in Phase I there is no sampling of balls from the urn, instead there is a pronouncement that the person who placed the balls in the urn likes white balls better than black balls, and consequently may have placed more white balls than black balls in the urn. Obviously, there is no guarantee that the person placed more balls of the liked than disliked colour in the urn, although it is somewhat likely.

What can be done now? It might depend on the nature of the Phase II information. In the case where there are 1,000,000 random draws, the obvious solution is to ignore the Phase I information and use only the posterior probability information. This would be what a frequentist would do anyway. A Bayesian might go in this direction too, on the grounds that with such a large sample size, posterior information would swamp prior information, unless one sets the prior probability of drawing a white ball at zero or one.

Let us suppose that there were only ten random draws in Phase II. In that case, the Phase II probability, used alone, is likely to be a poor estimate. Can the estimate be improved by using the Phase I information? Here is where frequentists and Bayesians would likely have a serious disagreement. A frequentist would insist on only using the Phase II probability, with an admission that it is likely to be an inaccurate estimate. A Bayesian would say that it is silly to settle for that, and would be comfortable using a prior probability distribution to improve the estimate. Other Bayesians might disagree about which prior probability distribution to use or where to set the parameters of the prior probability distribution, but they would at least agree that there ought to be a prior probability distribution. For example, a Bayesian

might suggest that as a beta distribution is conjugate to a binomial distribution, the researcher might commence with a prior beta distribution and then use the ten random draws to update.

5. Revisiting the distinction between philosophical and empirical incompatibility

We have seen that frequentists and Bayesians have different concepts of what probability means, thereby resulting in a basic philosophical incommensurability. However, philosophical incommensurability need not imply empirical incommensurability. The slipped disk example provided a philosophically trivial case where frequentists and Bayesians would obtain the same posterior probability. In addition, when Phase I and Phase II of the urn example are based on random draws, one can see again that there is empirical commensurability between frequentists and Bayesians, despite the philosophical incommensurability. In contrast, when Phase I and Phase II information were qualitatively different, we finally see that there is empirical incompatibility between frequentists and Bayesians. When the Phase II sample size is not large, it seems debatable whether or not to use Phase I information. To a frequentist, the qualitative difference between Phase I and Phase II constitutes an apples-versus-oranges problem, and the two cannot be combined. To a Bayesian, it is a fruit problem, with apples and oranges counting as fruit, and so it makes sense to combine them. We are finally at the crux of the empirical difficulty.

5.1. Laplace's Demon

It is now time to invoke Laplace's Demon, who is omniscient and always truthful, to help clarify this thinking. Let us consider three urn scenarios, but this time involving the Demon who knows the frequencies of black and white balls in the urn.

- *Scenario 1*. A researcher suggests to the Demon a particular prior probability distribution, and the parameters of that distribution too. The researcher then asks the Demon whether the resulting probability estimate is more accurate instantiating those assumptions into Bayes theorem, or whether the probability estimate is more accurate only using the Phase II information. The Demon replies that the estimate is more accurate using only the Phase II information. In this scenario, it would be best to be a frequentist and use only the Phase II information.
- *Scenario 2*. This scenario is similar to Scenario 1, but with one exception. The Demon informs the researcher that the Bayesian estimate, including both the Phase I and Phase II information, is superior to the frequentist estimate, using only the Phase II information. In this scenario, the Bayesian prevails over the frequentist.

- *Scenario 3*. This scenario is similar to Scenario 2, but with one exception. That is, in addition to favouring the Bayesian estimate as more accurate than the frequentist estimate, the Demon also explains that the frequentist definition of probability is correct, and the Bayesian definition of probability is wrong. Thus the Demon elaborates: "I know the proportion of white balls in the urn, and that is the criterion probability I am using to determine whether combining Phase I and Phase II information provides a better or worse estimate than using Phase II information alone. Note that I am using a frequentist probability as the criterion, thereby demonstrating, in my omniscience, that the frequentists are philosophically correct in their conception of probability."

Scenario 3 is extremely interesting. The Demon is, in essence, telling us that Bayesians are philosophically inferior, but that the frequentists are empirically inferior with respect to the case at hand. Does one favour philosophical superiority over empirical superiority, or empirical superiority over philosophical superiority?

The answer here is to keep them separate, which was the reason for introducing the issue via the three scenarios, i.e. it is perfectly consistent to simultaneously hold that (a) the frequentists are philosophically correct and (b) using the philosophically incorrect Bayesian approach nevertheless results in a better probability estimate even by frequentist standards. An elegant characteristic of this approach is that it recognizes philosophical incommensurability while simultaneously asserting that it does not force empirical incommensurability. In turn, the lack of empirical incommensurability leaves it open that researchers might, at times, be better-off using the philosophically wrong approach. The philosophically wrong approach sometimes leads to a better empirical result, and that is that!

5.2. The judgment call

With philosophical and empirical superiority distinguished, what happens when the Demon is no longer available for questioning? Without the Demon it is necessary to make judgment calls, and the sciences would be better off admitting to making judgment calls, without dogmatic assertions that there is only one right way to go. In essence, the judgment calls come down to the issue of the quality of the Phase I and Phase II information, and because the Bayesian posterior probability is influenced by what prior probability distribution is assumed, and what parameters are assumed, the issue might not be frequentist versus Bayesian, but rather frequentist versus Bayesian_1 versus Bayesian_2 versus Bayesian_3 … versus Bayesian_N.[2] In essence, if the quality of the Phase I information is so low, relative to the quality of the Phase II information, that using it decreases the accuracy of the empirical estimate, then it is better to be an empirical frequentist than an empirical Bayesian. However, if the quality of the

[2] There can be multiple frequentist approaches too with respect to assumptions pertaining to different distributions.

Phase I information is not so low, so that using it in conjunction with the Phase II information increases the accuracy of the estimate, then it makes more sense to be an empirical Bayesian than an empirical frequentist. Once again, it is possible to be an empirical Bayesian even while remaining a philosophical frequentist. Hence, at the empirical level, it is a judgment call whether to be an empirical frequentist or an empirical Bayesian, and, within the Bayesian sphere, it is a judgment call exactly as to what kind of empirical Bayesian to be.

However, the conclusion that whether to be an empirical frequentist or an empirical Bayesian, with respect to any particular case at hand, is a judgment call, does not force one to base the decision on a coin toss or one's affective predilections. Here is a place where mathematical or computer simulations can help to play out the consequences of various possibilities. Suppose that simulations, under all possibilities that the researcher considers at least mildly realistic, favour Bayes theorem as resulting in a more accurate estimate than the frequentist computation. In that case, it is not difficult to conclude that a Bayesian approach would be empirically beneficial, although there would remain the issue of which Bayesian approach to use. Naturally, simulations might address the issue of which Bayesian approach to use, as well as whether to be Bayesian at all with respect to the case at hand. In contrast, simulations might indicate that it is unlikely that Bayesian approaches will increase the accuracy of the estimate for the case at hand, which would be a reason for not being an empirical Bayesian. Otherwise, there might be reasons to fear being misled by assumptions about prior distributions, even in cases where there also are reasons to suspect that prior distributions might be helpful. In such relatively ambiguous cases, and depending on the particular research contexts, intelligent and well-meaning researchers could disagree about whether to maximize the expected accuracy which might be improved with a Bayesian approach, or whether to minimize the maximum possible inaccuracy, in which case a frequentist approach might be preferred.

To conclude this section, it sometimes is not possible to completely eliminate judgment calls at the empirical level because the researcher might not have sufficient information to make an unambiguous judgment about how likely a Bayesian estimate, or competing Bayesian estimates, are to seriously mislead the researcher. Different research contexts might call for maximizing expected accuracy, minimizing maximum inaccuracy, or yet other strategies. It would be desirable for scientists and philosophers to simply admit empirical ambiguity and make case-by-case judgments about what is empirically appropriate, and grant researchers the privilege (and also duty) of providing arguments to support why their judgment calls went in one direction or another. It is also possible to make probability estimates using both perspectives, or even including multiple Bayesian perspectives along with a frequentist perspective, so the reader can gain an idea of how different judgment calls lead to different or similar estimates.

6. An argument for Bayesian philosophical superiority

Let us consider again Scenario 3, where the Demon used its omniscient knowledge of the relative frequency of white balls to total balls in the urn to arrive at a frequentist criterion probability. That this was the obviously correct course for the Demon to take, thus far, supports frequentist philosophical superiority. Yet, there is a potential counterargument that can be generated, still assuming that the Demon is correct to use its omniscient relative frequency knowledge to set the criterion probability.

To approach the counterargument, consider the historical context of probability, as documented by Stigler (1990). People have often used probability to indicate that particular events might come about, even in the absence of relative frequency information. It is not difficult to imagine a person living in England during 1587 saying something like, "It is probable that there will be a war with Spain next year."[3] The person would not have evaluated relative frequencies and likely would not even have known about the concept of relative frequencies, but would have been indicating a state of belief based on knowledge of the various tensions between England and Spain at that time. Such knowledge might include assumptions about Phillip II's desire to add England to his empire, the provocative effect of English privateering, impatience at Elizabeth I's political games, and others. The lack of relative frequency knowledge need not prevent the person from making a shrewd estimate of a likely war. This example highlights that probability is best considered to be a state of belief.

Where does relative frequency apply? A Bayesian could argue that when quality frequency information is available, it is rational to base one's belief upon it. Thus, a Bayesian could agree that the Demon's knowledge of the relative frequency of white balls in the urn provides the best basis for a criterion probability, while at the same time insisting that the underlying reason for caring about the relative frequency information is the desire to have the most rational state of belief possible. From this perspective, it is obvious that relative frequency information, if available, should be used, but this is a special case of the more basic, and more general, point that the crucial question to be asked is, "What should one believe?" Having asked that question, Bayes theorem, with subjective priors, provides a clear answer. In fact it is even possible to generate an argument that, in the absence of relative frequency information in Phase I and Phase II, Bayes theorem could still be used to suggest the implications of subjective beliefs pertaining to both phases. In any event, the Demon's preference for relative frequency information does not indicate the philosophical superiority of the frequentist probability concept, only that the frequentist probability concept is a particularly precise instantiation of the more general Bayesian probability. Thus, a Bayesian accusation could be that frequentists confuse precision of information with philosophical superiority. That the relative

[3] Spain attempted an invasion of England in 1588 that was repulsed by the English navy.

frequency information is precise, renders it preferable to less precise information in those cases where it is possible to instantiate relative frequency information into Bayes theorem. Yet the bottom-line philosophical concept of a probability as a state of belief still holds.

6.1. Philosophical and empirical superiority and the rationality problem

It was shown earlier that under an assumption of frequentist philosophical superiority, it is nevertheless possible to have particular cases of Bayesian empirical superiority. By invoking Laplace's omniscient Demon, it was possible to know, in particular cases, that a frequentist or Bayesian estimate was closer to the relative frequency of white balls in the urn. It would be convenient now to reverse matters and assume Bayesian philosophical superiority and use the Demon to nevertheless find cases of frequentist empirical superiority, however this leads to a problem.

By insisting on the Bayesian concept that a probability is a state of belief, there is no longer an absolute standard that the Demon can use to judge frequentist or Bayesian empirical estimates. Although the Demon can assert its knowledge of the proportion of white balls in the urn, the Demon would also have to admit that different people, with different knowledge or lack thereof, would be justified in asserting different values. There is a vagueness here that can be dramatized by considering that (a) from a frequentist point of view, the relative frequency of white balls in the urn **is** the probability and (b) from a Bayesian perspective the relative frequency of white balls in the urn merely engenders one's belief state with more **specificity**. A frequentist would insist that the Demon's value for the criterion probability is de facto correct whereas it is not clear what a Bayesian would insist. One Bayesian might insist that the Demon's value is merely precise, but it is not correct in any absolute sense because there is no absolute standard of correctness, however this argument leads in an uneasy direction. For example, it opens the door for a frequentist to sarcastically assert: "Well, if you are going to insist that there is no absolute standard of correctness, why use Bayes theorem at all? Why not just have people move directly to their belief states regarding posterior probabilities?" The Bayesian might answer by bringing up standards of rationality, but these are difficult to defend. Even the assertion that there is something good about having the Demon's precise relative frequency information might be difficult to justify when there is a lack of absolute standards, as a frequentist could query how the Bayesian defines that which is good in a subjective context.

Another Bayesian might argue that the Demon's value is correct, and so there is an objective standard after all, but this seems tantamount to accepting the frequentist definition of probability and to go against a potential Bayesian assertion that the frequentists are confusing specificity with philosophical correctness.

Matters become worse when considering the scenario involving potential war between England and Spain, where it is not clear that probability even has a meaning

in the context of absolute knowledge. Consider that in 1587, the omniscient Demon knows there will be war in 1588, and so the Demon's prior probability assignment of war would be 1.00. Yet, most Bayesians consider probability assignments of 1 or 0 to be irrational because such an assignment of prior probabilities leaves no room for modification in light of posterior information.[4] For example, it is easy to imagine a person insisting that the prior probability of the existence of God equals 1.00, and thereby declaring that there is no need to consider any other information. A Bayesian would insist that such a prior probability assignment is irrational because it allows no room to change one's mind, but such an insistence denies the Demon's probability assignment of 1.00 in 1587 to war in 1588.

6.2. Straightforwardness and implications

The previous section clarifies that a Bayesian insistence, at the philosophical level, implies difficulties. The Bayesian probability concept is not straightforward and implies a necessity to bring in outside arguments about what it means to be rational. There is no intellectually honest way to play down the lack of straightforwardness. There are implications for possible philosophical avenues.

One avenue is to be a philosophical frequentist while admitting that Bayes may be empirically superior on particular occasions. This was the tack taken in a previous section. Another possibility is to insist on being a philosophical Bayesian and take on the associated difficulties regarding the lack of objective standards and having to demarcate that which is rational from that which is irrational. This includes addressing squarely the issue of what justifies a prior probability of 1 or 0, or why there is no justification. Either argument is susceptible to rejoinders. If prior probabilities of 1 or 0 are deemed acceptable, then the Bayesian admits that there is no possibility of updating, and probability updating is usually considered the Bayesian glory, or if prior probabilities of 1 or 0 are not deemed acceptable, then it is not clear why not. Why might it be rational to have a prior probability of 0.9999999999999999999999999999999 but not to have a prior probability of 1?

A further philosophical difficulty is that there are many more ways to be a Bayesian than to be a frequentist. Thus, disagreements among Bayesians are much more frequent than disagreements among frequentists. An argument against using a Bayesian approach is that commencing with a probability concept that leads to more disagreement is a bad idea when it is possible to commence with one that leads to less disagreement. A possible Bayesian rejoinder might be that such disagreement

---

[4] To see that this is true, consider Bayes theorem in its simplest form: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$. If we instantiate 1.00 for $P(A)$, we have the following: $P(A|B) = \frac{(1.00)\ P(B|A)}{P(B)} = 1.00$, as $P(B|A) = P(B)$ in the special case where $P(A) = 1.00$. Or, if we use 0, we would conclude that $P(A|B) = 0$. Either way, it is impossible to update the prior probability with posterior information.

is the price to be paid for having a more powerful probability conception and besides, disagreement can be interpreted as indicating dynamism.

The issue of what to believe might be less fundamental than Bayesians admit. Consider that there was a time prior to the evolution of humans, and without them, human beliefs were irrelevant. However, quantum probabilities and entropic probabilities nevertheless functioned,[5] thereby suggesting that it is possible to parse probabilities from human beliefs. From this point of view, calibration of degrees of beliefs might remain important; but is distinguishable from probabilities that are 'out there', thereby indicating that conceptions of probability should not be subservient to concerns about belief calibration. However, if one is a determinist, a rejoinder might be that there were no probabilities in the absence of humans and their imperfect knowledge states and computational limitations. Thus, if one is a determinist, it makes sense to be a Bayesian because it is only imperfect human knowledge or processing capacity that contributes to uncertainty. Hence probabilities can only refer to resulting subjective beliefs, whereas, if one is an indeterminist, probability is separable from human knowledge, and so a Bayesian concept of belief is insufficient. Put another way, it would make sense for an indeterminist to maintain an ontological distinction between epistemic estimates–estimates of one's state of knowledge – and probabilities.

The remainder of this article will be agnostic to the issue of philosophical superiority of probability concepts because of the difficulty in handling issues regarding rationality and determinism. The following two sections concern sample size estimation and hypothesis testing, respectively.

7. Sample size estimation

In the slipped disk example, the goal was to estimate a conditional probability: What is the probability of a slipped disk given a positive MRI? In the urn example, the goal was to estimate a population proportion or unconditional probability: What is the proportion of white balls in the urn? Yet these are not the only possibilities. Researchers might be interested in estimating other parameters such as the population mean, standard deviation, shape, kurtosis, and others, but they might be also interested in estimating sample sizes needed to obtain satisfactory estimates of population parameters.

Recently, a procedure was developed for sample size estimation, termed the a priori procedure (APP). As the present subject is a philosophical discourse, APP equations are not presented, but Trafimow (2019a) provided a review, and Li et al. (2020) created user-friendly computer programs for making APP calculations. The APP researcher makes two specifications.

[5] Baggott (2015) has provided quantum and entropic histories of the universe.

- Precision: How close do I want my sample statistics to be to their corresponding population parameters?
- Confidence: What probability do I want to have of meeting the precision specification?

Given the specifications for precision and confidence, APP equations or computer programs can be used to find required sample sizes for meeting the specifications under a variety of assumptions. For example, if the distribution is assumed normal, and the researcher desires to have 95% confidence that the sample mean will be within one-tenth of a standard deviation of the population mean, at least 385 participants are required. Trafimow and Myüz (2019) showed that parameter estimates in five areas of psychology are quite imprecise; and Trafimow, Hyman, and Kostyk (2020) provided a similar demonstration in the marketing field.

Just as it is possible to perform frequentist or Bayesian estimation with respect to population parameters, it also is possible to perform frequentist or Bayesian estimation with respect to the sample sizes needed to obtain good estimates of population parameters; but what constitutes good sample size estimates?

From the perspective of scientific conservatism, it is possible to argue that the largest estimate is the best one. That is, ceteris paribus, collecting a larger sample renders better parameter estimation than collecting a smaller sample. Alternatively, from the point of view that participants are expensive, it is possible to argue that the smallest estimate is the best one because it is most feasible. Therefore, it is possible to query whether frequentist or Bayesian procedures result in larger or smaller sizes of samples, and also whether it is good or bad to have larger or smaller samples.

Wei et al. (Wei, Wang, Trafimow, and Talordphop, 2020) provided a nice demonstration of the Bayesian possibilities. These researchers used mathematical and computer simulations to show that under good assumptions about prior probability distributions, the resulting Bayesian sample size designations are smaller than corresponding frequentist sample size designations. In essence, the availability of prior information confers an advantage to the researcher, thereby reducing the sample sizes necessary to reach the same levels of precision and confidence. Again, we can see that philosophical incommensurability need not engender empirical incommensurability. Obviously, as Wei et al. performed the simulations, they were able to employ user-defined parameters. In the practice of substantive science, these are unknown, and it is not clear, without data, whether a Bayesian or frequentist APP is better for particular substantive research projects. This may well depend on the idiosyncrasies of substantive research projects. As usual, within the Bayesian sphere, there are the issues of which prior distributions to use and how to set the parameters of those distributions.

Before concluding this section, it is worth noting that there are alternative frequentist procedures as well as alternative Bayesian ones. Even frequentists might disagree with each other on whether the necessary sample size to meet precision and confidence specifications should be based on assuming normality, skew

normality, lognormality, uniformity, or others. As usual, however, there is more room for Bayesians to disagree with each other because they can disagree about prior distributions too.

8. Hypothesis testing

Undoubtedly, the most contentious aspect of the frequentist-Bayesian debate concerns hypothesis tests (e.g. Briggs, 2016; 2019; Cohen, 1994; 1997; Greenland, 2017; Halsey, Curran-Everett, Vowler, and Drummond, 2015; Hubbard, 2016; McShane et al., 2018; Nickerson, 2000; Trafimow, 2003, 2019b; Trafimow and Marks, 2015; 2016; Wasserstein and Lazar, 2016; Valentine, Aloe, and Lau, 2015; Ziliak and McCloskey, 2016). Frequentists cannot calculate the probability of a hypothesis. For a frequentist, the probability of a hypothesis is meaningless as the hypothesis is right or wrong, though the researcher might not know which. For a Bayesian, hypotheses have meaningful probabilities and it is sensible to perform Bayesian calculations to determine those probabilities. Hence, one can see again philosophical incommensurability at work.

Even some Bayesians feel uncomfortable with probabilities of hypotheses, but Bayes factors can be used to index how much more likely the data are under one hypothesis than another. Let us consider all these issues.

8.1. Dealing with hypotheses based on frequentist thinking

There are two main frequentist modes for dealing with hypotheses, and neither assumes that it is sensible to compute probabilities for a hypothesis. One method is to compute a *p*-value, based on a hypothesis, and conceptualize the *p*-value as indicating the incompatibility of the data with the hypothesis. The other method is to engage in error control, where the researcher sets a threshold level, and rejects or does not reject the hypothesis depending on whether the *p*-value is under threshold or not. The idea is that if one sets the threshold, say, at 5%, then the researcher will wrongly reject hypotheses at a 5% rate or less. Thus, 'error control' is achieved at the threshold level that the researcher sets.

However, based on the distinction between estimation and hypothesis testing, both methods can be argued to be problematic. To see the distinction, consider that when the researcher is using a sample statistic to estimate a corresponding population parameter, or even to estimate sample sizes required to meet precision and confidence specifications, there is no need for all the researcher's assumptions to be exactly correct. On the contrary, the expectation is that the researcher will not be exactly correct but might be close enough to correctness for the estimation to be useful. However, matters are different for hypothesis testing, as explained below.

To start, let us consider that there are two classes of entities necessary to engage in computations involving hypotheses. There is the hypothesis itself. Secondly,

however, there are additional assumptions. These can be distributional assumptions, the ubiquitous assumption of random selection from the population (Berk and Freedman, 2003), and many others. There are so many additional assumptions that Bradley and Brand (2016) and Trafimow (2019b) proposed taxonomies of these assumptions. One can use the term *model* to refer to the conjunction of the hypothesis and the additional assumptions. As it is tantamount to an impossibility that all model assumptions are exactly correct, it should be clear that the model is wrong. One might counter, however, that the model is no more 'wrong' in a hypothesis testing context than in an estimation context, so there is no reason to make such a big deal about model 'wrongness' in a hypothesis testing context.

However, there are two problems with the counterargument. One problem is that in an estimation context, the model need not contain a hypothesis, whereas it must contain a hypothesis in a hypothesis testing context. Thus, there necessarily is more to the model in a hypothesis testing context than in an estimation context, so there is more that can go wrong with the model in a hypothesis testing context than in an estimation context. There is also a second, and more important problem with the counterargument. Specifically, in an estimation context one assumes the model is wrong but hopes that it is sufficiently close to correct to render the model useful, but in a hypothesis testing context, there are only two possibilities: the hypothesis is correct or incorrect; there is no such thing as being close. Yet because the hypothesis is embedded in a known to be wrong model, it should be clear that whatever the status of the hypothesis, the model is wrong either way.

To address the consequences of knowing that the model is wrong, whatever the status of the hypothesis, consider that something stated earlier – that *p*-values are based on hypotheses – is not true, but rather that *p*-values are based on models. A *p*-value does not index the incompatibility of the data with the hypothesis, instead a *p*-value indexes the incompatibility of the data with the model. Therefore, if a researcher obtains an impressively low *p*-value, it is not clear whether the incompatibility is due to the hypothesis being wrong, one of the additional assumptions in the model being wrong, or due to multiple problems. The only sound conclusion that can be drawn is that the data are incompatible with the model. This is of very little value because the model is known to be wrong from the start. To ask a pointed and rhetorical question: "Why index evidence against a model that is already known to be wrong?"

Are matters improved by moving to an error control concept? Not only is the answer in the negative, but error control increases the problems. In the first place, again remembering that the *p*-value is based on the model rather than on just the hypothesis; the best that can be obtained is error control with respect to models, not with respect to hypotheses. However there is no point in error control at the model level because all models are known to be wrong. Therefore, if the *p*-value comes in under the threshold, and hence the researcher rejects the model, nothing is gained because the model is already known to be wrong. In contrast, if the *p*-value does not come in under the threshold, the researcher fails to reject the already known

wrong model, which is deleterious. Hence, the best outcome that can be obtained is no gain, the worst outcome that can be obtained is negative, and so the expected utility of error control with respect to models is negative. To ensure that there is no misunderstanding, it is worth reiterating that there is no sound way to perform error control at the level of hypotheses because hypotheses are embedded in known wrong models.

Yet the negative expected utility of error control is even worse than explained in the previous paragraph, which becomes clear upon consideration of what is required to obtain *p*-values under the threshold. There are two factors: sample size and sample effect size. As the model is always wrong, one can practically guarantee a statistically significant *p*-value, merely given a sufficiently large sample size. However, it is possible to make a more interesting argument based on the fact that if a researcher were to collect a large number of moderately sized samples from a population, the sample effect sizes would vary from sample to sample due to the nature of random sampling. As *p*-values are based, in part, on sample effect sizes, they too vary from sample to sample – the so-called dance of the *p*-values. As statistically significant findings – findings where the *p*-values are under threshold – are overwhelmingly more likely to be published than statistically insignificant findings, insisting on a threshold level sets the process of regression to the mean into motion. It is largely a matter of luck whether a particular researcher's experiment results in a sufficiently large sample effect size for statistical significance, and so the scientific literatures contain many lucky findings, i.e. inflated sample effect sizes. Following replication attempts, getting lucky is not expected to be often replicated successfully, and thus sample effect sizes in replication studies tend to regress in the direction of the mean. This is not an esoteric mathematical point, either. The Open Science Foundation (2015) examined approximately 100 studies in top psychology journals and found that the average sample effect size in the replication cohort was less than half the average sample effect size in the original cohort. That error control leads to dramatically inflated effect sizes in scientific literatures adds importantly to negative expected utility (see Grice, 2017; Hyman, 2017; Kline, 2017; Locascio, 2017a; 2017b; Marks, 2017 for further discussion).

8.2. Dealing with hypotheses based on Bayesian thinking

That there is no sound way to use frequentist thinking to decide what hypotheses to reject or not reject may seem a compelling reason to go Bayesian. Furthermore, an advantage of Bayesian thinking, in the context of hypotheses, is that because hypotheses are assumed to have probabilities between zero and one, it is possible to integrate estimation and hypothesis testing in a way that is impossible under frequentist thinking, in which the hypothesis is either correct or incorrect, and there is nothing to estimate; one can only be correct or incorrect. It is this dichotomy that is at the bottom of why frequentist thinking does not work in a hypothesis

testing context. Yet a Bayesian could reasonably assume that although the model is wrong, it might be close enough to being correct that the estimated probability of the hypothesis is reasonably close to the true probability of the hypothesis. As stated earlier, a wrong model need not be fatal for estimation whereas it is always fatal for hypothesis testing.

Still, some sobering considerations temper the optimism of the previous paragraph. The first consideration harks back to the issue of what probability is. Imagine asking Laplace's omniscient Demon about the hypothesis that the difference in means equals zero. Would the Demon assign an intermediate number, such as 0.65, or would the Demon assign an extreme number, such as 0 or 1? It seems obvious that because the Demon is omniscient, it would assign 0 to indicate that the hypothesis is definitely wrong, or 1 to indicate that the hypothesis is definitely right. Moreover, suppose the Demon did come back with an answer of 0.65. How would we interpret that number? One interpretation is that 0.65 indicates the Demon's uncertainty, but how can an omniscient Demon be uncertain? An alternative interpretation is that 65% of hypotheses are correct, and so the Demon is interpreting the question as asking for the relative frequency of correct hypotheses, but such an interpretation is no help whatsoever in drawing a conclusion about the present hypothesis. A third interpretation is that 65% of the present researcher's hypotheses are correct, but again, this is not useful for drawing conclusions about the present hypothesis. It is possible to invent yet more interpretations for an answer of 0.65, but these interpretations are not useful. Thus, the Demon's answer of 0 or 1 contradicts the notion of Bayesian probability, and an intermediate answer is difficult to interpret. Either way, there are serious problems with going Bayesian.

That is not all. Even from a computational perspective, it is not clear what to do. How does one assign a prior probability for the hypothesis? How does one assign a probability of the finding given that the hypothesis is not true? Regarding the latter, how does one sum up the probabilities of the findings given all possible hypotheses that are not the hypothesis under investigation?

A Bayesian way out from under these difficulties is to focus on Bayes factors. The idea here is to compute a Bayes factor that indexes the relative conditional probabilities of data given one hypothesis or a competing hypothesis. If the data are much more likely given Hypothesis 1 than given Hypothesis 2, then that is considered a strong reason for favouring Hypothesis 1 over Hypothesis 2.

However, the previous paragraph is unfortunately a misstatement. The truth of the matter is that, just as we saw in the frequentist section with *p*-values, Bayes factors are based on models, not just on hypotheses. Thus, what Bayes factors index is not relative support for one hypothesis against another; but rather relative support for one model against another. This may seem like a too-subtle distinction but consider again that both models are wrong. To ask a pointed and rhetorical question: "What is gained by saying that the data are more likely given one wrong model than given another wrong model?"

Worse yet, both models might be quite unimpressive. For example, suppose the probability of the data given Model 1 is 0.01 and the probability of the data given Model 2 is 0.001. The resulting Bayes factor would be 10, which is considered by most Bayesians to be a respectable value in favour of the hypothesis embedded in Model 1, despite the fact that both models perform badly when considered individually.

The author could now admit painting himself into a corner, as frequentist hypothesis testing and Bayesian hypothesis testing are unsound. What is left? The next subsection addresses that question.

8.3. Returning to estimation as an initial stage in hypothesis testing

From the point of view of the statistician who insists on using inferential statistics for hypothesis testing, one remains with a terrible problem that both frequentist and Bayesian hypothesis testing are unsound. Fortunately, there is an alternative potential point of view that although hypothesis testing is important for scientists, the exercise cannot be carried out in the absence of the substantive researcher's expert judgment. From this point of view, the previous century when statisticians have insisted on being able to perform hypothesis tests via automated procedures are weeds on the lawn of science. There should have been much more of a focus on estimation, and this is the direction future researchers should take.

It is worth taking a brief detour invoking the Demon, but this time with respect to estimation. Suppose the Demon informed us that all our sample statistics, in all our studies, have absolutely nothing to do with corresponding population parameters. This would bring on the most dramatic crisis in the history of science and nobody would have confidence in our ability to use such data to test hypotheses. The detour renders salient the necessity of the tacit assumption that sample statistics are reasonable indicators of corresponding population parameters, as a prerequisite for hypothesis testing.

For instance, suppose that a researcher hypothesizes, based on a theory, that the difference between two means should be 27.00. To be sure, this is a point prediction that is much more precise than is typical, at least in the social sciences, but let us continue anyhow. Suppose that the researcher obtains a difference of 35.83. Does this finding confirm or disconfirm the hypothesis? As research is performed today, the typical researcher would perform a significance test, and if the obtained difference of 35.83 is statistically significantly different from the hypothesized value of 27.00, the conclusion would be that the data contradict the hypothesis. We have already seen this is problematic for multiple reasons, including that the low p-value could be due to model wrongness rather than hypothesis wrongness.

An alternative approach would be to consider two stages, namely an estimation stage and a hypothesis testing stage. In the estimation stage, the researcher or statistician would estimate whether the sample size used meets specified criteria for precision and confidence, and we have already seen that the APP, whether used in

a frequentist or Bayesian way, can aid the researcher in proceeding in this direction. If the sample size used is deemed sufficient to engender confidence that the sample difference in means is close to the population difference in means, the researcher can move on to the hypothesis testing stage. Obviously, the researcher would need to consider issues such as whether the model is 'good enough for government work'.

In the hypothesis testing stage, there is increased subjectivity as there are questions that require impressive substantive expertise. A few of the many possible questions are bullet-listed below.

- What is the range of possible mean differences that could be considered consistent with the hypothesis, inconsistent with the hypothesis, or somewhere between?
- Are there alternative hypotheses that do a better, or worse, job of accounting for the obtained mean difference?
- What is the quality of the added (auxiliary) assumptions used to reduce the theory to an empirical hypothesis?
- What is the quality of the assumptions used to reduce the empirical hypothesis to a statistical hypothesis?

It is important not to exaggerate the difference between the estimation and hypothesis testing stages. Both the statistician and the substantive researcher have roles to play in both stages. Nevertheless, it should be clear that the substantive role is more pronounced in the hypothesis testing stage.

9. Conclusion

The subject of this paper is that philosophical incommensurability, that is, conflicting probability conceptions do not force empirical incommensurability. It is not contradictory for a person to be a frequentist at the philosophical level and nevertheless also be an empirical Bayesian for those cases where Bayes theorem is expected to improve estimation accuracy. It also is not contradictory for a philosophical Bayesian to insist on empirical frequentist probabilities when high quality relative frequency information is available. Thus, there is room for both frequentist and Bayesian estimation, regardless of whether the researcher wishes to estimate population parameters or sample sizes required to meet specifications for precision and confidence. In contrast, we have seen that neither frequentist nor Bayesian methods work well for hypothesis testing. This is because statistical hypotheses are always embedded in known wrong models. In an estimation context, that the model is wrong need not be fatal provided that it is good enough to provide a reasonable estimate. Yet in a hypothesis testing context, there is no such thing as 'good enough', there is only 'right or wrong'. Consequently, there can be no automated hypothesis testing procedures that the statistician can perform soundly, in the absence of crucial input from the substantive expert. The substantive expert must dominate and the statistician, though useful, should be the assistant and not the master. Hypothesis testing is necessarily a much more substantive process than

is estimation, and a much more substantive process than statisticians have admitted over the last century.

Lest the reader feel uncomfortable with the philosophical theme that higher level philosophical incommensurability does not force lower level empirical incommensurability, consider that a similar phenomenon occurs in other domains. For example, a foundational Kuhnian premise is that different scientific theories are incommensurable and therefore cannot be tested against each other as a Popperian might wish to do. For example, *mass* has a different meaning for Newton and Einstein, they are speaking different languages, and so there is no way to perform a definitive test of the two theories. However, consistent with the present theme, incommensurability at the theoretical level does not imply incommensurability at the empirical level (Godfrey-Smith, 2003; Trafimow, 2020). Two researchers who disagree about how to conceptualize mass nevertheless can agree on a clock reading. If Newton and Einstein made different predictions about the clock reading, they could be tested against each other despite incommensurability at the theoretical level.[6]

In conclusion, although frequentist and Bayesian probability conceptions are incommensurable – frequentists and Bayesians are not speaking the same language – philosophical incommensurability does not force empirical incommensurability. It is possible for a philosophical frequentist to be a practical Bayesian for particular cases where Bayesian estimation is likely to be superior to frequentist estimation. This is not naïve call for frequentists and Bayesians to get along, they should not get along at the philosophical level any more than scientists following Newton's characterization of *mass* should get along with scientists in favour of Einstein's characterization of *mass*. Yet frequentists and Bayesians can, and should, get along at the empirical level, just as scientists in theoretical opposition can agree on empirical clock readings. Hence, although there is no room in a single philosophical tent for both frequentists and Bayesians, the empirical tent is much larger, and does contain sufficient room. Empirical commensurability, even in the presence of higher level philosophical or theoretical incommensurability, entails consequences.

References

Andersson, G. B. (1991). Epidemiology of spinal disorders. In J. Frymoyer (Ed.), *The adult spine: Principles and practice* (pp. 107-146). New York, NY: Raven Press.

Baggott, J. (2015). *Origins: The scientific story of creation*. Oxford, UK: Oxford University Press.

[6] A similar point can be made with respect to two major schools of moral philosophy. Deontological and utilitarian perspectives are often considered to be incommensurable. However, a rule utilitarian could argue that the utilitarian calculations are too difficult to make, on a routine basis, so that at the level of actual behaviour, acting in a deontological manner is the best way to approach the utilitarian ideal of the best for the largest number of people.

Berk, R. A., and Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg and S. Cohen (Eds). *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (2nd Ed, pp. 235-254). Aldine de Gruyter.

Boos, N., Rieder, R., Schade, V., Spratt, K. F., Semmer, N., and Aebi, M. (1995). The diagnostic accuracy of magnetic resonance imaging, work perception, and psychosocial factors in identifying symptomatic disc herniations. *Spine, 20*, 2613-2625. http://dx.doi.org/10.1097/00007632-199512150-00002

Borenstein, D. G., Wiesel, D. G., and Boden, S. D. (1995). *Low back pain-medical diagnosis and comprehensive management* (2nd ed.). Philadelphia, PA: WB Saunders.

Box, G. E. P., and Draper, N. R. (1987). *Empirical model-building and response surfaces.* New York*:* John Wiley & Sons*.*

Bradley, M. T., and Brand, A. (2016). Significance testing needs a taxonomy: Or how the Fisher, Neyman-Pearson controversy resulted in the inferential tail wagging the measurement dog. *Psychological Reports*, *119*(2), 487-504. doi: 10.1177/0033294116662659

Briggs, W. (2016). *Uncertainty: The soul of modeling, probability and statistics*. New York: Springer.

Briggs, W. (2019). Everything wrong with p-values under one roof. In V. Kreinovich, N. N. Thach, N. D. Trung, and D. Van Thanh (Eds.), *Beyond traditional probabilistic methods in econometrics* (pp. 22-44). Cham, Switzerland: Springer.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997-1003. doi: 10.1037/0003-066X.49.12.997

Cohen, J. (1997). The earth is round (p < .05). In L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Eds.), *What if there were no significance tests*? (pp. 21-36). Hillsdale, NJ: Erlbaum.

Frymoyer, J. W. (1988). Epidemiology. In J. W. Frymoyer, and S. L. Gordon (Eds.), *New perspectives on low back pain* (pp. 19-33) (Symposium, Workshop, Airlie). Chicago, IL: American Academy of Orthopedic Surgeons.

Gillies, D. (2000). *Philosophical theories of probability*. London: Taylor and Francis.

Godfrey-Smith P. (2003). *Theory and reality: An introduction to the philosophy of science*. Chicago: The University of Chicago Press.

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology*, 186, 639-645. doi: 10.1093/aje/kwx259

Grice, J. W. (2017). Comment on Locascio's results blind manuscript evaluation proposal. *Basic and Applied Social Psychology*, *39*(5), 254-255. https://doi.org/10.1080/01973533.2017.1352505

Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015). The fickle *P* value generates irreproducible results. *Nature Methods*, *12*, 179-185. doi:10.1038/nmeth.3288

Hubbard, R. (2016). *Corrupt research: The case for reconceptualizing empirical management and social science*. Los Angeles, California: Sage Publications.

Hyman, M. (2017). Can 'results blind manuscript evaluation' assuage 'publication bias'? *Basic and Applied Social Psychology*, *39*(5), 247-251. https://doi.org/10.1080/01973533.2017.1350581

Jensen, M. C., Brant-Zawadzki, M. N., Obuchowski, N., Modic, M. T., Malkasian, D., and Ross, J. S. (1994). Magnetic resonance imaging of the lumbar spine in people without back pain. *New England Journal of Medicine, 331*, 69-73. http://dx.doi.org/10.1056/NEJM199407143310201

Lawrence, R. C., Feseon, D. T., Helmick, C. G., Arnold, L. M., Choi, H., Deyo, R. A., … Wolfe, F. (2008). Estimates of the prevalence of arthritis and other rheumatic conditions in the United States Part II. *Arthritis & Rheumatism*, *58*, 26-35.

Li, H., Trafimow, D., Wang, T., Wang, C., and Hu, L. (2020). User-friendly computer programs so econometricians can run the a priori procedure. *Frontiers in Management and Business*, *1*(1), 2-6. doi: 10.25082/FMB.2020.01.002

Kline, R. (2017). Comment on Locascio, results blind science publishing. *Basic and Applied Social Psychology*, *39*(5), 256-257. https://doi.org/10.1080/01973533.2017.1355308

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge, England: Cam-

bridge University Press.
Locascio, J. (2017a). Results blind publishing. *Basic and Applied Social Psychology*. *39*(5), 239-246. https://doi.org/10.1080/01973533.2017.1336093
Locascio, J. (2017b). Rejoinder to responses to "results blind publishing." *Basic and Applied Social Psychology*. *39*(5), 258-261. https://doi.org/10.1080/01973533.2017.1356305
Marks, M. J. (2017). Commentary on Locascio 2017. *Basic and Applied Social Psychology*. *39*(5), 252--253. https://doi.org/10.1080/01973533.2017.1350580
McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2018). Abandon statistical significance. arXiv preprint arXiv:1709.07588.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241-301. doi: I0.1037//1082-989X.S.2.241
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716
Popper, K. R. (1983). *Realism and the aim of science*. London: Routledge.
Stigler, S. M. (1990). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.
Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, *110*(3), 526-535. doi: 10.1037/0033-295X.110.3.526
Trafimow, D. (2006). Using epistemic ratios to evaluate hypotheses: An imprecision penalty for imprecise hypotheses. *Genetic, Social, and General Psychology Monographs*, *132*, 431-462. doi: 10.3200/MONO.132.4.431-462
Trafimow, D. (2019a). A frequentist alternative to significance testing, p-values, and confidence intervals. *Econometrics*, *7*(2), 1-14. https://www.mdpi.com/2225-1146/7/2/26
Trafimow, D. (2019b). A taxonomy of model assumptions on which P is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*, *22*(6), 571--583. doi: 10.1080/13645579.2019.1610592
Trafimow, D. (2020). Our intellectual children: Kuhnian ants or Feyerabendian questioners? *Advances in Educational Research and Evaluation*, *1*(2), 88-92. doi: 10.25082/AERE.2020.02.005
Trafimow, D., and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1-2. http://www.tandfonline.com/doi/full/10.1080/01973533.2015.1012991
Trafimow, D., and Marks, M. (2016). Editorial. *Basic and Applied Social Psychology*, *38*(1), 1-2. doi: 10.1080/01973533.2016.1141030
Trafimow, D., and Trafimow, J. H. (2016). The shocking implications of Bayes' theorem for diagnosing herniated nucleus pulposus based on MRI scans. *Cogent Medicine*, *3*: 1133270. doi: 10.1080/2331205X.2015.1133270
Valentine, J. C., Aloe, A. M., and Lau, T. S. (2015). Life after NHST: How to describe your data without "p-ing" everywhere. *Basic and Applied Social Psychology*, *37*(5), 260-273. http://dx.doi.org/10.1080/01973533.2015.1060240
Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2). doi:10.1080/00031305.2016.1154108
Wei, Z., Wang, T., Trafimow, D., and Talordphop, K. (2020). Extending the a priori procedure to normal Bayes models. *International Journal of Intelligent Technologies and Applied Statistics*, *13*(2), 169--183. doi: 10.6148/IJITAS.202006_13(2).0004
Ziliak, S. T., and McCloskey, D. N. (2016). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, Michigan: The University of Michigan Press.

ZWOLENNICY CZĘSTOŚCI A ZWOLENNICY PODEJŚCIA BAYESOWSKIEGO. SPÓR O NIEWSPÓŁMIERNOŚĆ W ZNACZENIU FILOZOFICZNYM I EMPIRYCZNYM

Streszczenie: Zwolennicy częstości i podejścia bayesowskiego nie zgadzają się co do rzetelności wykonywania obliczeń opartych w istotnej części na wcześniejszych informacjach. Niezgoda powraca do podstawowego sporu filozoficznego dotyczącego tego, jak określać znaczenie prawdopodobieństwa. Ponieważ obydwie grupy naukowców używają tego terminu w inny sposób, powstaje kluczowa filozoficzna niewspółmierność. Jednak w efekcie nie musi ona oznaczać empirycznej niewspółmierności. Możliwe jest, że jednocześnie może istnieć niewspółmierność filozoficzna z empiryczną współmiernością. W artykule omówiono konsekwencje przedstawionej sytuacji.

Słowa kluczowe: niewspółmierność, współmierność, procedura *a priori*, estymacja, testowanie hipotez, wskaźniki bayesowskie.