Revisiting the two predominant statistical problems: the stopping-rule problem and the catch-all hypothesis problem

Yusaku Ohkubo*,**,***

Abstract

The history of statistics is filled with many controversies, in which the prime focus has been the difference in the "interpretation of probability" between Frequentist and Bayesian theories. Many philosophical arguments have been elaborated to examine the problems of both theories based on this dichotomized view of statistics, including the well-known stopping-rule problem and the catch-all hypothesis problem. However, there are also several "hybrid" approaches in theory, practice, and philosophical analysis. This poses many fundamental questions. This paper reviews three cases and argues that the interpretation problem of probability is insufficient to begin a philosophical analysis of the current issues in the field of statistics. A novel viewpoint is proposed to examine the relationship between the stopping-rule problem and the catch-all hypothesis problem.

Key words: epistemology, statistics, Frequentist, Bayesian replication crisis

1. Introduction

Statistics has been an indispensable tool for scientific inquiry. It is also a field characterized by intense controversy among researchers and philosophers. In the history of statistics and the philosophy of statistics, the focus of the most intense conflict has been two schools of statistical thought[1]: Frequentist (or error statistician) and Bayesian.

The Frequentist-Bayesian conflict has taken place at three different levels:

* Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University
Email: ohkubo.yusaku1989@gmail.com
** Center for Data Assimilation Research and Applications, Joint Support Center for Data Science Research, Research Organization of Information and Systems
*** the Institute of Statistical Mathematics

[1] In the philosophy of statistics, likelihoodism also has a huge influence. However, I will leave this issue aside in this paper because of its, at least relatively, limited applications in science thus far.

methodology, epistemological justification, and interpretations of results. For a historical example, Neyman and Pearson refused to set prior probability (methodological level), assumed an expected property under a "true" distribution to show validities of a statistical method (epistemological justification level) and took results as optimal decision rules (interpretation level). In contrast, Savage applied Bayes' theorem with prior probability, based on an axiom of the rational agent, and took results as a "degree of belief" of this rational person (see later sections for details). These positions have been the influential bases of many philosophical works and textbooks.

Currently, however, as discussed later, many "hybrid" approaches have been proposed at different levels; some "Bayesian" statisticians reject subjective interpretation and adopt Frequentist properties to justify their procedures. Additionally, some "Frequentists" reject that the results of their method should be interpreted in the sense of long-run properties. Dichotomizing statistical theories based on Frequentist/Bayesian methods might be inappropriate for analyzing the current practice of statistics because these "hybrid" approaches complicate the relationships between the three different levels. In this paper, I examine the relationship between two schools of statistics—Frequentist and Bayesian—and propose a novel point of view to clarify the connection between the two schools of statistical thought. I then argue that the two well-known problems among philosophers of statistics—the stopping-rule problem and the catch-all hypothesis problem—are merely different sides of the same coin when approached from the proposed perspective.

The rest of the article is organized as follows. In sections 2 and 3, the stopping-rule problem of Frequentist theory and the catch-all hypothesis problem of Bayesian theory are reviewed, respectively. Then, it is argued that both problems pose serious difficulties in the practice of science and that they have still not been solved. In section 4, a revision is discussed of the dichotomized account of Frequentist/Bayesian methods that is needed to overcome them and consider the status of a recent "hybrid" statistical approach as a natural candidate to do so. In section 5, it is argued that this alternative is not the solution and in showing why not, a novel perspective of statistics is proposed through which the two problems are mirror image of the same problem. Finally, in section 6, the article is concluded, and the implications for practice are discussed. Even though a complete solution to the stopping-rule problem and the catch-all hypothesis problem is not presented, a possible direction for future work is examined.

2. The Frequentist and the stopping-rule problem

Frequentist statistics have been the predominant approach to data analysis among scientists. Its scope includes point estimation, interval estimation, hypothesis testing, model selection, model averaging, and so on. These procedures are, as its name suggests, justified on the frequency properties of a method; that is, a performance of the method when it is repeatedly applied to a data. Take the maximum likelihood estimator (MLE), for example.

Suppose we have a collected data x. Frequentist methods are based on the assumption that there exists an unobservable probability distribution of X, called the "true distribution", and that x is a random sample from it[2]. Then, MLE is defined as

$$\boldsymbol{\theta}_{MLE} = \arg\max L(x; \boldsymbol{\theta}),$$

where $L(\ )$ is the likelihood function of the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ given the model. In other words, MLE is the value at which the likelihood of the given model is maximized. Under some conditions (Wald 1949; Huber 1967), MLE is known to satisfy the following asymptotic property:

$$\sqrt{n}(\boldsymbol{\theta}_{True} - \boldsymbol{\theta}_{MLE}) \sim Normal(\mathbf{0}, \boldsymbol{I}^{-1}),$$

where $\boldsymbol{\theta}_{True}$ is the unknown true value, which we want to estimate by a finite sample at hand, $Normal(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})$ is the normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{I}$ is a metric called Fisher Information. Thus,

$$\mathbb{E}_{x \sim X}[\boldsymbol{\theta}_{True} - \boldsymbol{\theta}_{MLE}] = \mathbf{0},$$

where $\mathbb{E}_{x \sim X}[\ ]$ is the expectation over x. Note that each x is a finite random sample from a (typically unknown) probability distribution. The above property shows that even though the MLE could produce a "wrong" estimation of the true value because the MLE is a random variable, it is still a "good" guess of the true value because the relative frequency of the MLEs, which correctly specifies the true value, is highest among other MLEs when we repeat the process of data sampling and estimation. It is why the MLE has a Frequentist justification[3]. Even though there exist many other Frequentist methods and each of them requires elaborating mathematical analyses, their main argument is similar to this strategy (take Akaike Information Criterion for example; Sec. 4.2 below).

This strategy is the ground of a typical claim that the Frequentist inference merely offers a "behavioristic rule" or is based on a "long-run performance". Now, we see in what sense Frequentist inferences have been seen as "behavioristic" or a "long-run" performance, although this interpretation is reconsidered later. Suppose a 95% confidence interval (CI), which is a Frequentist method of interval estimation of a parameter $\boldsymbol{\theta}$. Typically, it takes the form

[2] This metaphysical commitment would be another point where Frequentist and Bayesian theories does not agree, although this is beyond the scope of this article.

[3] Note that MLE would have other justifications. See section 5 below.

$$\boldsymbol{\theta}_{MLE} - 1.96 * \boldsymbol{SE} < \boldsymbol{\theta} < \boldsymbol{\theta}_{MLE} + 1.96 * \boldsymbol{SE},$$

where **SE** is the standard error of the estimator. A commonly held "correct" interpretation of a CI, regardless of whether one is Frequentist or Bayesian, is that, if we repeat the process of obtaining the data of size n, estimating the unknown parameter, and calculating the $100 -$ alpha% CI, then the proportion of the interval, which covers the true value will be $100 -$ alpha (as long as the specified model is well designed). A particular interval of a CI encountered in scientific works is a realization of this stochastic event. Thus, the following typical statement is a "misinterpretation": "The specific 95% confidence interval presented by a study has a 95% chance of containing the true effect size" (Greenland et al. 2016; p.343), although many practitioners commit this fallacy (as found in an empirical study about psychologists conducted by Hoekstra et al. 2014). This claim comes from the fact that, for strict Frequentist, "it no longer does, once we insert numerical values for the endpoints. Then, no probability (except 0 or 1) can be attached to the event that the interval contains the parameter: either it does or it doesn't (Pratt 1961; p.165)" because, for Frequentists, a probability is defined only as a relative frequency of an event. This is in contrast to the Bayesian confident interval.

The above argument also applies to hypothesis testing problems. Suppose a researcher conducts experiments to examine the treatment effect of a newly developed drug. She randomly assigned n people into two groups where she prescribed the new drug for the treatment group and a placebo for the control group. Using the obtained data, she needs to decide which hypothesis

$$H_0\colon \theta = 0,\ H_1\colon \theta \neq 0,$$

should be accepted where θ, the true effect of the new drug. She conducts hypothesis testing to decide whether a null hypothesis should be rejected or accepted. In the Frequentist hypothesis testing framework, originally established by Neyman and Pearson (1928 a, b)[4], we can control error probabilities of this testing when this procedure is applied to repeatedly collected data, but we cannot talk about how strongly each particular hypothesis is rejected or accepted.

Frequentist methods, however, have been criticized by Bayesian proponents. The stopping-rule problem is one of the major arguments against Frequentist statistics in this hypothesis testing situation (Lindley 1957, Edwards et al. 1963, Berger and Wolpert 1988, Royall 1997, Sober 2008). Suppose, again, a hypothesis testing scenario of a developed drug. For example, when the experimenter finished collecting

[4] Fisher developed another Frequentist framework called "significance testing". The stopping-rule problem matters for both hypothesis testing and significance testing as long as these methods violate the Likelihood Principle (see below for more about this principle).

100 samples for each group and submitted the whole dataset to a statistician, it was still not possible to calculate the correct error probability and decide whether the null hypothesis should be rejected as long as the statistician did not know the stopping rule of this experiment, that is, whether she had determined the sample size to be 100 before the experiment and conducted the experiment accordingly or whether she changed the plan for some reason (e.g., she continued experimenting and obtained more data when the first results did not support her desired hypothesis). Frequentists have argued that the results of testing should be adjusted if she decided when the experiment is to be stopped, because the nominal error probability would not be satisfied in this case. If she had this option, the experiment is considered an "optional stopping" case, for which Frequentists argue that the significance level/error probability must be adjusted to account for the optional stopping.

One problem of this adjustment is that a stopping rule is just an intention in the experimenter's mind (Berger and Wolpert 1988). Although some theoretical studies have formulated adjustment rules for optional stopping (e.g., Armitage 1958), they might not work well as long as it is not possible to access the exact predetermined plan, which is not always feasible. Although some scientific journals are introducing preregistration systems and it would be helpful to enhance the transparency of data collection protocols, this type of system may not be suitable for some fields of research (e.g., medical doctors of a rare disease may find it difficult to predict the number of available subjects at the time they started a new research project). Worse, the preregistration approach cannot prevent unethical scientists from hiding the existence of the actual data at hand, which is used to dredge hypotheses and deceive editors as if the "predetermined" plan is still not carried out[5]. Establishing a formal statistical framework that is indifferent to the stopping rules may provide practical benefits.

Bayesians have argued that Bayesian statistics offer such a framework. According to them, the source of these problems is the fact that the Frequentist procedure violates the Likelihood Principle (LP). The LP is a normative statement that "[t]he information brought by an observation x about θ is entirely contained in the likelihood function $L(\theta|x)$. Moreover, if x_1 and x_2 are two observations depending on the same parameter θ, such that there exists a constant c satisfying $L_1(\theta|x_1) = c\ L_2(\theta|x_2)$ for every θ, they then bring the same information about θ and must lead to identical inferences" (Robert 2007, p. 16; notations modified to the subscript).

This principle has played a central role in the history of Bayesian statistics since it was first formulated by Birnbaum (1962) and generalized by Berger and Wolpert (1984). According to the LP, any statistical inferences should depend only on what has happened. Adjusting based on a stopping rule, however, means that adequate

[5] One might argue that this is a kind of data fabrication and that any statistical account cannot prevent this case. This is partly true, but cheating a data collected date is easier than misrepresentations of the dataset itself, and harder to detect.

inferences are determined by the experimental design and its sample space. In other words, Frequentist inferences are based not only on what happened but also on what could have happened. Berger and Wolpert (1984) discussed the virtue of the LP and concluded that the Frequentist's dependency on the stopping rule seems "very strange" because they are unable to analyze data even when they already have the data at hand.

This poses another related problem in the practice of science: the need to consider that stopping rules hinder the flexibility and simplicity of data analysis (Edwards, Lindman and Savage 1963). Rouder, an empirical phycologist, argues that:

> there is a sense in which optional stopping seems like a smart thing to do. We seemingly should monitor our results as they come in, and we should end early when the results are clear and perhaps keep going when they are not. (Rouder 2014, p.301)

In any case, dynamic increases in empirical data are a fundamental nature of scientific inquiry because "[i]n practice, of course, we typically receive data in successive stages" (Bernardo et al. 1994; p.47). Bayesians have argued that incorporating a formal account of "update-by-data" into statistics is needed in the scientific community and that Bayesian philosophy offers these dynamics in a straightforward and coherent manner.

In summary, the stopping-rule problem poses serious difficulties for the Frequentist theory of statistical inference, and its solution has not been attained on its own grounds. Indeed, typical responses by Frequentists are to simply dismiss the issue or to try to construct arguments that dependence on a stopping rule is required (Mayo 2018) to control the error probability. Although several procedures for controlling risk induced by optional stopping have been proposed, they restrict the applicability and flexibility of the Frequentist methods within the limited fields of scientific practice. Worse, these procedures are not the answer to the critics' concerns that the exact stopping rule, which is required for the correct adjustment of the significance level, is merely an intention in the experimenter's mind and not accessible in general.

3. The Bayesian and the catch-all hypothesis problem

Bayesian statistics , as previously mentioned, is another school of statistics and is becoming more popular among scientists. Bayesian statistics offers formal procedures for point estimation, interval estimation, hypothesis testing, model selection, and model averaging, but with a different justification; at least historically, subjective Bayesian theory has played a central role. Here, we briefly review some basic concepts of this theory.

Suppose a researcher wants to evaluate how a hypothesis, H, is plausible given her prior degree of belief in H, and collected observations *Data*. The Bayes theorem gives how she should update her belief on H:

$$P(H|Data) = \frac{P(Data|H)P(H)}{P(Data)},$$

where $P(H|Data)$ is the posterior probability of H given *Data*, $P(Data|H)$ is the "likelihood", $P(H)$ is prior probability and $P(Data)$ is the probability of total evidence. Unlike Frequentist methods, it offers a direct way to evaluate a scientific hypothesis at hand, in the sense that probability gives quantitative plausibility of this hypothesis; according to Bayesian theory, elaborated arguments[6] have tried to guarantee a rationale to express epistemic uncertainty by a probability function. From this perspective, Bayes' theorem describes the dynamics of changing uncertainty and thus acts as a common ground to express the plausibility of a hypothesis before/after data are obtained. Bayesianism argues that these coherent natures are appealing to scientists.

There exists, however, the catch-all hypothesis problem, which is an argument against the Bayesian approach, in which a posterior probability of a hypothesis H is not calculable in many cases of scientific practice. When the set of candidate hypotheses is finite and known, $P(Data)$ can be "decomposed" in the following manner:

$$P(Data) = \sum_i P(Data|H_i)P(H_i).$$

In scientific practice, however, it is not straightforward to determine how we should apply this rule. Suppose the posterior probability that the general theory of relativity (GTR) is true (Sober 2008). According to Bayes' Theorem,

$$P(GTR|Data) = \frac{P(Data|GTR)P(GTR)}{P(Data|GTR)P(GTR) + P(Data|\neg GTR)P(\neg GTR)}.$$

A set of "$nonGTR''$ theories contains *all* possible theories rather than GTR and is called a catch-all hypothesis. It includes a theory that no one had come up with thus far. How can we obtain $P(Data|\neg GTR)$, the likelihood of data under the catch-all hypothesis? This is called the catch-all hypothesis problem (Sober 2008). Again, as discussed about the stopping-rule problem, we see why the catch-all hypothesis problem matters. In practice, the emergence of a novel theory is a fundamental way by which scientific inquiry is driven. It is not realistic to assume that all the candidate hypotheses are known and that their likelihood is calculable. Some might reject the claim that the probability of total evidence is not calculable. Even when a researcher

[6] For example, the Dutch-book argument by de Finetti; 1972 or an axiomatized rationality by Savage; 1954. See, for example, Sprenger and Hartmann 2019 for a brief review.

cannot exactly obtain $P(Data|\neg GTR)$, she might elaborate an approximation, which is conditional on all the *known* non-GTR hypotheses. However, it does not give a solution because, as Fitelson (2008) pointed out, a seemingly improbable (very low likelihood) hypothesis could have a huge influence on the catch-all likelihood, and thus a finite approximation of the catch-all likelihood is not reliable.

Some might respond that a catch-all likelihood is calculable in the practice of statistical analysis when the set of hypotheses is well defined on the space of parameters of a statistical model. Suppose $\theta \in \Theta$ is the parameter of interest. The catch-all likelihood is exactly defined as

$$P(Data) = \int P(Data|\theta)P(\theta)d\theta,$$

even though some numerical problems are known to solve the integrations. This is called the marginal likelihood and plays a great role in model selection, model averaging, and hypothesis testing in Bayesian theory.

Again, it is not a promising approach. To interpret the posterior based on this form as the degree of belief that the hypothesis is true, it is necessary that the true value exists in Θ. In other words, the supposed class of model Θ is true. However, this is not realistic because, in the practice of statistical analysis, many researchers believe the model at hand is *not* true, even as an approximation. Rather, Bayesians admit there would be a case where one has several candidate models but no one candidate model is assumed to be true (some have called this situation an "M-open case"). Constructing useful statistical inferences in such scenarios is a focus of both theoreticians and practitioners (e.g., Bernardo et al. 1994; Walker 2013; Kelter 2021). While the catch-all hypothesis problem is avoided in these cases, because $P(Data)$ is exactly defined as the marginal likelihood, it comes with a dilemma; we cannot interpret a posterior distribution as "the degree of belief that the hypothesis is true" at face value (see also Vassend 2019). This poses a serious problem for the traditional account of the subjective Bayesian method.

In summary, the catch-all hypothesis problem is also the source of serious troubles for Bayesians and is still not solved. While the typical statement "the posterior probability gives the degree of the belief that a hypothesis is true" has been the foundation of the traditional Bayesian philosophy, this virtue will be lost as soon as it is apparent that the catch-all likelihood cannot be calculated. Even though current practitioners appear to avoid this problem by limiting the set of hypotheses and obtaining the marginal likelihood, this comes with another difficulty: to make the catch-all likelihood calculable is not the same thing as solving the catch-all hypothesis problem nor *keeping the validity of the "degree of belief" interpretation intact.*

4. Motivating examples (case studies from recent theories of statistics and the philosophy of statistics)

There are, as discussed above, differences in strategy between how Frequentists and Bayesian methods are justified, and we considered the stopping-rule problem and catch-all hypothesis problem separately. We observed that these two problems pose serious difficulties for each Frequentist and Bayesian method and concluded that neither is still solved within their own traditional philosophical foundation. As both problems have practical consequences, some treatment is needed. To do so requires a revision of the predominant account of statistics that regards Frequentist and Bayesian methods as exclusive and/or conflicting.

A natural idea for this revision is to articulate unified methodologies of statistics. Indeed, philosophers have good reasons for thinking seriously about this matter. First, there are some *hybrid* methods of Frequentist and Bayesian statistics in the recent theory of statistical analysis. Second, even some philosophers of statistics have cast doubt on a traditional view of the one-to-one relationship between interpretations of probability and schools of statistics. These facts pose a fundamental question: how should we interpret the presumed dichotomy of Frequentist and Bayesian statistics? How do these facts affect the implications of two problems in the practice of the scientific community? In this section, some cases of hybrid methods are introduced, which will, in turn, be a motivating case for section 5. Some philosophical works are also discussed later.

4.1. Several insights from recent methodologies

First, we consider insights from recent discussions by practitioners. We focus on Frequentist aspects of recent Bayesian methods. Some Bayesian statisticians do not accept the traditional "update-of-degree-of-belief" interpretation of probability and statistical methods. For example, Gelman and Shalizi (2013) discussed, from a practitioner's view, that,

> a Bayesian inference was not for computing the posterior probability that any particular model was true – we never actually did that. (Gelman and Shalizi 2013, p.13)

Furthermore, prior distributions used in applied work represent statisticians' states of knowledge and beliefs before examining their data (ibid. p.19) because, according to them, it is "like a regularization device" (ibid, p.19; i.e., a statistical method to stabilize an estimate). In fact, some Bayesian methods adopt a "Frequentist" justification for parameter estimation and model selection.

4.1.1. Parameter estimations

Suppose, for example, a linear regression model to investigate the relationship between an objective variable and explanatory variables.

$$\boldsymbol{y} \sim Normal(\boldsymbol{X}\boldsymbol{\beta}, \sigma_y^2),$$

where $\boldsymbol{y}$ is the data (n-dimensional vector) of the objective variable, $\boldsymbol{\beta}$ is the d-dimensional vector of the parameters to be estimated by the data, $\boldsymbol{X}$ is the data matrix of explanatory variables ($n \times d$), and σ_y is the SD of $\boldsymbol{y}$ given $\boldsymbol{X}$. This regression is one of the most basic models and has been widely applied among practitioners. To estimate $\boldsymbol{\beta}$ and to evaluate its uncertainty, both the MLE and Bayes estimator with a prior is applicable.

However, technical problems arise when d is very large, the so-called high-dimensional case. In particular, the MLE is not uniquely defined when $n < d$, although many practical scientists demand this situation. For example, biologists want to explore which single nucleotide polymorphism (SNP) is relevant to the development of a disease. While a set of candidate SNPs (explanatory variables) contains $>$ 10,000 sites, the number of subjects can be no more than $\approx$ 100. Establishing a valid statistical estimator in such cases is an active research field.

Carvalho et al. (2010) proposed the Horseshoe prior distribution for these situations. This prior puts a very large mass around $\boldsymbol{\beta} = 0$ but has heavy tails. From the subjective Bayesian point of view, this prior distribution represents the prior belief that most of the candidate variables are irrelevant to $\boldsymbol{y}$, but some of them might take a large value. A Bayes estimate with this prior gives a unique estimate of $\boldsymbol{\beta}$ even when the MLE is not obtained. Some researchers, however, found that this prior has desirable properties even in the Frequentist sense. Van der Pas et al. (2017) found that the Bayesian credible interval obtained by this prior has good “frequentist coverage” (Van der Pas et al. 2017, p.1221) properties, as does the 95% CI of usual cases. This can be viewed as a Frequentist justification of a Bayesian method.

Furthermore, the fact that they take a hybrid approach (a mixture of Frequentist and Bayesian) is explicitly stated by some authors of high-dimensional statistics studies. Castillo et al. (2015) considered another type of prior and explained their proposal as follows:

> We study a Bayesian approach based on priors that set a selection of coeffi cients β_i [an element of the vector $\boldsymbol{\beta}$] a priori to zero; equivalently, priors that distribute their mass over models that use only a (small) selection of the col umns of X. Bayes’s formula gives a posterior distribution as usual. We study this under the “frequentist” assumption that the data Y has in reality been gen erated according to a given (sparse) parameter β^0 [the true value]. The expecta tion

under the previous distribution is denoted $\mathbb{E}_{\beta^0}$. (Castillo et al. 2015, p.1986; double-quotation original)

We see that the same strategy is adopted to justify a parameter estimation method to the MLE of section 2 (the expected performance of a method under the "replication" of true data distribution).

4.1.2. Model comparisons by information criteria

Another example of a hybrid Bayesian method is found in research field on information criteria. Suppose a researcher has multiple statistical models to be deployed and wants to select the "best" model. This type of task is called the model selection problem. One way to select a model is to compare the predictive performance (typically, it is defined as the Kullback-Leibler divergence between the model and the true data generating distribution) when these models are applied to unobserved future data. Varieties of information criteria offer a tool to examine prediction accuracy without obtaining actual future data (Konishi and Kitagawa 2008). The most well-known criterion is the Akaike information criterion (AIC; Akaike 1974):

$$\mathrm{AIC} = -2 * \log[L(x; \boldsymbol{\theta}_{MLE})] + 2k,$$

where k is the dimension of the parameters for which the maximum likelihood estimator is applied. Akaike (1974) showed that AIC is a good measure to compare the prediction performance among different models because $\mathbb{E}_{x \sim X}[\mathrm{AIC}]$ is an asymptotically unbiased estimate of the Kullback-Leibler divergence up to a constant (this constant is irrelevant to model comparisons). Note that, again, the expectation is over x, the true data generating distribution. We see, again, that the Frequentist strategy is applied to evaluate the method, as in the maximum likelihood method above.

Some Bayesian statisticians have adopted this strategy in exchange for the loss of their traditional justification of Bayesian theory. Sprenger (2013) reviewed the mathematical backgrounds of three major information criteria for Bayesian models—Minimum Massage Length, Bayesian Information Criterion, and Deviance Information Criterion—and pointed out that they lack the fundamentals of "Bayesian philosophy". The deviance information criterion (DIC; Spiegelhalter et al. 2002, 2014) was, for example, one of the most applied information criteria among practitioners for comparing hierarchical Bayesian models.

$$\mathrm{DIC} = D + pD,$$

where D is the posterior mean deviance $\mathbb{E}_{\theta|y}[-2 \log p(y|\theta)]$, and pD is the "posterior mean deviance - deviance of posterior means", defined as

$$\mathbb{E}_{\theta|y}[-2\log\{p(y|\theta)\}] + 2\log[p\{y|\tilde{\theta}(y)\}],$$

where $\tilde{\theta}(y)$ is a point estimator of θ using data y (typically, posterior mean). This criterion aims to give a measure of predictive performance when AIC is not useful because 1) the model structure is too complex (e.g., hierarchical model) and 2) the Bayesian method (not the maximum likelihood estimator of AIC) is applied to estimate unknown parameters. Again, the objective of DIC is to construct an unbiased estimate of predictive accuracy (defined in terms of Kullback-Liebler divergence) given a particular model.

Having considered these issues, Sprenger (2013) pointed out the implicit assumption "that Bayesian model selection is firmly anchored in Bayesian philosophy is usually not satisfied" (Sprenger 2013, p. 113); that is, the model selection is usually conducted outside the scope of the subjective Bayesian theories and that "ideas and techniques from different philosophies [. . .] may be mixed" (ibid. p.112) in DIC. In other words, many pragmatic rationales rather than the subjective Bayesian theories play an indispensable role in justifying this "Bayesian" procedure, including the Frequentist approach.

After Sprenger's work, statisticians have pointed out several limitations of DIC (reviewed in Spiegelhalter et al. 2014) and a novel criterion from a machine learning literature gained popularity among some practitioners (e.g. Widely Applicable Information Criterion; Watanabe 2010, Gelman et al. 2013). Nevertheless, the main rationale behind the model selection is similar to DIC in the very sense that its objective is to provide a measure of the predictive performance of a Bayesian model based on the Kullback-Liebler Divergence.

4.2. Philosophical grounds of the Neyman-Pearson framework revisited

The second reason to reconsider the traditional interpretations of Frequentist-Bayesian relationships is found in some philosophical works. Recently, some philosophical works of statistics have cast doubt on a common view of a one-to-one relationship between interpretations of probability and schools of statistics. Take Mayo (2018), for example., who has long defended the Neyman-Pearson framework of hypothesis testing, which we mentioned as a Frequentist method in section 2 (Mayo 1996, 2018). However, she adopts a two-stage strategy to accomplish her objective. The first approach is to criticize Bayesian philosophy. She starts her argument by the formulation of an abstract principle named the "(weak and strong) severe requirement":

> Severity Requirement (weak): One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data x agree with a claim C but the method used is practically guaranteed to find such agreement, and little or no capability of finding flaws with C even if they exist, then we have

> bad evidence, no test. (Mayo 2018, p.5)
> Severity Requirement (strong): We have evidence for a claim C just to the extent [that] it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing results, x is evidence for C. (ibid. p. 14)

These two requirements represent intuitive concepts regarding what kind of property is desired for scientific investigation (including statistical methods). The former weak version posits a qualitative concept of evidence: for an extreme case, it is natural to say you cannot trust a statement that claim C, according to method f, is probably true if f produces the same results for everything. The latter strong version posits the quantitative concept of evidence: if we have an objective measure of the stringency of method f, then this measure is useful to measure the strength of evidence for C given data. For most scientists, if not all, it seems natural to accept these two requirements.

Based on these concepts, she argues that the Bayesian philosophy of statistics violates the severe requirement. Note that she does *not* reject Bayesian *methods*, admitting that Bayesian statistics are useful and valid in many scientific applications. For example, she examined the argument considered above by Gelman and Shalizi (2013). She concluded that "[t]here is no difficulty with prior" as long as its role is pragmatic like "a regularization device" to smooth the likelihood[7]. Even though there can be counterarguments against such a "pragmatic Bayesian", a subjective Bayesian must consider the fact that at least some practitioners consider their prior distribution, not their subjective belief. Mayo claims that these methods should not be based on Bayesian philosophy or interpreted as a rational degree of belief, although most Bayesians try to do so. Rather, she argues that such methods should be based on the severe testing concept.

The second approach to defend the Neyman-Pearson framework is to reconsider well-accepted views, which have even been accepted by Frequentist philosophers and statisticians. As described above, this framework is usually construed as a method that gives a binary result (a hypothesis is accepted or rejected) and has valid meanings as long as the procedure is applied repeatedly (i.e., long-run performance). However, why does the fact that a method is justified on a property of expectation over x, the true distribution, yield the claim that the results of a Frequentist method are meaningless for each particular case? She rejects the prevailing long-run performance view of the Frequentist method and argues that the severe requirement gives another meaning to these procedures: if the Neyman-Pearson framework has high severity

[7] Mayo (2018, p.435). Shimatani (2021) showed another pragmatic virtue of the Bayesian statistics like convenience to construct a complex model and to conduct numerical optimizations. See this special issue.

(i.e., has a high probability of finding an error, if there are any), then a particular hypothesis that passes this framework is probably a well-justified claim.

5. Relationships between the two theories and problems

In the previous section, we saw three cases to reconsider a dichotomized view of statistics. There exists a "hybrid" Frequentist-Bayesian approach in terms of both theoretical and philosophical studies. Some might argue that since Frequentist/Bayesian distinctions are meaningless, so are their alleged problems; that is, the stopping-rule and catch-all hypothesis problems are irrelevant to any current statistical theory or practice. Alternatively, with more cautiousness, she might discuss that, hopefully, these "hybrid" approaches will be applied to simultaneously solve the two problems. However, there exists good reason to doubt this kind of statement because the philosophical status of these methods is not fully clarified and, thus, its relationships with the stopping-rule problem and the catch-all hypothesis problem are not obvious. Indeed, whether Bayesian methods are irrelevant to a stopping rule is recently attracting a renewed focus of practitioners since "Bayesian" methods are becoming popular as a remedy for replication crises (e.g., Erica et al. 2013; Rouder 2014). How should we place the stopping-rule problem and the catch-all hypothesis problem? To integrate the two theories of statistics and to solve both problems, the "hybrid" methods described above are not sufficient, and an integrated viewpoint is needed to see how these problems arise.

It is important to distinguish between a method itself and its justification. For example, the justification of the MLE method is not necessarily based on Frequentist philosophy; for Bayesian theory, the MLE may be an acceptable approximation of the posterior estimate when the prior is almost negligible. Likewise, the distinction between a method and its justification gives a more precise meaning of the case studies in the previous section. The first two cases of section 4, the high-dimensional estimation and the model selection, can be seen as a Frequentist justification of a Bayesian method because these methods adopt prior probability but show "good" performance in the sense of the coverage probability and the expected prediction performance. The third example, Mayo's severe testing account, tried to provide grounds for a common justification for both Frequentist (Neyman-Pearson hypothesis testing) and Bayesian methods based on her severe requirement. Establishing a relationship between the stopping-rule problem and the catch-all hypothesis problem based on how statistical justifications work would be a starting point to obtain philosophical insights and to make productive suggestions for both theorists and practitioners of statistics. In this section, the relationship between Frequentist and Bayesian statistics is examined from a novel viewpoint of "evaluation". It is then argued that the stopping-rule problem and the catch-all hypothesis problem are just the mirror image of the same problem from this viewpoint.

Let us leave specific statistical procedures aside for the moment and consider a more general problem: what is the fundamental problem of an evaluation? Suppose you evaluate a target X on a condition of Y. Such a type of evaluation takes place in various scenes of our daily life (e.g., inspecting a malfunction of home appliances), public probes (e.g., investigating a suspicious politician), and scientific inquiries. In these cases, the results of the evaluation depend on Y and thus change when a different situation is assumed; if a condition Y changed to Y', then the evaluation for X would also change (from X conditional on Y to X conditional on Y'). In this sense, we can compare how the evaluation of X changes assuming various hypothetical situations of Y. However, things are completely different if the target X changed. Even when Y is exactly the same, the meaning of Y also changes because how Y affects the evaluation depends on the purpose for which Y is used. Thus, the continuous update of the evaluation (from X conditional on Y to X' conditional on Y) might not be possible even when we have the same Y.

Now let us turn back to our objective, the examination of statistical methods: in the context of statistical analysis, what is the target of evaluation and what is the condition? A Frequentist strategy of justification is achieved by an evaluation of data based on a set of parameters. If the target of an evaluation is data, you can compare whether a certain data is likely under different sets of parameters but cannot change the data itself. A Frequentist justification cannot allow the continuous update of data because the target of the evaluation is the data itself. This is the very reason that the stopping-rule problem arises within the Frequentist strategy of statistics. Even when one uses a "hybrid Bayesian" method, as discussed in section 4, the curse of the stopping rule is inevitable as long as they adopt this strategy for their justification of inferences. For Bayesians, on the other hand, methods are based on the posterior probability distribution, which is an evaluation of hypotheses. Various inferences include point estimation (e.g., posterior mean, posterior median, maximum a posteriorti), confidence interval, and hypothesis testing. Here, the inference is based on the evaluation of parameters conditional on a set of data. If the target is a hypothesis, the set of hypotheses to be considered cannot be changed. Such a situation induces the catch-all-hypothesis problem.

6. Discussion and conclusion

In this paper, two well-known problems are reviewed, namely, the stopping-rule problem and the catch-all hypothesis problem of the Frequentist approach and Bayesianism, respectively. I argued that some treatments for both of the problems are required because they impose practical impacts. Then, I examined how recent studies, from both statistics (parameter estimation in high-dimensional cases and model selection by information criterion) and the philosophy of statistics, pursue "hybrid" approaches of Frequentist/Bayesian statistics and posed a question of whether these "hybrid" approaches could be a panacea for the current issues. Since the traditional account of the Frequentist-Bayesian distinction based on the interpretation of probability is not enough for this project, I propose a novel viewpoint to examine the differences between the two theories.

Using this concept, I revisited the relationship between the stopping-rule problem and the catch-all hypothesis problem and concluded that they are just a mirror image of the same problem. Furthermore, the analysis in this paper shows that the recent "hybrid" approach does not give the solution because, according to our analysis, the two problems stem from the fundamental difference in strategy rather than differences in technical details; that is how, or on what grounds, do we justify a statistical inference. As long as the current strategy of epistemological justification is taken for Frequentist/Bayesian methods, the two problems are inevitable. Thus, if we want to choose just one strategy of either the Frequentist or Bayesian justification, then we would have to abandon either the updating of data or the updating of the hypothesis. Of course, both of them are essential to the nature of science, and something that allows "bridging" the two evaluations are needed.

One approach to bridge them is the posterior predictive distribution. Gelman and Shalzi (2013) discussed how the traditional "belief updating by data" account of Bayesian analysis is problematic and how checking the model by the posterior predictive distribution plays an important role. According to them, the current practice of Bayesian data analysis involves a "goodness of model" assessment to ensure that the model accurately captures the characteristics of the data at hand. If not accurate enough, the structure of the model would be modified, and then the data would be reanalyzed by this new model and the Bayesian estimate. In practice, not only the parameters but also the set of models (i.e., a set of hypotheses) would be updated, which would open the possibility of both an update-of-data and an update-of-hypothesis. However, model checking by posterior predictive distribution suffers from several problems. Although Gelman and Shalzi (2013) stressed the role of qualitative checking (e.g., plotting the raw data and fitted regression curve), quantitative approaches would be preferred. While they also proposed a more quantitative approach by posterior predictive p-value (ppp) to decide whether to reject the model or not, it would cause the problem of an arbitrarily chosen threshold, which many researchers have pointed out with usual significance testing.

Another approach to fill the gap is marginal likelihood. Usually, marginal likelihood has been considered a Bayesian quantitative index of a model, by which various statistical inferences, including hypothesis testing, model selection and model averaging, are made (Kass and Raftery 1995). However, recent theoretical studies have revealed that it is also related to Frequent properties of prediction error (e.g., Fong and Holmes 2020).

There is still a long way to go, but surely we need to find some way to fill this gap not only for the sake of the theories but also for the sake of statistical practice. Nevertheless, we must first recognize the true difference between Frequentism and Bayesianism, as discussed in this paper.

Acknowledgments

The core idea of this paper was presented at the 16$^{\text{th}}$ Congress of the Logic, Methodology, and the Philosophy of Science and Technology, Aug. 10. Prague, Czech Republic. I thank all the feedbacks from the participants. I also thank many valuable comments from anonymous reviewers.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research is partly funded by the Grant for the Groundbreaking Young Researchers (2018) by Suntory Foundation and the JSPS KAKENHI (Grant Number 16H03050, 20H01736, 21K15170).

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716–723.

Armitage, P. (1958). Sequential methods in clinical trials. American Journal of Public Health and the Nations Health, 48(10), 1395–1402.

Berger, J. O., & Wolpert, R. L. (1988). The likelihood principle. Institute of Mathematical Statistics, Hayward.

Bernardo, J. M. & Smith, A. F. M. (1994). Bayesian Theory. John Wiley & Sons, Hoboken

Birnbaum, A. (1962). On the foundations of statistical inference. Journal of the American Statistical Association, 57(298), 269–306.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. Biometrika, 97(2), 465–480.

Castillo, I., Schmidt-Hieber, J., & Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. The Annals of Statistics, 43(5), 1986–2018.

Clyde, M., & Iversen, E. S. (2013). Bayesian model averaging in the M-open framework. Bayesian theory and applications, 14(4), 483–498.

de Finetti, B. (1972). Probability, Induction and Statistics, Wiley, New York.

Fong, E., & Holmes, C. C. (2020). On the marginal likelihood and cross-validation. Biometrika, 107(2), 489–496.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. Psychological Review, 70, 193–242.

Erica, C. Y., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. Psychonomic bulletin & review, 21(2), 268–282.

Fisher, R. A. (1925). Statistical methods for research workers. Genesis Publishing, Guildford.

Fitelson, B., & Thomason, N. (2008). Bayesians sometimes cannot ignore even very implausible theories (even ones that have not yet been thought of). The Australasian Journal of Logic, 6, 25–36.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology, 66(1), 8–38.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. CRC press, Boca Raton.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology, 31(4), 337–350.

Hacking, I. (1965). Logic of Statistical Inference. Cambridge University Press, Cambridge.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. Psychonomic bulletin & review, 21(5), 1157–1164.

Howson, C., & Urbach, P. (1991). Bayesian reasoning in science. Nature, 350(6317), 371–374.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 221–233.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773–795.

Kelter, R. (2021). Bayesian model selection in the M-open setting—Approximate posterior inference and subsampling for efficient large-scale leave-one-out cross-validation via the difference estimator. Journal of Mathematical Psychology, 100, 102474.

Konishi, S., & Kitagawa, G. (2008). Information criteria and statistical modeling. Springer Science & Business Media, Berlin.

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

Mayo, D. G. (1996). Error and the growth of experimental knowledge. University of Chicago Press, Chicago.

Mayo, D. G. (2018). Statistical inference as severe testing. Cambridge University Press, Cambridge.

Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika, 175–240.

Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. Biometrika, 263–294.

Pratt, J.W. (1961). Testing Statistical Hypotheses. Journal of the American Statistical Association, 56, 163–166.

Robert, C. P. (2007). The Bayesian choice: From decision-theoretic foundations to computational implementation. Springer-Verlag, New York.

Romeijn, J. (2017). Philosophy of Statistics, In. Edward N. Zalta (ed.) The Stanford Encyclopedia of Philosophy (Spring 2017 Edition).

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. Psychonomic bulletin & review, 21(2), 301–308.

Royall, R. (1997). Statistical Evidence: A Likelihood Paradigm. CRC Press, Boca Raton.

Savage, L. J. (1954). The Foundations of Statistics. Wiley, New York (2$^{\text{nd}}$ edition 1972), Dover, New York.

Sober, E. (2008). Evidence and evolution: The logic behind the science. Cambridge University Press, Cambridge.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B: Statistical Methodology, 64(4), 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2014). The deviance information criterion: 12 years on. Journal of the Royal Statistical Society: Series B: Statistical Methodology, 76(3), 485–493.

Sprenger, J. (2013). The role of Bayesian philosophy within Bayesian model selection. European Journal for Philosophy of Science, 3(1), 101–114.

Sprenger, J., & Hartmann, S. (2019). Bayesian philosophy of science. Oxford University Press, Oxford.

Vassend, O. B. (2019). New Semantics for Bayesian Inference: The Interpretive Problem and Its Solutions. Philosophy of Science, 86(4), 696–718.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. The Annals of Mathematical Statistics, 20(4), 595–601.

Walker, S. G. (2013). Bayesian inference with misspecified models. Journal of Statistical Planning and Inference, 143(10), 1621–1633.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11, 3571–3594.

(Received 2021.1.19; Revised 2021.5.11; Second Revised 2021.6.22; Accepted 2021.7.7)