Revisit to the Likelihood Principle

Masahiro MATSUO*

Abstract

Undoubtedly, whether to accept the likelihood principle or not has been, and still is, one of the most crucial issues for philosophical debates on statistics, though interests in it are waning from statistical debates due to general preferences for more practical issues of statistics. The principle says all you need in parameter analyses of a statistical model is found in likelihood for the data obtained. Bayesians and likelihoodists have traditionally regarded this principle as fundamental, declining any forms of statistics which violate it. Frequentism, on the other hand, try to reject this principle, upholding error probability as a more crucial factor for statistical analyses. But arguments made so far on the likelihood principle still seem to stay on those as to what principle we prefer to choose in statistical analyses. The validity of this principle seems to have never been explored fully enough through the arguments on either side. In this paper, I briefly review how these arguments have been made and show some difficulty in maintaining the principle. I think this has some impact upon statistical practices as well.

Key words: likelihoodism, frequentism, evidentialism, likelihood principle, law of likelihood, error probability, distance to the truth

1. What is the likelihood principle, and what's not?

The idea of the likelihood principle (LP) could be traced back to Fisher (1956), though he held back from using this as a main principle for statistical analyses (the word 'likelihood' can also be traced back to Fisher (1921)). The principle has been formulated in slightly different manners by different statisticians, but the essence is the same. Here is the formulation of LP by Berger (1985: 28).

In making inferences or decisions about θ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x . Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ).

---

* School of Science, Hokkaido University
E-mail: matsuou@sci.hokudai.ac.jp

As seen above, LP consists of two parts. In the first half, it says likelihood is the minimal sufficient statistic. The latter half shows what is involved in a statistical analysis if the minimal sufficient statistic T is likelihood. We can easily see this conditional hold. Suppose we have two data sets E and E' . And suppose for every single data $x \in E$, $x' \in E'$, the equation below holds.

$$P(x|\theta_i) = cP(x'|\theta_i).$$

‘ c ’ is a positive constant, and θ_i is a parameter in a statistical model, which is the target of a statistical analysis (the equation holds for every i). The basic assumption shared broadly among proponents of LP is that the statistical analysis of a parameter should be based on the ratio between two likelihoods: a relative comparison of two hypotheses (what we should derive is the ratio between likelihoods of two single hypothetical values of a parameter). If for a θ_i , a hypothesis $H_1: \theta_1$ and $H_2: \theta_2$, and if the above equation holds for $x \in E$, $x' \in E'$, then the likelihood ratio (LR) for E is always equal to the one for E' as shown below.

$$\text{LR}_E = \frac{P(E|H_1)}{P(E|H_2)} = \frac{cP(E'|H_1)}{cP(E'|H_2)} = \text{LR}_{E'}.$$

This equation holds not only for likelihoodism, which maintains straightforwardly likelihood is a minimal sufficient statistic (with some exception), but also for Bayesianism, whose updating mechanism solely depends on the likelihood ratio when two hypotheses are compared (this ratio is usually called ‘Bayes Factor’)¹.

Probably the most important implication of LP is that any statistical analysis cannot be influenced by those data which are not actually obtained: counterfactual dependence on possible but not obtained data is irrelevant to statistical analyses (Royall 1997). Consequently, frequentism’s sampling theories, which depend on a model of sample space, cannot consist with LP (we will see how frequentists avoid (not deny) LP in the next section). As long as we hold on to LP as the basic principle in statistics, we cannot incorporate counterfactual data into the analysis (‘evidential’ analysis) in any way. I will show finally this implication will affect badly the assessment of the validity of LP.

We need to see also what’s *not* LP. Likelihoodists hold up another principle besides LP: the law of likelihood (LL). LP clearly shows what a minimal sufficient statistic should be like, but does not show how to use it. LL shows the usage of likelihood in the way already mentioned above: for hypotheses $H_1: \theta_1$ and $H_2: \theta_2$, and for observed data x ($x \in X$, X is a random variable), x is evidence supporting

---

¹ Likelihood ratio is not always the same as Bayes Factor, because when parameters are treated as continuous random variables in Bayesian statistics, Bayes Factor is the ratio between marginal likelihoods, which are averaged by prior probability distributions. Likelihoodists generally take priors as unwarranted and unnecessary.

H_1 over H_2 if and only if $L(x, \theta_1) > L(x, \theta_2)$. And the measure of the strength of evidential support is given by $L(x, \theta_1) / L(x, \theta_2)$ (Hacking 1965, Royall 1997, Sober 2008).

Here the word ‘evidence’ is used on the assumption that evidential judgment should be formed not by an evaluation of a single hypothesis, but by a relative comparison of two hypotheses (in spite of the difference in statistic, this idea is also shared with some traditional frequentists, such as Gosset or Neyman-Pearson)². Note that although LL provides explicitly how to use likelihood in the analysis, it doesn’t say anything about the statistical status of likelihood (it doesn’t refer to likelihood as a minimal sufficient statistic). Then we have two variations of likelihoodism: one is that holding both LP and LL (Royall 1997, Edwards 1992, etc. Edwards integrated LP and LL into the likelihood axiom), and the other holding just LL (Sober (2008) appears to fit with this stance). The latter seems rather exceptional among proponents of likelihoodism, but if it is tenable, the scope of likelihoodism would not be confined to actually obtained data in a statistical analysis. This means it might be exempt from LP’s defect, if there’s any. We will get back to this point later.

In the next section, we show first how LP was ‘proved’ by a frequentist, Birnbaum, and how elaborately this proof has been evaded by frequentists including Birnbaum himself. In the third section, we show how LP has been endorsed, in contrast, particularly by likelihoodists. But we also point out a non-negligible deficiency in LP in relation to ‘error rates’, which has been the frequentists’ escape from LP. And in the final section, we discuss some possible compensation for this deficiency and a turning point likelihoodists should face over whether to maintain LP or not.

2. How frequentists evade LP?

In this section, we see how LP was first conceived but at the same time evaded by frequentists.

As mentioned above, LP originated from frequentists’ idea but this principle is opposed to their sampling theories. The one who tackled this twisted problem first is Birnbaum. His solution is twofold: firstly he ‘proved’ LP from two premises which any frequentist is supposed to accept (Birnbaum 1962), but secondly nevertheless abandoned LP for the sake of fruitfulness brought by sampling theories (Birnbaum 1969:). Traditionally frequentists dealt with LP based on this argument and mostly

---

² What an evidence means is not necessarily self-evident even in the context of likelihoodism. But it is clear that those who deal with statistical ‘evidence’ focus on the ‘true value’ of the model parameter in some way or other, which is in distinct contrast with sampling theorists. Evidentialists are particularly in earnest about this, who aim to expand likelihoodism conceptionally and methodologically, providing desiderata of a theory of evidence based on the conception of relative distance to truth (Taper & Lele 2011).

attacked Birnbaum's proof on some different grounds.

2.1. Birnbaum's 'proof'

The proof of Birnbaum is briefly as follows. The goal of his proof is to show that statistical evidence which satisfies LP condition ($P(x|\theta) = cP(x'|\theta)$) is also experimental (inferential) evidence. Suppose we have two possible experiments E_1 and E_2 , and the choice of which is determined by a randomizer (flipping a coin) with some probability for each. Here we can define a mixture experiment E^* as comprised of these two possible experiments. Birnbaum tried to prove LP based on this assumption of experiment. The proof is given deductively from two plausible premises, both of which are called principles.

- (1) *Principle of sufficiency.* Let E_1 denote an experiment and E_2 be its derived experiment with the same parameter space, such that any outcome x of E finds its corresponding sufficient statistic $T = T(x)$. Then for each x , $\text{Ev}(E_1, x) = \text{Ev}(E_2, T)$, where Ev is an evidential inference.

There are several different ways to give a principle for sufficient statistics (and their relation to data). Above seems somewhat devious, but it's sufficient and convenient to the proof.

Now suppose that x and x' be an outcome of respective experiment E_1 and E_2 , and that for some positive c and for all θ , $P(x|\theta) = cP(x'|\theta)$, that is, x and x' have the same likelihood function. From the general theory of sufficient statistics (which was independently introduced by Bahadur), 'if two outcomes x, x' of one experiment E determine the same likelihood function, then there exists a (minimal) sufficient T such that $T(x) = T(x')$ '. Based on this consequence and the principle of sufficiency, we have a lemma: if x and x' of E have the same likelihood function, they have the same evidential inference. That is,

$$\text{Ev}(E, x) = \text{Ev}(E, x').$$

Birnbaum applied this to the mixture experiment E^* above. x and x' are outcomes of respective experiments, E_1 and E_2 , but E_1 and E_2 compose 'one' experiment E^* . Then, if x and x' determine the same likelihood function,

$$\text{Ev}(E^*, (E_1, x)) = \text{Ev}(E^*, (E_2, x')). \quad [1]$$

This constitutes one half of the deductive proof.

- (2) *Principle of conditionality.* This principle is not so mathematically rigorous but what 'many statisticians are inclined to accept for the "evidential inference"'. If an experiment E_1 comprises a mixture experiment E^* with E_2 , the choice of

which is determined by a randomizer, once an outcome (E_1, x) is obtained, evidential inference should be based on this outcome alone, not on other component experiments, which are not actually performed (other parts of E^* are irrelevant to the inference).

It seems quite natural to accept this principle. If we accept it, we have the following equivalence relation.

$$\text{Ev}(E^*, (E_1, x)) = \text{Ev}(E_1, x). \quad [2]$$

Then from [1] and [2], we can deductively conclude:

$$\text{Ev}(E_1, x) = \text{Ev}(E_2, x').$$

What this means is that the proportional relation of likelihoods implies equivalence of evidential inference. This is how Birnbaum ‘proved’ LP (he also showed two principles can be derived from LP).

2.2. How frequentists have been subsequently confronted with LP?

As mentioned, despite this proof, Birnbaum himself ‘abandoned’ LP in favor of usefulness of error probability (sampling theories). ‘The likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretations’ (Birnbaum 1969: 128).

What about other frequentists? Most of their arguments about LP have cast doubt on the validity of Birnbaum’s proof and concluded we don’t have to accept LP as a necessary consequence of sound premises. It’s natural for frequentists to try to get rid of LP by way of finding defects in the proof, since the proof is claimed to be precisely based on the premises assumed by ‘frequentists’. Let’s see some of their criticism.

Durbin (1970) criticized the proof for the principle of sufficiency being equivocal and for the principle of conditionality being misguided. According to the principle of sufficiency, he says, evidential meaning should depend only on the ‘minimal’ sufficient statistic, which requires that in the mixture experiment above, the index h of E_h ($h = 1, 2$) should be retained in evidential inference. Then this also requires a modification to the principle of conditionality: if $\{E_h\}$ are components of an mixture experiment E^* , ‘ h depends only on the value of the minimal sufficient statistic’ (396). But in Birnbaum’s experimental setting, the conditional probability that $h = 1$, $x = x_1$ given that $h = 1$, $x = x_1$ or $h = 2$, $x = x_2$ is given by $m_1c/(m_1c + m_2)$, provided $P(x_1|\theta) = cP(x_2|\theta)$ for all θ , where m_h is the density of the distribution of the mixture E^* of h . This means that ‘the actual value of h cannot be part of the minimal sufficient statistic’ (397), which renders the proof invalid ([2] cannot be derived in the setting, given the modified principle of conditionality).

We find Mayo (2014) in line with this argument of Durbin's in claiming that the unconditional and conditional assessments are not evidentially equivalent. But Mayo argues that even if we accept two Birnbaum's principles without any revisions, his argument 'in any of its forms, necessarily applies these principles in a self-contradictory manner' (2014, manuscript: 3). So violation of LP doesn't mean violation either of the two principles. Kalbfleisch (1975), on the other hand, pay much more attention to the order of application of principles, though he also requires revision of them in terms of minimal sufficient statistic as well. He made a distinction between experimental conditionality and mathematical one: the former (C_E) implies one should condition on the particular experiment physically performed; the latter (C_M) is Birnbaum's original one including assumption that E is mathematically equivalent to a mixture E^* . Kalbfleisch contends that in evidential inference we should start from C_E , which leads to reduction of reference set from E^* to the actually performed experiment, E_h (a minimal experiment). The principle of sufficiency should be introduced only after that with a necessary revision which was made by Durbin (S'). The point is that LP does not follow from C_E and S' .

Thus, frequentists have called for some appropriate revisions of the two principles or the proof's logic, or else they just 'ignored conditionality' (Evans 2014). But Birnbaum's proof has not been invalidated (or not replaced with other candidates of principle) completely. For example, Evans indicates that what Birnbaum originally pursued in his proof was the measurement of statistical *evidence*, so that it's not enough to seek a strong foundation for *principles* but rather we need a *theory* of statistical inference. We will not explore here anymore how promising these frequentists' refutations of Birnbaum's proof are, but it's enough for us now to know the type of argument they adduce against LP. After all, frequentists do not denounce LP plainly but avoid it in favor of the usefulness of error probability³, which is almost in the same vein as Birnbaum's argument.

3. LP and likelihoodism

We now turn to how LP has been endorsed, particularly by likelihoodists, which is in sharp contrast with frequentists'.

3.1. Endorsement of LP by likelihoodists

Historically speaking, Bayesians were without doubt its advocates, but they put as much emphasis on priors as on LP (even Savage (1962), who was an ardent advocate of LP, considered it to provide a good motivation for personalistic statistics).

---

³ As Mayo (2011) summarizes citing Cox and Hinkley, 'Frequentists have long responded that having a good chance of getting close to the truth and avoiding error is what matters.'

So in this paper we focus on likelihoodists, who seem to be rooted in nothing other than LP.

As the designation shows, it seems a matter of course for likelihoodists to uphold LP as their main principle. But why do they think LP can be the main principle of statistical analyses? I think there are three lines of argument for this. One reason for their endorsement of LP is the fact that Fisher, one of the founders of LP, started his statistical arguments from LP (or likelihood as the essential part of statistics) and seemed to have returned to this in his later years making a detour through construction of p -value statistics. For example, Edwards (1992) refers to Fisher (1956) which made emphasis on the importance of likelihood, saying ‘Apart from the simple test of significance, ... there are to be recognized and distinguished ... two well-defined levels of logical status for parameters ... that in which the probability is known for the parameter ... and that in which ... the Mathematical Likelihood of all possible values can be determined from the body of observations available’ (27), ‘The likelihood supplies a natural order of preference among the possibilities under consideration’ (*ibid.*)⁴. Edwards also refers to Birnbaum’s deductive proof of LP from two more basic principles, which he thinks would ‘resolve the issue of the most convincing axiomatic basis for likelihood’ with further discussion. We find in Royall (1997) as well such a reliance on antecedent arguments supportive of LP by frequentists. Thus, we can derive part of the reasons for likelihoodists to endorse LP from traditional arguments which was once provoked by but not succeeded to favorably by frequentists.

The second type of argument for LP in likelihoodists has recourse to a counterintuitive consequence implied by violation of LP. As is well known, p -values (for example in the case of investigation of new drug) can be easily affected by the choice of stopping rule, while the likelihood ratio for two parameters retain the same value irrespective of it (not only likelihoodists but Bayesians like Savage (1961), who advocate LP, make full use of this type of argument as well).

The third argument for LP offered by likelihoodists is fulfillment of basic requirements for statistical inference. Edwards shows a list of six requirements for an appropriate measure of support in statistical analyses: transitivity, additivity, invariance under transformation of data or of parameter, relevance and consistency, and compatibility. ‘Relevance and consistency’ means that a measure of support must be intuitively acceptable and consistent in different applications (it must not be affected by intuitively irrelevant information). ‘Compatibility’ says it’s convenient if a measure of support has a simple link to Bayesian statistics when a valid prior exists. And LP is claimed to meet all the requirements. These requirements presuppose a relative

---

⁴ Fisher (1930) also says, ‘Likelihood serves all the purposes necessary for the problem of statistical estimation.’

support for a hypothesis and include some requirements of logical and mathematical consistency, which seem to be good criteria for the choice of statistical measure. However, even if these offer a sufficient condition for the choice of likelihood (and a ratio of likelihood) as a measure, they are not a necessary condition for the choice of an appropriate measure of support. The relative distance to the truth may well constitute one of the requirements, and so does a control of erroneous support of a hypothesis. And these requirements may very well result in a different choice of measure. Obviously, Edwards' list of requirements is not meant to be the one minimally required but one which is just applicable or intuitively correct.

Sober (2008)'s 'modest principle' might be taken as another example of this third argument for LP. It says, 'If learning that E is true justifies you in accepting the proposition P, and you were not justified in accepting P before you gained this information, then E must be evidence for P' (He also adds to this a negative principle of evidence against P). Sober puts this principle as a basic premise for the choice of measure of statistics. Note that, similarly to Edwards', this does not state necessary and sufficient conditions for evidence but just a sufficient one (data which are actually obtained can be qualified as evidence). True, LP fulfills this principle if LP is held to be an evidential statement, but it is still open to debate what evidence means in statistics and how statistical inference should be made even when the main concern of it is evidential (and as we will see, Sober holds LP as something different from LL, the latter of which he seems to relate as the main principle to likelihoodism).

So none of the reasons above succeed in providing a definitive answer comparable to Birnbaum's proof to the crucial question whether LP constitutes an indispensable part of statistical inference or not (as to Birnbaum's proof, they just refer to it). Rather likelihoodists seem to just show their preference to LP either based on their intuition or on relative advantages over frequentism in some respects.

Then it appears that any argument on LP, whether by advocates or by opponents, is no more than a demonstration of their preferences and that the arguments might result in sheer parallelism. And likelihoodists' attitude toward LP may seem even less productive when we see most of their arguments emphasize just the positive aspects of LP (they seem to focus almost exclusively on how they can evaluate the strength of evidential support for a hypothesis by setting a threshold of likelihood ratio, or on how they can replace safely *p*-value-based analyses with LP based ones), while frequentists have made efforts to seek a way out of dilemma they face between LP and their statistical standard.

In order to clarify a real issue of statistics in LP, I think we should examine first what is left almost unexplored by likelihoodists: rates of error, that is, how often we obtain erroneous conclusions when we follow the rule of statistical inference. As told above, frequentists give priority to error probability over LP. On the other hand, likelihoodists have not compared explicitly LP with control of error probability as a

basic principle, and just found fault with some consequences derived from frequentists' methods such as Neyman-Pearson (NP)'s hypothesis testing. That is to say, whether or not statistics based on LP (and LL) is undermined when error rates or error probabilities are dismissed is almost off the table of likelihoodists'. Can this really be negligible?

As an exceptional argument, Royall (1997) introduced 'misleading probability' into likelihoodism. With this probability, error rate is taken into consideration and even some parallelism can be seen between this and NP's error probability, though Royall stresses that there is a clear distinction between the two and that the misleading probability can be adopted as a suitable supplement to LP without undermining the principle's consistency. This argument provides a good foothold for our examination.

3.2. How error probability is tied to LP?

Why likelihoodists need error rates for the likelihood ratio might not be so clear. In likelihoodism, evaluation of a hypothesis is given primarily by the degree of how data actually obtained favor (support) the hypothesis, not by the degree of confirmation or by error rates. This comes from LL (law of likelihood), which is, as mentioned above, unanimously adopted as an essential law by likelihoodists and is usually considered to be in natural combination with LP (and not to mention, a likelihood of a hypothesis, when taken just as it is, cannot derive any meaningful statistical inference). As is palpable by the form of likelihood ratio, error probability cannot be tied directly to the evaluation of a hypothesis. What the ratio can indicate is simply which hypothesis is relatively favored by the data and to what degree the data favor one hypothesis (this is indicated by the value of likelihood ratio).

It is quite natural for likelihoodists to set some threshold to make a distinction between a strong and a weak favor: one of the typical threshold is 8 (and its counterpart, $1/8$). If the ratio is greater than 8 (less than $1/8$), one hypothesis is strongly favored by the data⁵, and otherwise ($1/8 < LR < 8$) weakly favored (Royall calls the data in the former case 'strong evidence', and the latter 'weak evidence'). Not only do we find positive sense in this evaluation, but also negative sense can be attached to the weak evidence, in that any interesting scientific conclusion cannot be drawn from it. The evaluation seems complete with this distinction. Then does it have any

---

⁵ Whether this threshold is generally applicable or not can be a problem. Royall shows why 8 is suitable only through a simple example of picking out balls from an urn. The more moderate view is that the threshold value can be different depending on the situation the likelihood ratio is applied to. More serious might be the problem raised by Hacking (1972). Hacking questioned whether the same likelihood ratio has the same degree of favor in all cases, which seems to be held uncritically in likelihoodism. Though it's worth arguing, I will focus for now on the relational issue between LP and error rates.

need to be made coupled with error probability?

It seems that the greater the value of the ratio, the more convincingly one hypothesis is favored by the data. However, whether the favor is genuine or not cannot be judged by the value of the ratio. Although it is known as the sample size increases, the log-likelihood ratio converges to the difference between the Kullback-Leibler divergences (the difference between the average log-likelihood of the true distribution and that of predictive distribution) of the two hypotheses compared (Taper and Lele 2011), no matter how large it is, the favor can always be erroneous. In what sense the favor can be erroneous is the point of issue.

Royall (1997) suggests what should be considered as erroneous is the probability that strong evidence is obtained even though the hypothesis which is not favored by the data is the true one. This probability can be calculated in some cases (when the data are considered to be normally distributed, for example), and it can be very easily shown that this probability is always less than a certain threshold as follows⁶. Suppose we obtain a likelihood ratio for two hypotheses, $H_1: \theta_1$ and $H_2: \theta_2$, and H_1 is favored by data x by k ($k > 1$). Then the probability that H_2 is a true hypothesis though the opposite one is favored by k is:

$$P(L(x, \theta_1)/L(x, \theta_2) \geq k) \leq 1/k.$$

This inequality⁷ is easily derived: the probability that such misleading data are obtained is simply given by the true hypothesis, that is, $P(x|\theta_2)$. From the assumption, $P(x|\theta_2) \leq P(x|\theta_1)/k$, and from the axiom of probability, $P(x|\theta_1) \leq 1$. Then $P(L(x, \theta_1)/L(x, \theta_2) \geq k) = P(x|\theta_2) \leq P(x|\theta_1)/k \leq 1/k$.

Royall calls this probability ‘misleading probability’. That this probability has the upper threshold inversely proportional to the degree of favor means that the higher the degree of favor is, the lower misleading probability we have. Surely this is a cogent way to introduce error rates to likelihoodism, and it has indeed a practical meaning. But we should note that this probability is literally a rate (frequency) of erroneous cases: other possible data than the ones actually obtained have to be assumed so as to calculate the upper threshold. Then it appears opposed to LP, which stipulates that statistical inference is totally irrelevant to data which are possible but not actually obtained.

This apparent discrepancy becomes larger when Royall tries to make likelihoodism more fully equipped with two probabilities, which can be compared to error probability in Neyman-Pearson’s hypothesis testing. As mentioned, the value of misleading probability (M) is calculable, and so is the probability that we obtain weak

---

⁶ Cox (2004) points out the first to give inequality as to error rates to likelihood ratio was probably Barnard.

⁷ The inequality was first introduced in Robbins (1970).

evidence (W). Royall considers M is comparable (analogous) to α (probability of erroneously rejecting the null hypothesis in NP testing) in that the true hypothesis fails to be favored in the misleading case ('strong evidence against H_1 is observed'). And he considers $M + W$ is comparable to β (probability of erroneously rejecting the alternative hypothesis in NP testing) in that the probability is that of failing to reject the null (H_1) in favor of the alternative (H_2) when the alternative is true. The aim of this analogy is to show that even though we can see some correspondence between the two negative kinds of probability, the ones of misleading and weak evidence are superior to those of NP testing since the former never have what likelihoodists (and Bayesians) think a fateful flaw in the latter (for example, a paradoxical acceptance of the alternative hypothesis with high power, which can happen when the sample size is very large. Howson & Urbach 2006). But what seems to be a serious problem to likelihoodists about M and W is that these probabilities are calculable only when the sample space is given (for example, when the model is a normal distribution, both M and W are given by the cumulative distribution function of standard normal distribution, that is, integration of probability distribution function with respect to x is needed).

Royall has a very simple but ingenious answer to this problem: M and W are taken into consideration *before* data are obtained, while evidential inference is made exclusively by a likelihood ratio *after* data are obtained. In other words, M and W can be separated from LP and their role is confined to a referential one in the experimental design. Then M and W are *not* incompatible with LP, but rather they are considered to comprise complementarily (as related to pre and post data) the whole statistical analysis. This answer may be a good solution to the problem Royall is faced with (the apparent mismatch of M and W with LP). But is this strategy really a solution to the original problem that LP fails to be coupled with error rates for its validation?

I think the answer is no. The point is what the error rates (probabilities) mean when LP is the principle of statistics. What Royall tries to control, particularly in a misleading case, is the probability that the unfavored hypothesis is *true*. Indeed, it can be counted as error when the true hypothesis is not favored. But it's hardly the case that one of the hypotheses in LR is true: in most cases both hypotheses treated in LR are, to be precise, false. The case in which the true hypothesis is unfavored (with a very high ratio) should be taken for *the worst case scenario* in that the ultimate goal of the evidential inference is to pin down the true hypothesis when it is within reach. But how should we treat errors when the two hypotheses are both false? Although whichever hypothesis is favored, we can say the result is erroneous in a strict sense, it still makes sense to pursue errors since it is natural for us to avoid the case in which a hypothesis very far apart from the truth is favored when the other hypothesis is fairly close to the truth. Then it seems what we need is a measure

regarding the relative distance of each hypothesis to the truth. This measure is not precisely an error probability, but something which can effectively substitute for it in the sense that an undesirably favored hypothesis in LR could be avoided with high probability.

Before moving on, we should note one thing. Though likelihoodists cannot simply couple LP with NP's error control without compromising their principle (which is shown in Sober (2008)), there is another important reason that they should not join NP when dealing with errors. The problem is, NP's error probability is also calculable on the assumption that a hypothesis is true⁸. What is imposed upon likelihoodists for their advantage over traditional frequentism is to go beyond such an unrealistic assumption. In this respect as well, the most plausible measurement needed is about a relative distance to truth.

Our question is whether this relative distance is attainable without recourse to sample space or not. Or if sample space is unavoidable, what we want to see is whether or not the distance is attainable at least without compromising LP. Of course, the truth is usually unknown. But there are some statistical measures to estimate the distance to the truth. Let us consider two cases in which such measures can possibly be combined with likelihoodism, and then let's see what will become of LP.

4. LP and a relative distance to the truth as a substitute for error probability

A well-known measure of the distance to the truth is Kullback-Leibler divergence. But this measure in itself is not applicable directly to practical issues, since it calculates the divergence assuming the truth (true value of parameter).

Akaike's Information Criterion (AIC) is one of the practical tools to measure a comparative distance to the truth between models, based on Kullback-Leibler divergence. First, we see whether AIC can be a good measure of the distance for likelihoodists, through examining Forster and Sober's approach. And next, we see another practical approach, bootstrap method, advocated by evidentialists and how efficacious it is for the current problem.

4.1. LP and AIC

First candidate of the measure is Akaike's Information Criterion, AIC. AIC is a

---

⁸ One more reason that likelihoodists refuses to be combined with NP in error control is that NP's error probability does not indicate uncertainty of the hypothesis itself but an expected error rate when the same method is applied to the test. Royall's misleading probability evades this problem since M is determined by sample size and the values contained in the two hypotheses compared (for example, $M(k) = \Phi(-\frac{\sqrt{n}\delta}{2\sigma} - \frac{\sigma \log k}{\sqrt{n}\delta})$, when the statistical model is $N(\theta, \sigma^2)$, and two hypotheses are $H_1: \theta = \theta_1$, $H_2: \theta = \theta_1 + \delta (\delta > 0)$, with $LR = k(k \geq 8)$).

measure of the distance of a predictive distribution from the true one, which is calculated by penalized maximum likelihood estimates (Akaike 1973. $-2 \sum_{i=1}^n \log p(x_i|\hat{\theta}) + 2k$, where $\hat{\theta}$ is the maximum likelihood estimate, and k is the number of parameters in the model). Interestingly, Sober (2008) adopts a strategy of combining LL (not exactly LP) and AIC, but for another reason. His strategy aims to complement likelihood ratio, which cannot essentially deal with composite statistical hypotheses, by way of AIC, that deals with them as ‘models’ and provides a criterion of each model’s predictive accuracy. Indeed, if the target of statistical analyses is aimed at both simple and composite hypotheses (though this assumption is different from other likelihoodists’), this combination seems to cover all the targets. Furthermore, as AIC is based on the maximum likelihood estimate for each model, it seems to be little problem to combine this with likelihoodism⁹. But does this combination overcome Royall’s problem at the same time?

We cannot answer it positively for two reasons. First, as AIC is confined to the comparison of composite hypotheses, LR for simple hypotheses are left untouched, and what is worse than Royall’s, no care is taken for errors there in Sober’s. It’s possible to apply Royall’s M (and W) to Sober’s likelihoodism, but the error control in LR still remains to be solved. Then there is no counterbalancing merit of incorporating AIC.

Secondly, there is a problem about compatibility with LP. AIC is composed of log-likelihood and the penalty term k , which is the number of parameters in the model. Seen on the surface, there is nothing to contradict LP in this composition, since k seems just the number proper to each model, determined independently of data. Then some say AIC is an approach that *implement* LP (Robert 2007). But it is simply misreading. ‘ k ’ is a penalty term for doubly using the same data to estimate the true parameter and to estimate average log-likelihood. k happens to almost coincide with the number of parameters in the model, but to be precise, the term is a bias correction (average difference between average log-likelihood and log-likelihood) given by $E[I(\hat{q}, p(\hat{\theta})) - E\{I(q, p(\hat{\theta}))\}]$ (where p : a predictive distribution, q : the true distribution, $\hat{q}$: an empirical distribution, $\hat{\theta}$: the maximum likelihood estimate, and I : Kullback-Leibler divergence) (vid. Konishi & Kitagawa 2004), and k is just an approximation to it (that is, k is used just for a practical reason). As shown in the mathematical form, the bias correction derives from *averaging over the sample space* (Boik (2004) points it out but with a different view from above). And AIC score is meaningful for evidential inference only when taken as a whole (k cannot be detached from it, and separation of pre and post data analysis as in Royall is impossible). Then we have to say Sober’s strategy with AIC goes against LP.

However, we should note that this does not directly go *against likelihoodism*.

---

⁹ Indeed, the maximum likelihood estimate accords well with LP.

AIC still has a chance to accord with the other cornerstone of likelihoodism, LL. LP is often taken as accompanied by LL, but as I mentioned in section 1, we can make a distinction between these two principles. LP says just the likelihood is the minimum sufficient statistic with no mention of how to use it in statistical (evidential) inferences. LL, on the other hand, just defines how to use likelihood in evidential inference without mentioning what the minimum sufficient statistic is. Grossman (2011) emphasizes the importance to distinguish between LP and LL, but his point is that LL entails LP; in other words, LP is much weaker than LL, which I think is misleading. Though the condition ‘if and only if $L(x, \theta_1) > L(x, \theta_2)$ ’ in LL may seem to imply that likelihood is a sufficient statistic, what it exactly refers to is how evidential (relative) support is realized, and does not refer to how evidential inference is *completed* with likelihood. On the other hand, LP proclaims that complete evidential inference is to be made with likelihood alone (since likelihood is the minimum sufficient statistic) without specifying how. Then it is possible to say LP is much stronger than LL as a principle, and that, unlike LP, sample space is not plainly ruled out of statistical inference by LL when we treat it in isolation¹⁰.

But when it comes to AIC, how can LL be weakened so as to subsume the averaging over the sample space? To this question, Forster & Sober (2011) have a very skillful answer as follows. If model M_2 is nested in M_1 , and data y ($y > 0$) is observed, for all z ($z \neq x$), there always exists a positive number x , which satisfies the inequality below (PA is a predictive accuracy).

$$\begin{aligned} &P\{AIC(M_2) - AIC(M_1) = y | PA(M_2) - PA(M_1) = x\} \\ &> P\{AIC(M_2) - AIC(M_1) = y | PA(M_2) - PA(M_1) = z\}. \end{aligned} \quad (1)$$

The inequality has the same mathematical structure as $L(x, \theta_1) > L(x, \theta_2)$ in LL, meaning that a *hypothesis* that $PA(M_2) - PA(M_1) = x$ ($x > 0$) is favored than any other hypothesis by the *data* that $AIC(M_2) - AIC(M_1) = y$ ($y > 0$). This shows predictive accuracy of AIC is validated by way of LL, that is, AIC is claimed to be compatible and methodologically linked with LL. Thus, they manage to circumvent a cumbersome problem of sample space contained in AIC by way of reversing

---

¹⁰ There are several different interpretations of the relationship between LP and LL. Joshi (1983) and Edwards (1992) argued we need no distinction between them (Edwards combines them into one axiom), while Hacking (1965), who first introduced formally LL, argued LP is too strong in that LP has no mention of the statistical model (and also criticize its insufficiency). Even Hacking (though he later abandoned likelihoodism) may not go so far as to permit sample space in evidential inference explicitly, but if it makes sense to distinguish LP and LL, and if we take LL as a weaker principle than LP (which is possible), then it’s not preposterous to try to weaken it to the extent that sample space is subsumed into a generalized LL. Interestingly, whether it’s intentional or not, Royall (2000) removes ‘only if’ condition from LL, which makes this interpretation much more likely.

the relationship between data and hypotheses. Though their proof is still tentative and holds under limited conditions, it opens the possibility of making LL weak (or general) enough to subsume AIC¹¹, and consequently makes it possible to combine likelihoodism with AIC. (Once again, their proof might be aimed at a different goal, but it's wholly effective for our current problem.)

That being said, the fact remains that LP must be abandoned in this strategy. Having LL alone as a principle, and also having the weaker LL means having likelihoodism inevitably weaker. In other words, saving likelihoodism from the problem of error rates by way of AIC is only possible at the cost of strong likelihoodism; above all, LP.

4.2. Evidentialists' approach

Another candidate to compensate for the insufficient error control of likelihoodism is evidentialists's approach. Let's see briefly how it works for our problem. Evidentialism is literally an evidence-centered statistics, which inherits likelihoodism in that it is based on the likelihood ratio of two simple hypotheses in inference, but at the same time it aims to overcome defects of likelihoodism.

Taper & Lele (2011), one of the leading groups of evidentialism, attempts to overcome Royall's problem by applying bootstrap method to likelihood ratio. The bootstrap is a statistical method introduced by Efron (1979), in which inference is made about an estimate of parameter θ by resampling data repeatedly with the same sample size n with replacement from the existing data set. We can evaluate the value of θ for each sample, and then we obtain a sampling distribution, from which we can obtain a confidence interval for θ , for example (the sampling distribution converges to a normal distribution based on central limit theorem, and the mean of it converges to the true value of θ as the number of sampling tends to ∞). With this method, Taper & Lele say, we can obtain a confidence interval for the likelihood ratio (by which the upper or lower limit of the plausible divergence is shown), or we can calculate the proportion of times that one hypothesis will be favored over the other¹².

What they emphasize about this method is that unlike Royall's M, the assessment is made 'under truth, and not under either model' (not counterfactually). And indeed, not just theoretically but also practically with the aid of computer, the estimated accuracy can be improved as much as we like since it is assured we approach the truth (the true likelihood ratio, that is, the difference between Kullback-Leibler

---

¹¹ Boik (2004) denies AIC is compatible with LL since AIC is about comparison of models while LL concerns that of specified parameters. Sober & Forster's proof also circumvents Boik's problem.

¹² The bootstrap method is applied to information criteria as well: EIC (Ishiguro, Sakamoto, and Kitagawa 1997). But this does not provide a solution to Sober's problem since it still stays at 'model' selection.

divergences). Moreover, unlike Sober, Taper & Lele are consistent in the target: what is pursued is constantly the likelihood ratio of two simple hypotheses.

Does this mean the problem of error rates in likelihoodism is overcome with this as a solution? And is this solution compatible with LP? The answer to the former question depends on what we desire in evidential inference. If we are content with obtaining *some* reliable information based on the truth, the answer is yes. But if we want to obtain the *precise* probability of having erroneous likelihood ratio for the given data, the answer may be no. That's why Taper & Lele say their method is just a practical one.

Then how about the second question? Taper & Lele refer to LP in their paper in relation to Royall's likelihoodism, but do not make explicit its relation to their method. However, the answer must be no, since even if we resample data set from the given data set over and over again with computer, we always have to depend on the resampled data set which are possible but not actually obtained (note that as long as we resample some data set from the existing data set, each of the *data* is admittedly already given, but what *data set* and how many of them will be resampled is not fixed).

There is a chance that likelihoodism will be rescued from the error problem more safely with this bootstrap method than with AIC¹³, but again, LP cannot be sustained in a strict sense. This explains in part why they call this approach afresh, evidentialism.

Conclusion

The likelihood principle has been discussed traditionally just like the issue of preferences and remain ambiguous in the evaluation of its validity, but if we focus on the problem of error rates of likelihood ratio in likelihoodism, the deficiency of the principle becomes clear and it seems difficult to make up for the deficiency even with the most plausible statistical measures of the relative distance to the truth, AIC or the bootstrap method. It does not mean that this is fatal to likelihoodism. Rather, the traditional likelihoodism is still useful in some practical applications (perhaps in the clinical trial, for example, where likelihoodism seems to become now more popular than frequentists' tests). However, in order to make inference rigorously in science based on likelihood ratio, likelihoodism would be urged to be replaced with a more generalized one in which the likelihood principle is invalidated.

---

¹³ Sober (2015) seems to depend on multiple data to obtain a precise likelihood ratio rather than on the combination with AIC. I think this would be more reasonable in terms of retention of likelihoodism.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP16H03050 and JP20H01736. I would like to thank anonymous referees for their insightful comments. I am also grateful to Elliot Sober for his valuable comments.

References

- Akaike, H. (1973). Information Theory as an Extension of the Maximum Likelihood Principle, Petrov, B. and Csaki, F. (eds.), *Second International Symposium on Information Theory*, pp. 267–281.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag.
- Birnbaum, A. (1962). On the Foundations of Statistical Inference, *Journal of the American Statistical Association*, Vol. 57, No. 298, pp. 269–306.
- Birnbaum, A. (1969). Concepts of Statistical Evidence, Morgenbesser, S. et al. (eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, St. Martin's Press, pp. 112–143.
- Boik, R. J. (2004). Commentary on ‘Why Likelihood?’ (Forster, M. and Sober, E.), Taper, M. L. and Lele, S. R. (eds.), *The Nature of Scientific Evidence*, The University of Chicago Press, pp. 167–180.
- Cox, D. R. (2004). Commentary on ‘The Likelihood Paradigm for Statistical Evidence’ (Royall, R.), Taper, M. L. and Lele, S. R. (eds.), *The Nature of Scientific Evidence*, The University of Chicago Press, pp. 138–140.
- Durbin, J. (1970). On Birnbaum's Theorem on the Relation between Sufficiency, Conditionality and the Likelihood, *Journal of the American Statistical Association*, Vol. 65, No. 329, pp. 395–398.
- Edwards, A. W. F. (1992). *Likelihood*, The Johns Hopkins University Press.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, Vol. 7, No. 1, pp. 1–26.
- Evans, M. (2014). Discussion of “On the Birnbaum Argument for the Strong Likelihood Principle,” *Statistical Science*, Vol. 29, No. 2, pp. 242–246.
- Fisher, (1921). On the ‘Probable Error’ of a Coefficient of Correlation Deduced from a Small Sample, *Metron*, Vol. 1, pp. 3–32.
- Fisher, R. A. (1930). Inverse Probability, *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 26, Issue 4, pp. 528–535.
- Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*, Oliver and Boyd.
- Forster, M. & Sober, E. (2011). AIC Scores as Evidence – A Bayesian Interpretation, Bandyopadhyay, P. S. and Forster, M. R. (eds.), *Philosophy of Statistics (Handbook of the Philosophy of Science 7)*, Elsevier, pp. 535–549.
- Grossman, J. (2011). The Likelihood Principle, Bandyopadhyay, P. S. and Forster, M. R. (eds.), *Philosophy of Statistics (Handbook of the Philosophy of Science 7)*, Elsevier, pp. 553–580.
- Hacking, I. (1965). *Logic of Statistical Inference*, Cambridge University Press.
- Hacking, I. (1972). Likelihood, *The British Journal for the Philosophy of Science*, Vol. 23,

No. 2, pp. 132–137.

- Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach* (3rd ed.), Open Court.
- Ishiguro, M., Sakamoto, Y., and Kitagawa, G., (1997). Bootstrapping Log Likelihood and EIC, An Extension of AIC, *Annals of the Institute of Statistical Mathematics*, Vol. 49, pp. 411–434.
- Joshi, V. M. (1983). Likelihood Principle, Kotz, S., Johnson, N. L., and Read, C. B. (eds.), *Encyclopedia of Statistical Sciences*, Vol. 9, Wiley.
- Kalbfleisch, J. D. (1975). Sufficiency and Conditionality, *Biometrika*, Vol. 62, No. 2, pp. 251–259.
- Konishi, S. & Kitagawa, G. (2004). *Information Criteria*, Asakura Shuppan (available only in Japanese).
- Mayo, D. G. and Spanos, A. (2011). Error Statistics, Bandyopadhyay, P. S. and Forster, M. R. (eds.), *Philosophy of Statistics (Handbook of the Philosophy of Science 7)*, Elsevier, pp. 153–198.
- Mayo, D. G. (2014). On the Birnbaum Argument for the Strong Likelihood Principle, *Statistical Science*, Vol. 29, No. 2, pp. 227–239. (Also, the manuscript is available on the Web. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.362.9758&rep=rep1&type=pdf>)
- Robbins, H. (1970). Statistical Methods Related to the Law of the Iterated Logarithm, *The Annals of Mathematical Statistics*, Vol. 41, No. 5, pp. 1397–1409.
- Robert, C. P. (2007). *The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation*, Springer.
- Royall, R. (1997). *Statistical Evidence: A likelihood paradigm*, Chapman & Hall/CRC.
- Royall, R. (2000). On the Probability of Observing Misleading Statistical Evidence, *Journal of the American Statistical Association*, Vol. 95, No. 451, pp. 760–768.
- Savage, L. J., et al. (1962). On the Foundation of Statistical Inference: Discussion, *Journal of the American Statistical Association*, Vol. 57, No. 298, pp. 307–326.
- Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*, Cambridge University Press.
- Sober, E. (2015). *Ockham's Razors: A User's Manual*, Cambridge University Press.
- Taper, M. L. and Lele, S. R. (2011). Evidence, Evidence Functions, and Error Probabilities, *Philosophy of Statistics (Handbook of the Philosophy of Science 7)*, Elsevier, pp. 513–534.

(Received 2021.1.20; Revised 2021.6.30; Accepted 2021.7.10)