Science & Philosophy Volume 7(1), 2019, pp. 3-20

Why is Bayesian Confirmation Theory rarely Practiced?

Robert W.P. Luk*

Abstract

Bayesian confirmation theory is a leading theory to decide the confirmation/refutation of a hypothesis based on probability calculus. While it may be much discussed in philosophy of science, is it actually practiced in terms of hypothesis testing by scientists? Since the assignment of some of the probabilities in the theory is open to debate and the risk of making the wrong decision is unknown, many scientists do not use the theory in hypothesis testing. Instead, they use alternative statistical tests that can measure the risk or the reliability in decision making, circumventing some of the theoretical problems in practice. Therefore, the theory is not very popular in hypothesis testing among scientists at present. However, there are some proponents of Bayesian hypothesis testing, and software packages are made available to accelerate utilization by scientists. Time will tell whether Bayesian confirmation theory can become both a leading theory and a widely practiced method. In addition, this theory can be used to model the (degree of) belief of scientists when testing hypotheses.
Keywords: Bayesian confirmation theory; hypothesis testing; induction problem; probability modeling.
2010 AMS subject classifications: 62F03;62F15. [1]

*Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong; csrluk@comp.polyu.edu.hk

[1]Received on January 24th, 2019. Accepted on June 17th, 2019. Published on June 30th, 2019. doi: 10.23756/sp.v7i1.449. ISSN 2282-7757; eISSN 2282-7765.

1 Introduction

In the philosophy of science, Bayesian confirmation theory is one of the leading theories to decide the confirmation or refutation of a hypothesis based on probability calculus. The theory has supporters who try to rescue (e.g., [Schippers and Schurz, 2018]) it from challenges (e.g., [Chihara, 1987],[Wayne, 1995], [Shaffer, 2001], [Huber, 2005] and [Brössel and Huber, 2015]) or who try to extend (e.g., [Myrvold, 2003], [Crupi et al., 2008], [Henderson et al., 2010] and [Festa and Cevolani, 2017]) it for greater generality and applicability. Dawid [Castelvecchi, 2015] noted that it may be used to test whether string theory is science. Norton [2011] enumerated three theoretical advantages of Bayesian confirmation theory as follows:

> "First, the theory reduces the often nebulous notion of a logic of induction to a single, unambiguous calculus, the probability calculus. Second, the theory has proven to be spacious, with a remarkable ability to absorb, systematize and vindicate what elsewhere appear as independent evidential truisms. Third is its most important virtue, an assurance of consistency. The larger our compass, the more we must digest evidence of diverse form and we must do it consistently. Most accounts of evidence provide no assurance of consistency in their treatment of larger bodies of evidence."

Given many theoretical advantages of Bayesian confirmation theory, one would have expected that many scientists apply it for making decisions to accept/reject hypotheses. However, a casual sampling of scientific research articles (e.g., in Nature and Science journals) reveals that almost all such articles did not use Bayesian confirmation theory for hypothesis testing at present. Therefore, why is Bayesian confirmation theory rarely practiced by scientists for hypothesis testing?

2 The practical problem with the Bayesian confirmation theory

To confirm a theory in Bayesian confirmation theory, it is often required that the conditional probability, $P(H|E)$, of the hypothesis H happening given the evidence E should be larger than the prior probability of the hypothesis H without any evidence, i.e., $P(H|E) > P(H)$ where $P(\cdot)$ is the probability. This requirement is based on the notion that the scientist belief in the hypothesis H is revised with more degree of belief after seeing the evidence E compared with her/his initial degree of belief of hypothesis H. To calculate $P(H|E)$, it is based on the

conditional probability:

$$P(H|E) = P(E|H) \times P(H)/P(E).$$

It may not be difficult to estimate $P(E|H)$ if H is the null hypothesis. Unfortunately, H is typically not the null hypothesis in this case, so there may be difficulties to estimate $P(E|H)$. In addition, there are real problems to estimate the prior probabilities, $P(H)$ and $P(E)$ as indicated by Earman [1992] who offered three ways to deal with the problem. The first proposal is the hope that the priors 'wash out' as evidence accumulates. This is not useful for the scientists because they need to justify the priors in order to come up with a conclusion in their papers for publication. If more and more evidence accumulated adjusts the priors, then the conclusion drawn may depend on the stage of the investigation. Later work may find earlier work drawing different conclusions because the prior probabilities have changed! This is not very desirable for scientists. The second proposal is to provide rules to fix the initial degrees of belief. Earman [1992] commented that "none of the rules cooked up so far is capable of coping with the wealth of information that typically bears on the assignment of priors". So, scientists cannot rely on this proposal as there may be debate over which rules to use as well as whether the rules are suitable. The third proposal is based on plausibility argument. Again, the scientists cannot rely on such argument because this would open up for debate when they draw conclusions in their paper. Therefore, there is no remedy for the prior probability problem for scientists.

Another strategy for the scientists is to try to cancel out or embed the prior probabilities so that we do not need to estimate them. In this case, not all Bayesian confirmation measures [Fitelson, 1999] can be used to cancel out the prior probabilities. For example, Carnap's measure [Carnap, 1962] cannot cancel out the prior probabilities. On the other hand, measures forming a ratio of probability may be able to cancel out the prior probabilities. For example, Keynes [1921] is interested in the ratio $P(H|E)/P(H)$ so that one can consider the odds in favor of the hypothesis H given the evidence E is known and one does not need to estimate $P(H)$. However, one still has the problem of estimating $P(E)$ which is not trivial. One may argue that instead of comparing $P(H|E)$ and $P(H)$, we compare the conditional probabilities between two hypotheses, $H0$ and $H1$, so that $P(E)$ is canceled as follows:

$$P(H1|E)/P(H0|E) = P(E|H1)/P(E|H0) \times P(H1)/P(H0).$$

The above ratio on the immediate right of the equal sign is called the Bayse factor and the ratio on the far right is the prior ratio. If the above ratio on the left is larger than one then we have a higher degree of belief for $H1$ over $H0$. However, how can the scientists estimate $P(H0)$ and $P(H1)$? One solution is to invoke the principle of indifference so that $P(H0) = P(H1) = 0.5$ as there are two hypotheses.

However, these are the two hypotheses tested and not the total number of hypotheses that are in existence. This is the point where there is debate about how to set the prior probabilities as it is uncertain how many alternative hypotheses there are to take into account. Unfortunately, this may affect the test and the conclusion drawn by scientists. We believe that this is why Bayesian confirmation theory is not widely used because its application to drawn (scientific) conclusion is open to debate.

On the other hand, if we invoke the principle of indifference then $P(H0) = P(H1) = 1/n$ for n hypotheses. Now, since $P(H0) = P(H1)$ for whatever number of hypotheses, we have

$$P(H1|E)/P(H0|E) = P(E|H1)/P(E|H0).$$

Therefore, this ratio can be used as the basis to accept or reject hypothesis $H1$ compared with $H0$ without any prior probabilities. However, there are (at least) three problems. One problem is that it may not be easy to estimate $P(E|H1)$ because $H1$ is not the null hypothesis any more. The second problem occurs when $P(H1|E)/P(H0|E) > 1$, but $P(H1|E)/P(H2|E) < 1$. In the absence of knowing all the hypotheses, we do not know whether $H1$ is the most likely hypothesis to be accepted. Which hypothesis has the highest probability ratio is important to scientists since that hypothesis is supposed to be the leading one to be confirmed. We may assume that the state-of-the-art theory or model is published in journals or conference proceedings, and it is the leading one to form $H0$ in order to compare with the proposed theory or model forming $H1$. So, the second problem may be resolved partially. The third problem is that we still do not have any measure of the risk involved in accepting $H1$, so it is difficult to appreciate the likelihood of making an error. Although for some distributions of the underlying probabilities of the ratio, we can deduce the distribution of the ratio of probabilities. In general, we cannot deduce the distribution of the ratio for any distributions of the underlying probabilities of the ratio. Likewise, although for nested models we can assume the log-likelihood ratio (i.e., $\log[P(H1|E)/P(H0|E)]$) multiplied by two to asymptotically follow a chi-square distribution so that we can translate between the log-likelihood ratio value and the estimated p-value, in general again it is an open research problem especially for non-nested models and for small samples. Therefore, given the first and last problems Bayesian confirmation theory is not very popular among scientists.

3 Scientists' solution

To avoid the debate about the prior probabilities and to estimate the risk of the decision making, scientists tend to use a different statistical method. The idea is

to compare the performance based on a control group using classical hypothesis testing, i.e. null hypothesis significance testing (NHST) and Neyman-Pearson hypothesis testing. The null hypothesis is that the performance of the particular theory or model has no difference with that of the control group. If the performance is different from the control group statistically significantly, then the scientists can report the p-value or significance level, and claim that the null hypothesis is rejected given a particular confidence/significance level.

For NHST, the scientists can know the risk, i.e., the type-I error of incorrectly rejecting the null hypothesis. In this way, there is no need to estimate $P(E)$ or $P(H)$ while at the same time, the risk in making the wrong decision is known. Many scientists are only concerned with the type-I error because they are interested in rejecting the null hypothesis. Otherwise, if they cannot reject the null hypothesis, then they usually may not be able to publish their scientific paper as they do not have a better model or theory. Also, it is relatively difficult to estimate the type-II error for composite alternative hypothesis because there may be more than one parameter value for the distribution of the composite alternative hypothesis. In this case, it is not clear how the Bayesian confirmation theory handles such composite alternative hypothesis as there may be many prior distributions that fit the composite alternative hypothesis.

The control group that has the lowest performance is based on a random model or random guessing. This provides the lower bound performance of a model that scientific models must perform better than according to the basic principle of modeling accuracy by Luk [2017]. To establish a new scientific model, this model is compared with the old scientific model that serves as the control group. Since the old scientific model is supposed to be better than the random model or guessing, the new scientific model is expected to perform better than the random model or guessing when the new model performs better than the old scientific model statistically significantly. Therefore, when there is an established scientific model, there is no need to compare the new one with the random one. It is sufficient to compare the new model with the established one.

Scientists make use of many statistical tests that would give some idea about the risk in the decision making process so that others know about the confidence level in arriving at the acceptance or rejection of the null hypothesis. For example, scientists may use paired tests to eliminate influence of other intervening factors in the comparison. Typical paired tests include the Wilcoxon paired signed rank test and the randomization test (e.g., [Smucker et al., 2007]). For testing whether laws or principles hold in the theory, regressions may be used. Without any control group to compare, the statistical test decides whether the null hypothesis that the coefficients of the regression are zero is true or not. With some control group, some scientists may use Chi-square to compare two distinct regression models, and some scientists use the F-test to compare two nested models (i.e., one being

the special case of the other). For a probability distribution like that specified by the Zipf law, the Chi-square test can be used. All these examples show that the statistical tests only compare with the null hypothesis and the reliability or risk is about making the wrong decision to reject the null hypothesis as most of the scientific papers report performance better than that mentioned in the null hypothesis. Therefore, many scientists do not use Bayesian confirmation theory to find support for their conclusion and it does not seem to be popular among many scientists for their work.

Statistical tests are done one at a time to compare with some state-of-the-art model or theory serving as the control group. The performance of this control group is used to set the null hypothesis that there is no performance difference between the control group and the new model or the new theory under test. By showing that the null hypothesis is rejected, scientists then claim that they have a better model or theory with statistically significant results, and this is the evidence for showing that scientific progress is made. To increase the reliability, more than one experiment reported in more than one scientific paper are used to obtain statistical significance results to support that scientific progress is made. For establishing a superior theory which is applied to build various models, a random model (serving as the control group in the null hypothesis) can be used to decide whether the new theory is better than the older theory by observing the number of models of the new theory that are better than the corresponding models of the old theory (similar to showing that the theory is true in [Luk, 2018]). For the random model in the null hypothesis, we may assume that the probability that the model of the new theory will perform the same as the corresponding model of the old theory is a half. After comparing with N different models, we can estimate the p-value based on the binomial distribution, so we can decide to reject the null hypothesis or not. Hence, scientific progress can be made to advance one theory over the other by using statistical tests in this way. However, for a widely held theory, usually, the newer theory is required to perform better than the old theory for every model they generate because there are not many models for the (expensive) experiments in the course of scientific development. For example, the experiment by LIGO team demonstrating gravitational waves (e.g., see [Bunge, 2018] for a discussion) is one more experiment to the existing few (e.g., Eddington and Gravitation Probe B experiments) that support Einstein's general relativity theory over Newtonian universal gravitation theory. Note that each experiment gives rise to a new model derived from the theory so that with several experiments, there are several models that make predictions according to the same theory.

4 Practical problems with the scientists' solution

Some philosophers are knowledgeable of how scientists perform statistical tests to claim the superiority of their model or theory. For example, Bird [2018] commented that scientists perform NHST and the randomized controlled trial (RCT), so he is aware that at least some scientists (at least those he examined) do not actually use Bayesian confirmation theory in practice. He did not provide any explanation as to why Bayesian confirmation theory is not used by the scientists. Instead, he focused on explaining the replication crisis [Baker, 2016] due to the low success rate of the concerned hypothesis.

The explanation by Bird [2018] suggests that in some scientific disciplines scientists may propose many implausible hypotheses that have low success rate, in poorly understood topics where the knowledge is highly incomplete. As a result, the actual false positive rate may be alarmingly high compared with that specified by the confidence level. Consequently, many experiments in such scientific disciplines may not be able to replicate or reproduce their results. He proposed three responds to this situation: (1) do nothing and keep quiet, (2) seek high-quality hypothesis (with high success likelihood) and (3) increase the confidence level.

Apart from these responses, scientists have other options that Bird did not mention. In some scientific disciplines, instead of just one data collection, the study may perform the experiment on several (highly different) data collections. Statistical tests are performed for different data collections. If all the data collections show statistically significant results, then it suggests that the proposed models or theories are more reliably better. This requires more resources but for some disciplines, this is the norm rather than the exception. Another option is to perform some kind of replication study but with some novel twist to the theory or model to throw some light on its generality. For example, Rainville et al. [2005] do not reproduce any experiment. Instead, they invent a new experiment to validate $E = mc^2$ thereby supporting or falsifying special relativity. In addition, their experiment tries to measure the precision that the famous equation holds. Yet another option is to perform some comparison study. In this case, many hypotheses may have been proposed to explain a phenomenon and the comparison study tries to isolate which hypotheses are critical to the observation of the phenomenon. Without further resources, another option is to partition the data into subgroups and perform statistical tests of the subgroups to see if reliable significance results can be obtained for each subgroup. Finally, instead of explanatory modeling, we can perform predictive modeling as suggested by Yarkoni and Westfall [2017] for psychology studies. In this case, we can perform N-fold cross-validation of the predictive model to ascertain the validity of the superiority of the proposed model or theory. Given that there are many remedies to classical hypothesis testing, the solution using classical hypothesis testing still has some advantage over Bayesian

confirmation theory because such theory does not provide the risk or its estimate in making the wrong decision.

Another problem with classical hypothesis testing is the issue about optional stopping rule (e.g., [Mayo, 1996] and [Howson and Urbach, 2006]). In this issue, it was found that for two different sampling plans, classical hypothesis testing can result in different conclusions for some specific set of data. Theoretically, this is undesirable. In practice, this can be circumvented so that it is not an insurmountable practical problem. The idea is to use a statistical plan that is less likely to be challenged by the reviewers. Therefore, the scientists play the "diligent researcher" role when selecting the sampling plan to show that they have used a commonly accepted sampling plan to sample that does not have many controversies. Only in special circumstances, when it is not feasible to play the diligent researcher role, a more controversial sampling plan may be selected and the researcher has to provide special justifications for such sampling plan in the research article to entice the reviewers to accept the paper.

For Neyman-Pearson hypothesis testing, it is not always possible to analytically derive the distribution of the likelihood ratio even though the likelihood ratio can be defined for composite alternative hypothesis [Casella and Berger, 2002]. This means that it is not simple to relate the significance level with the likelihood ratio value for some distributions. As a result, it is not always possible to know the significance level analytically given the likelihood ratio value although there may be some practical estimation method for discrete distributions. Therefore, one cannot claim that this test can always assess the risk of the wrong decision making as the likelihood ratio may not be able to translate to the significance level. In the case that the risk cannot be assessed by Neyman-Pearson hypothesis testing, scientists can always revert back to NHST to assess the risk of making the wrong decision so that this is not a great handicap for classical hypothesis testing. Alternatively, for nested models, scientists can gather a large amount of data if that is possible. This is because due to a theorem by Wilks [1938], as the sample size approaches infinity, the log-likelihood ratio (i.e., $\log[P(H1|E)/P(H0|E)]$) multiplied by two for a nested model asymptotically follows a chi-square distribution so that an approximate statistical test can be made in practice with knowledge of the approximate risk involved in the decision making. However, for the general case, this is still an open research problem.

5 Alternative theories and models

Ideally, when a paper about a theory or a model is proposed, the paper should report a better theory or better model than existing ones by providing evidence that better predictions are made. Accompanied with better results, there should

also be some assessment of the reliability of the results and so some statistical testing should be done. Typically, the null hypothesis that there is no difference in performance is rejected. For clear cut cases, papers proposing better theories or better models should be published. However, in real life, the proposed theory or model may not always be better than existing ones. Worst still, some proposed theory or model performs with no statistical significant difference from the existing ones. Should such theory or model be published by the journal or conference proceedings?

We answer this question by recalling an example of the problem of induction proposed by Bertrand Russell [1912]. In this example problem, a chicken (or a turkey) observes that the farmer keeps feeding him every day. So, by induction, the chicken concludes that the farmer will keep feeding him in the future. Until one day, the farmer slaughters the chicken for meat. This has been a problem for the believers of induction, as induction cannot guarantee that the future will occur identically as the past. However, there are few guarantees in life. Therefore, should we just accept induction as a limitation of our ability to know? With such a drastic life or death consequence, perhaps the chicken should think twice before accepting induction. What can the chicken do?

What the chicken should do is not to be satisfied with the only conclusion that the farmer keeps feeding him in the future. The chicken should hypothesize alternative theories or models explaining why the farmer feeds chicken in general and then observe whether these alternative theories or models can provide an alternative understanding as to why the farmer keeps feeding the chicken. Based on the existing evidence, the chicken may not be able to find a better theory or model but it is important for the chicken to keep in mind alternative theories and models in order to assess what are the possible consequences. With these alternative understandings, the chicken can look for evidence to support the surviving theory or model or weed out the other theories or models.

In science, we are faced with a similar situation as the chicken. The existing theory or model may perform well but if we rely on them only, we may only find what these theory or model predict (as in the confirmation bias). Instead, we should actively seek alternative theories or models to provide an alternative understanding of the topic so that we may assess the different impact. Initially, these alternative theories or models may not be able to perform better than existing ones, but they provide an alternative understanding of the topic. Therefore, they should be published so that other scientists can find evidence to determine which theory or model should survive. If such alternative theories or models are weeded out of the publication process during the review, then other scientists cannot help to find the surviving theory. In science, proposing a new theory or model, or finding a surviving theory or model may take a life time. Therefore, it is important for papers about alternative theories and models to be published and archived so

that in the future they can be tested. Therefore, even theories and models that only perform without significant difference from the existing theory and model are worthy to be published.

If alternative theories or models are allowed to be published, will we face a deluge of them with many junk theories and models archived? Will this pollute the field making it hard for the research to find the signal from the noise? Our proposal is that not all alternative theories or models should be published. We should publish those that perform at least with no statistically significant difference from the existing theory or model. If the existing theory or model is highly effective in terms of their predictions, then this will avoid lots of theories or models getting published. Apart from this criterion, we should also demand that the author should provide an alternative understanding of the topic and give some prediction that would distinguish the proposed theory or model from the existing ones. This would give a lead to other scientists to find evidence to weed out the theories or models, thereby accelerating the process of falsifying theories or models.

Note that when the null hypothesis is about comparing with the random model or random guessing, we require the proposed model or theory to be better than the random model or random guessing because random model or guessing represents that there is no knowledge about the specific topic or issue. Therefore, if we have some (scientific) knowledge, then we should get better results than no knowledge. However, when the null hypothesis is comparing with some state-of-the-art theory or model that is known to be performing better than the random model or random guessing, we only require the proposed theory or model to be the same or better than the state-of-the art theory or model for publication. In this case, similar to the example problem, we keep a look out for an alternative theory or model that can eventually perform better than the state-of-the-art theory or model. Thus, we should allow such alternative theory or model to be published.

While the classical hypothesis testing can test theories or models that are performing similarly with no statistically significance difference, Bayesian confirmation measures may have problems testing alternative theories or models. For example, the ratio of the probabilities (i.e., $P(H1|E)/P(H0|E)$) needs to be exactly one for the alternative theory or model to be performing similarly as the existing theory or model. In practice, getting one exactly is very difficult. If we relax the requirement of getting one exactly, we need to know the distribution of the ratio of probabilities. Sometimes, it is possible to deduce the distribution of the ratio of probabilities if we know the distribution of the underlying probabilities of the ratio. However, sometimes we do not know. Likewise, for nested models and large samples, the log-likelihood ratio (multiplied by two) asymptotically follows the chi-square distribution due to Wilks' theorem so that the ratio value can be translated into a p-value, but it is still an open research problem to find the distribution for non-nested models or for small samples. Therefore, this creates a

practical problem for scientists who may not rely on using Bayesian confirmation theory.

6 Modeling scientists' decisions

Bayesian confirmation theory can be modeling the degree of belief of the scientists when testing a hypothesis rather than a practical procedure to perform decisions to accept or reject the hypothesis of an experiment. Here, we provide a sketch of how this can be done. As there is uncertainty about some of the prior probabilities, it is difficult to know the estimated conditional probability $P(H|E)$. To verify whether the ratio, $P(H|E)/P(H)$, is accurate, we can ask scientists in a survey on how much more confident they are that $P(H|E)$ is compared with $P(H)$ before and after they have read about the scientific work. Then, we have the open research issue about how to translate the confidence to the ratio, $P(H|E)/P(H)$, which is open to yet another debate (e.g., [Kaplan, 1989] and [Huber, 2005]). Even if we have an estimate of $P(H|E)/P(H)$, we have the additional difficulty to determine the p-value from the null hypothesis because we have only a single probability for one experiment. The Bayesian confirmation theory requires M experiments to decide. In this case, the null hypothesis is that $P(H|E)/P(H) =$ confidence score in the survey. Note that $P(H)$ and $P(E)$ vary for different experiments because the number of hypotheses may vary with different experiments and the nature of evidence for each experiment may be different. As a result, we need to compare the predicted ratio $P(H|E)/P(H)$ and the confidence score by a paired test so that the null hypothesis is that the $P(H|E)/P(H)$ minus the confidence score is zero. A Wilcoxon paired signed rank test can be used to obtain the p-value for instance so that the scientist can obtain the risk in making the decision to reject or accept the null hypothesis. However, to make such a decision, more than one experiment is needed, and it is not clear whether the ordinary scientists are willing to put in the extra effort in addition to controversies in setting $P(H)$ and $P(E)$ for each experiment in order to predict the ratio, $P(H|E)/P(H)$. Although we have a mechanism to confirm whether the Bayesian confirmation theory predicts the subjective degree of belief of the scientists by running this kind of statistical tests, it is unclear whether ordinary scientists will perform such task. Most likely, this is the task for the experimental philosophers or researchers on science of science [Fortunato et al., 2018] to verify whether Bayesian confirmation theory makes good prediction about the relative degree of belief of the scientists for favoring hypothesis H but this confirmation is open to debate as how to set $P(H)$ and $P(E)$ for each experiment is an open issue.

Instead of subjective probability, $P(H|E)$ and $P(H)$ can be interpreted as

objective probability. One can count the number of scientists who belief in the hypothesis H after weighing on the evidence E and before the evidence E is made available. Next, $P(E)$ can be interpreted as the proportion of scientists who can access the evidence E. In this way, we can model the scientist decision making process. Therefore, Bayesian confirmation theory can model a scientist decision making process in hypothesis testing rather than using it to do hypothesis testing.

7 Jeffreys-Lindley paradox

This paradox [Lindley, 1957] has been debated in philosophy of science ([Spanos, 2013], [Sprenger, 2013] and [Robert, 2014]) and statistics for some time. It occurs when the sample size is large. For a point value null hypothesis, as the sample size tends to infinity, the point value may approach a particular value. For very large samples, the deviations from the point value may be small for the significance level in NHST, so that one may reject the null hypothesis because of tiny deviations due to the large sample. However, for some prior probabilities, the Bayesian confirmation theory suggests that the posterior probability of the null hypothesis approaches one instead of rejecting the null hypothesis. This incompatibility between NHST and Bayesian confirmation theory is the Jeffreys-Lindley paradox.

We resolve this paradox by recalling what these probabilities are supposed to model. For NHST, the probability is modeling the chance that the data occurred with the particular value deviating from the value specified by the null hypothesis. For Bayesian confirmation theory, we are modeling the belief of the scientists in accepting/rejecting the hypothesis. These two different kinds of modeling do not necessarily imply that their probabilities have to be consistent with each other. So, there is no paradox. For NHST, the probability is about the chance of data having some particular value whereas for Bayesian confirmation theory, the probability is about the belief of the scientist. In addition, the Bayesian confirmation theory is only about modeling the degree of belief of the scientist instead of getting the exact correct degree of belief. Therefore, even if Bayesian confirmation theory says the probability is 1.0, it is only an estimate. It can actually be 0.6 instead of one. Hence, whether Bayesian confirmation theory produces a probability that is consistent with NHST is not a real practical issue. In practice, scientists follow what the data tells them assuming that the assumed distribution is appropriate as this is a decision based on evidence; they are not concerned about the modeling of their belief about the hypothesis as that would mean that they are making decisions based on the degree of belief rather than based on evidence. Therefore, this paradox does not have an impact in practice for NHST but it has a negative impact on Bayesian confirmation theory.

8 Software packages?

Another plausible reason why Bayesian hypothesis testing is less utilized than classical hypothesis testing is that software packages are available for classical hypothesis testing but no software package was available for Bayesian hypothesis testing in the past. Recently, Wagenmakers et al. [2018a] favor the use of Bayesian hypothesis testing over classical hypothesis testing and gave examples [Wagenmakers et al., 2018b] of the use of Bayesian hypothesis testing using an open source package called JASP [Wagenmakers, 2017] trying to entice psychologists to use Bayesian hypothesis testing. Less radical is Quintana and Williams [2018] who advocated the use of Bayesian hypothesis testing in conjunction with classical hypothesis testing in order to be more informative about the hypothesis testing as these two methods are thought to complement each other. Apart from JASP, Quintana and Williams noted that Bayesian hypothesis testing is also available in the 'BayesFactor' R package [Morey and Rouder, 2018].

Despite such software packages are now available, there is still the thorny issue of choosing the prior distribution or prior probabilities for Bayesian hypothesis testing. At present, Quintana and Williams [2018] suggested to perform some robust or sensitivity analysis to check whether changing the prior distribution will have a great impact on the results and the conclusion drawn. If not, the conclusion drawn would be robust to different prior distributions. Otherwise, care should be taken to interpret the results. This is, however, what a responsible scientist should do. Alternatively, an irresponsible scientist may perform prior distribution hacking similar to p-value hacking, trying to obtain the most favorable results based on finding the suitable prior distributions to reject the null hypothesis. Therefore, Bayesian hypothesis testing is not immune to abuse.

Apart from the issue about choosing the prior distribution, sometimes the prior distribution specification is vague so that it is not possible to perform the hypothesis testing. However, NHST may still be able to perform the significance test. Similarly, sometimes there may not be an alternative hypothesis apart from the negation of the null hypothesis. For example, when the alternative hypothesis is $\theta \neq \theta_o$, then it is not clear how to specify the alternative hypothesis for Bayesian hypothesis testing since the alternative hypothesis does not specify θ to take on a specific value but rather to indicate that it should not take on a specific value. In Bayesian hypothesis testing, the practical solution for this kind of problem is to specify a number of prior distributions that can satisfy this general condition (e.g., $\theta \neq \theta_o$) of the alternative hypothesis, and determine whether the results are robust to these different prior distributions. As can be seen, it is not clear whether such Bayesian hypothesis testing has real advantage over NHST when presenting results and inferences in a paper as such results or inferences may be inconclusive.

Apart from problems with the prior distributions, using the software packages

for Bayesian hypothesis testing does not directly indicate the risk that we make the wrong decisions whereas classical hypothesis testing gives us some idea based on the p-value (or significance level). The Bayes factor reported by such packages gives us the odds comparing the alternative hypothesis with the null hypothesis but this is not the same as how likely we made the wrong decision in accepting/rejecting the null hypothesis. To do this, we need to know the underlying distributions of $P(E|H)$ which cannot guarantee to derive the distributions of the Bayes factor. In addition, it is unclear what the final distribution is since the Bayes factor is multiplied by the prior ratio (which has its own distribution). Likewise, although for nested models, the log-likelihood (i.e., $\log[P(H1|E)/P(H0|E)]$) multiplied by two asymptotically follows the chi-square distribution so that we can translate the approximate likelihood ratio value (estimated by multiplying the Bayes factor with the prior ratio) to a p-value, it is still an open research problem to find the distribution in the general case especially for non-nested models and for small samples. Therefore, there is no general solution to provide the risk in accepting/rejecting the null hypothesis for Bayesian hypothesis testing.

Given these practical shortcomings, it is not clear whether Bayesian hypothesis testing can become as widely utilized as classical hypothesis testing since Bayesian has many theoretical advantages. Time will tell whether Bayesian hypothesis testing will be as popular as classical hypothesis testing, or even more popular than classical hypothesis testing or still remains obscure in practice compared with classical hypothesis testing.

9 Conclusion

This article does not compare classical hypothesis testing with Bayesian confirmation theory as there are better suited articles for this (e.g., [Romejin, 2017]). It is also not an advocate of classical hypothesis testing as it has many unresolved statistical and philosophical issues (e.g., violation of the likelihood principle). Instead, this article is about the practical problems that explain why Bayesian confirmation theory is not practiced by scientists. Specifically, (a) Bayesian confirmation theory does not measure the reliability or risk of the decision making, (b) the assignment of probabilities to some of its (prior) probabilities may be open to debate affecting the conclusion drawn by the scientists, and (c) software package for Bayesian hypothesis was not available before whereas software package for classical hypothesis testing is widely available. Therefore, scientists use other methods to support their conclusion in scientific discourse. Specifically, scientists use classical hypothesis testing to obtain the significance level or p-value measuring the risk of rejecting the null hypothesis (based on some control group). Evidence of scientific progress is made when scientists found their theory or model obtaining

statistical significant results when they compare the performance of their model or theory with the state-of-the-art model or theory (serving as the control group in the null hypothesis) or in the absence of any model or theory a random model is used. To avoid the problem of induction, it is suggested attention should be paid not to just better theories and models but similarly performing theories and models so that the alternative theory may guide us to look for what kind of evidence to weed out the unsuccessful theory or model and the better surviving theory or model may be found in the future. While at present Bayesian confirmation theory may not be widely used by scientists to make decisions about which model or theory is better, time will tell whether it will be widely utilized due to the availability of software packages that accelerate utilization of Bayesian hypothesis testing. In addition, Bayesian confirmation theory has the advantage that it is not biased against the null hypothesis, so that scientists showing significance results based on Bayesian hypothesis testing appears to be more robust than classical hypothesis testing. However, the prior distributions of Bayesian hypothesis testing need to be examined carefully, and the classical hypothesis testing can adjust its significance level for better robustness, so that the advantage of Bayesian hypothesis testing is not that apparent. Finally, the Bayesian confirmation theory offers a probabilistic model of scientists making decisions in accepting or rejecting hypotheses.

Acknowledgements

I would like to thank Dr. Edward Dang for proof-reading the paper.

References

M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

A.J. Bird. Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, Forthcoming(doi:10.1093/bjps/axy051), 2018.

P. Brössel and F. Huber. Bayesian confirmation: a means with no end. *The British Journal for the Philosophy of Science*, 66(4):737–749, 2015.

M. Bunge. Gravitational waves and spacetime. *Foundations of Science*, 23(2): 399–403, 2018.

R. Carnap. *Logical foundations of probability, 2nd Edition*. Chicago Univeresity Press, Chicago, 1962.

G. Casella and R. L. Berger. *Statistical inference*. Thomson Learning, University of Michigan, 2002.

D. Castelvecchi. Feuding physicists turn to philosophy for help. *Nature*, 528 (7583):446–447, 2015.

C.S. Chihara. Some problems for Bayesian confirmation theory. *The British Journal for the Philosophy of Science*, 38(4):551–560, 1987.

V. Crupi, R. Festa, and T. Mastropasqua. Bayesian confirmation by uncertain evidence: a reply to Huber [2005]. *The British Journal for the Philosophy of Science*, 59(2):201–211, 2008.

J. Earman. *Bayse or bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, Cambridge, MA, 1992.

R. Festa and G. Cevolani. Unfolding the grammar of Bayesian confirmation: likelihood and antilikelihood principles. *Philosophy of Science*, 84(1):56–81, 2017.

B. Fitelson. The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, 66:S362–S378, 1999.

S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Peterson, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A-L. Barabási. Science of science. *Science*, 359(6379): doi:10.1126/science.aao0185, 2018.

L. Henderson, N. D. Goodman, J. B. Tenenbaum, and J. F. Woodward. The structure and dynamics of scientific theories: a hierarchical Bayesian perspective. *Philosophy of Science*, 77(2):172–200, 2010.

C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle, 2006.

F. Huber. Subjective probabilities as basis for scientific reasoning? *The British Journal for the Philosophy of Science*, 56(1):101–116, 2005.

M. Kaplan. Bayesianism without the black box. *Philosophy of Science*, 56(1): 48–69, 1989.

J. Keynes. *A Treatise on Probability*. Macmillan, London, 1921.

D. V. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

R. W. P. Luk. A theory of scientific study. *Foundations of Science*, 22(1):11–38, 2017.

R. W. P. Luk. On the implications and extensions of Luk's theory and model of scientific study. *Foundations of Science*, 23(1):103–118, 2018.

D. G. Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago, Chicago, 1996.

R. D. Morey and J. N. Rouder. *BayseFactor: computation of bayes factors for common designs*. https://CRAN.R-project.org/ package=BayesFactor, 2018.

W. C. Myrvold. A Bayesian account of the virtue of unification. *Philosophy of Science*, 70(2):399–423, 2003.

J. D. Norton. Challenges to Bayesian confirmation theory. *In Philosophy of Statistics: Volume 7 in Handbook of the Philosophy of Science*, 7:391–439, 2011.

D. S. Quintana and D. R. Williams. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using JASP. *BMC Psychiatry*, 18:178–185, 2018.

S. Rainville, J. K. Thompson, E. G. Myers, J. M. Brown, M. S. Dewey, E.G. Kessler Jr, R. D. Deslattes, H. G. Börner, M. Jentschel, P. Mutti, and D. E. Pritchard. A direct test of $E = mc^2$. *Nature*, 438(22):1096–1097, 2005.

C. Robert. On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2):216–232, 2014.

J-W. Romejin. *Philosophy of Statistics*. E.N. Zalta (ed.), https://plato.stanford.edu/archives/spr2017/entries/statistics/, 2017.

B. Russell. *The Problems of Philosophy*. Williams and Norgate, London, 1912.

M. Schippers and G. Schurz. Genuine confirmation and tacking by conjunction. *The British Journal for the Philosophy of Science*, Forthcoming (doi:10.1093/bjps/axy005), 2018.

M. J. Shaffer. Bayesian confirmation theories that incorporate idealizations. *Philosophy of Science*, 68(1):36–52, 2001.

M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. *In Proceedings of the 16th ACM CIKM '07*, pages 623–632, 2007.

A. Spanos. Who should be afraid of the Jeffreys-Lindey paradox. *Philosophy of Science*, 80(1):73–93, 2013.

J. Sprenger. Testing a precise null hypothesis: the case of Lindley's paradox. *Philosophy of Science*, 80(5):733–744, 2013.

E. J. Wagenmakers. *JASP*. https://jasp-stats.org, 2017.

E. J. Wagenmakers, M. Marsman, T. Jamil, A. Ly, J. Verhagen, J. Love, R. Selker, Q.F. Gronau, M. Šmíra, S.Epskamp, D. Matzke, J. N. Rouder, and R. D. Morey. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psyconomic Bulletin & Review*, 25(1):35–57, 2018a.

E. J. Wagenmakers, M. Marsman J. Love, A. Ly T. Jamil, J. Verhagen, R. Selker, Q. F. Gronau, D. Dropman, B. Boutin, F. Meerhoff, P. Knight, A. Raj, D-J. van Kesteren, and J. van Doorn. Bayesian inference for psychology. Part II: Example application with JASP. *Psyconomic Bulletin & Review*, 25(1):58–76, 2018b.

A. Wayne. Bayesianism and diverse evidence. *Philosophy of Science*, 62(1): 111–121, 1995.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9(1):60–62, 1938.

T. Yarkoni and J. Westfall. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspective on Psychological Science*, 12(6): 1100–1122, 2017.