

Likelihood Principles, Sequential Sampling, Stopping Rules, and Jeffreys's Rule

John P. Burkett
burkett@uri.edu

November 4, 2009

Outline

Likelihood Principles

- Bayesian likelihood principle

- Generalized likelihood principles

Sequential sampling and stopping rules

Jeffreys's rule revisited

Likelihood principle

Bayesian likelihood principle

In Bayesian statistics, as we have seen in earlier classes, data influence inferences via a likelihood fn. The posterior density of a parameter θ is proportional to the product of a likelihood fn and a prior density: $p(\theta|x) \propto L(\theta; x)p(\theta)$, where x denotes data. An immediate implication is that two data sets x and y that produce proportional likelihood fns [$L(\theta; x) \propto cL(\theta; y)$, where the proportionality constant is independent of θ] lead to the same posterior density if the same prior density is used in both cases. (Every posterior density must integrate to 1; thus proportional posterior densities are identical.) Two data sets that lead to the same posterior distribution for θ lead Bayesians to the same conclusions about θ . This implication is the gist of a proposition that Hill (1987, p. 98) calls “the Bayesian likelihood principle.”

Bayesian likelihood principle I

A more formal version the Bayesian likelihood principle can be expressed in terms of experiments. Writing \tilde{x} for a random variable to be observed during an experiment, x for a particular value of this random variable, θ for a parameter, and $p(x|\theta)$ for the data density, we can define an experiment as $E := \{\tilde{x}, \theta, p(x|\theta)\}$. Let E_1 and E_2 denote two possible experiments. Assume that the process (random or otherwise) that determines whether E_1 or E_2 is performed reveals nothing about θ . Now suppose that the outcomes x_1 and x_2 of experiments E_1 and E_2 are such that their likelihood fns are proportional—that is to say, $L(\theta; x_1) = cL(\theta; x_2)$, where c is a positive constant independent of θ . Then according to the Bayesian likelihood principle, any inference or decision concerned exclusively with θ that could be based on observing x_1 in E_1 could equally well be based on observing x_2 in E_2 . Hill (1987, p. 98) spells out the application to inference as follows:

Bayesian likelihood principle II

Let P be any proposition concerning the value of θ and nothing else—that is, that θ lies in some specified set. Then P should be regarded as equally valid whether x_1 is observed in E_1 or x_2 is observed in E_2 .

He specifies the application to decision in the following words:

In any decision problem where the loss function depends only on θ and the act taken, the same postdata preference for acts should obtain whether x_1 is observed in E_1 or x_2 in E_2 .

Hill's version of the Bayesian likelihood principle follows immediately from the fact that two experiments and outcomes that produce proportional likelihoods lead to the same posterior distribution $p(\theta|x_i)$, $i = 1, 2$. This posterior distribution fully determines Bayesian inferences about the probability of a proposition exclusively concerned with θ . That posterior

Bayesian likelihood principle III

distribution and a given utility fn $U(\theta, a)$ [or equivalently a loss fn $-U(\theta, a)$] also determine the expected utility

$$EU := \int U(\theta, a)p(\theta|x_i)d\theta$$

or expected loss $-EU$ associated with an act a . If E_1 and x_1 lead to the same posterior distribution for θ as do E_2 and x_2 , then an act that maximizes EU for E_1 and x_1 will also maximize it for E_2 and x_2 . Thus Bayesian decisions will be the same in either case. The Bayesian likelihood principle is sometimes called a *restricted* likelihood principle to contrast it to a more general one that we will consider next (Hill 1987).

The Bayesian likelihood principle is a tool for inference and decision within the framework of a given model. However, the correctness of the inference or decision may depend on the validity of the model. Further, no likelihood fn can express uncertainty

Bayesian likelihood principle IV

about the model on which it based (Pawitan 2001). For checking model assumptions, we may need to examine sampling distributions (Gelman et al. 2004).

Generalized likelihood principle I

Some statisticians who dislike use of prior distributions are nonetheless disposed to let likelihood functions play a central role in data analysis. An eminent early example is Ronald Fisher, who in fact first gave “likelihood” a technical statistical meaning. Attempts to formally justify likelihood-based inference without using Bayes’s theorem include Barnard (1949) and Birnbaum (1962). The latter generated much interest and controversy by asserting that a likelihood principle (LP) more general than that embraced by Bayesians was equivalent to a combination of two other principles which he found attractive, namely a “conditionality principle” (CP) and a “sufficiency principle” (SP). His claim, symbolically expressed, was that $LP \Leftrightarrow \{CP, SP\}$.

Birnbaum expressed the three principles in terms of abstract concepts of evidence and evidential meaning. An experiment E and its outcome x were said to constitute “evidence” denoted (E, x) having “essential properties” or “evidential meaning” denoted

Generalized likelihood principle II

$Ev(E, x)$ (Birnbaum 1962, pp. 269–270). He did not specify any particular mathematical form or properties for $Ev()$.

More recent expositors of conditionality and sufficiency principles, including Berger and Wolpert (1988) and Lee (2004), generally use a modified terminology and notation. In Lee's version—pertaining to a discrete parameter θ and a discrete random variable \tilde{x} with pmf $p(x|\theta)$ —an experiment E is defined as $E := \{\tilde{x}, \theta, p(x|\theta)\}$. If the experiment E generates an outcome x it is said to provide “evidence” $Ev\{E, x, \theta\}$ regarding θ . Again the form of Ev is not specified. Ev may be interpreted “to be *anything*, any collection of conclusions or reports” based on E and x (Berger and Wolpert 1988, 197).

We might expect that no very strong principles could be laid down that would apply so widely. Nonetheless, Birnbaum's principles make rather strong claims as we will see.

Generalized likelihood principle III

Birnbaum's theory involves a **mixed experiment** (a.k.a. mixture experiment) E that is

equivalent to a mixture of several other component experiments E_h , in the sense that observing an outcome x of E is mathematically equivalent to observing first the value of h of [a] random variable having a known distribution (not depending upon unknown parameter values), and then taking an observation x_h from the component experiment E_h .

Generalized likelihood principle IV

Birnbaum (1962, p. 271) expressed his general conditionality principle as follows:

If E is any experiment having the form of a mixture of component experiments E_h , then for each outcome (E_h, x_h) of E we have $Ev(E, (E_h, x_h)) = Ev(E_h, x_h)$. That is, the evidential meaning of any outcome of any mixture experiment is the same as that of the ... outcome of the corresponding component experiment, ignoring the over-all structure of the mixture experiment.

A slightly less restrictive postulate, called the *weak* conditionality principle, is used by Berger and Wolpert (1988) and Lee (2004). It can be expressed as follows: Let $E_1 := \{\tilde{x}_1, \theta, p(x_1|\theta)\}$ and $E_2 := \{\tilde{x}_2, \theta, p(x_2|\theta)\}$ be two experiments having in common an unknown parameter θ . Let E^* be a mixed experiment in which we observe j to be either 1 or 2, both outcomes having probability 1/2 independent of θ , and then perform component experiment E_j . In

Generalized likelihood principle V

more formal terms, let $E^* := \{\tilde{x}^*, \theta, p(x^*|\theta)\}$, where $x^* = (j, x_j)$ and $p(x^*|\theta) = \frac{1}{2}p(x_j|\theta)$. Then $Ev\{E^*, x^*, \theta\} = Ev\{E_j, x_j, \theta\}$.

The sufficiency principle as formulated by Birnbaum (1962, p. 270) reads as follows:

If E is a specified experiment, with outcomes x ; if $t = t(x)$ is any sufficient statistic; and if E' is the experiment, derived from E , in which any outcome x of E is represented only by the corresponding value $t = t(x)$ of the sufficient statistic; then for each x , $Ev(E, x) = Ev(E', t)$, where $t = t(x)$.

Generalized likelihood principle VI

Another version of this postulate, called the *weak* sufficiency principle, is given by Lee (2004, p. 195) in the following terms:

Consider the experiment $E := \{\tilde{x}, \theta, p(x|\theta)\}$ and suppose that $t = t(x)$ is sufficient for θ given x . Then if $t(x_1) = t(x_2)$, $Ev\{E, x_1, \theta\} = Ev\{E, x_2, \theta\}$.

Whether a statistic is sufficient for a parameter depends on the model in which they are embedded. For example, the sample mean and sample variance are sufficient for the population mean and population variance in the normal model but not necessarily in some other models. Uncertainty about model specification may result in difficulty identifying sufficient statistics or in a need to supplement sufficient statistics with other information.

Generalized likelihood principle VII

Most data analysts perform some sort of “model checking” when analyzing a set of data. Most model checking is, necessarily, based on statistics other than a sufficient statistic. For example, it is common practice to examine residuals from a model, statistics that measure variation in the data not accounted for by the model. . . . Such a practice immediately violates the sufficiency principle, since the residuals are not based on sufficient statistics. (Casella and Berger 2002, p. 295).

A generalized likelihood principle can be expressed as follows: Let $E_1 := \{\tilde{x}_1, \theta, p(x_1|\theta)\}$ and $E_2 := \{\tilde{x}_2, \theta, p(x_2|\theta)\}$ be experiments with a common parameter θ . Let x_1 and x_2 be particular outcomes of the experiments such that $L(\theta; x_1) = cL(\theta; x_2)$, where c is a

Generalized likelihood principle VIII

positive constant of proportionality independent of θ . Then $E_V\{E_1, x_1, \theta\} = E_V\{E_2, x_2, \theta\}$.

Birnbaum and later expositors of ideas similar to his offer mathematical proofs that the CP and SP are jointly equivalent to the generalized LP. Birnbaum (1962) seemed to expect that most statisticians would find the CP and SP intuitively obvious and thus be persuaded to accept the not-so-obvious generalized LP. For those reluctant to accept SP, the route to LP was cleared by subsequent work showing that CP alone implies LP Evans et al. (1986).

Birnbaum's three principles were met with criticism or indifference from many sides. Statisticians working in the tradition of Jerzy Neyman and Egon Pearson continued using test procedures that clearly violate conditionality and likelihood principles. Statisticians in the Ronald Fisher tradition were more sympathetic to likelihood based inference. However, a prominent member of this group has

Generalized likelihood principle IX

expressed reservations about the conditionality principle, noting an example in which a mixed experiment seems to yield more evidence than the component experiment it selects (Pawitan 2001, p. 211). He also noted that two data sets may lead to the same likelihood fn without necessarily leading to the same conclusions. (Prior distributions or sampling schemes could differ.) Birnbaum (1970) himself eventually rejected likelihood concepts. Some Bayesians expressed concern that the generalized LP made exaggerated claims. Hill (1987, p. 95), for example, argues that Birnbaum and his followers defined Ev so broadly that their principles make claims “stronger than can be justified, even within the Bayesian framework.” Hill notes that the Bayesian likelihood principle is narrower and more defensible than the generalized likelihood principle in two respects. First, it applies to statements concerned solely with θ , not statements about θ and other matters. Second,

Generalized likelihood principle X

it includes the qualification that “the choice of experiment not be informative as to the parameter” (Hill 1987, p. 98).

Sequential sampling and stopping rules I

Sampling is said to be **sequential** if observations are made one item or one group at a time and the experimenter can stop the sampling after any item or group. In the remainder of this section, sequential sampling is generally assumed to occur item by item.

Sequential sampling produces a sample that may be denoted by $\mathbf{x}^{(m)} = (x_1, x_2, \dots, x_m)$, where subscripts indicate the order in which the observations are drawn.

A **stopping rule** for sequential sampling is a sequence of functions $\tau_m(\tilde{\mathbf{x}}^{(m)})$ such that if $\mathbf{x}^{(m)}$ is observed then sampling terminates with probability $\tau_m(\mathbf{x}^{(m)})$; otherwise sampling proceeds to observation x_{m+1} . A stopping rule is **proper** if it guarantees that the final sample size (a.k.a. stopping time) will be finite. A stopping rule is **informative** if the stopping time yields information about parameters of interest beyond that given by the observed $\mathbf{x}^{(m)}$. Otherwise a stopping rule is **noninformative** (Lee 2004).

Sequential sampling and stopping rules II

A stopping rule is informative for a Bayesian if and only if it affects either the prior distribution or the likelihood fn value through channels besides $\mathbf{x}^{(m)}$. To see how a stopping rule might affect our prior distribution, suppose we are preparing to analyze data that that will be supplied by an experimenter who is conducting Bernoulli trials. The experimenter has a choice of specifying the sample size and letting the number of success be random (as in a binomial distribution) or specifying the number of successes and letting the number of trials or failures needed to get those successes be random (as in a negative binomial distribution). If the experimenter chooses the latter experimental design, we might surmise that he has prior information suggesting that successes are infrequent. This surmise might incline us to choose an informative prior distribution that gives high probability to low success rates (Hill 1987). Even if we use a noninformative prior, we may adapt it to express ignorance about what in particular the chosen

Sequential sampling and stopping rules III

experiment might reveal. (This possibility is explored further in the following section on Jeffreys's rule.)

To see how stopping rules, broadly interpreted, might affect a likelihood fn value other than through the observed $\mathbf{x}^{(m)}$, suppose that an experimenter records successes and failures in Bernoulli trials with parameter θ . The trials occur at random times governed by a Poisson process, whose rate also depends on θ . The experimental design calls for observing the trials for a certain period of time. The fixed observation time together with the Poisson process determines the number of trials. If the experimenter wants us to estimate θ , she should report the stopping rule (length of observation time) as well as the trial outcomes, because both affect the likelihood fn and thus the posterior distribution of θ Berger (1985).

If an experiment is conducted using one of several possible noninformative stopping rules, Bayesians can calculate the

Sequential sampling and stopping rules IV

posterior distribution $p(\theta|\mathbf{x})$ and thus make inferences about θ without knowing which of those rules was used.

Jeffreys's rule revisited I

We examined Jeffreys's rule for deriving uninformative priors in our fifth class when discussing the binomial distribution. We revisit it now to consider whether it conforms to Bayesian principles. Doubts about that issue arise because Jeffreys's rule allows experimental design to influence reference priors and the corresponding posteriors. Experiments with a common likelihood fn but different stopping rules can lead an analyst using Jeffreys's rule to different conclusions, in violation of the generalized likelihood principle. In the conflict between Jeffreys's rule and the generalized likelihood principle, Lee (2004, pp. 202-03) sides with the latter. However, some Bayesians disagree.

To see why some, perhaps most, Bayesians accept Jeffreys's rule, recall that although all Bayesians accept Bayes's theorem and its implication that new data can influence the posterior distribution only through the likelihood fn, they do not all accept the generalized likelihood principle. Jeffreys's rule is consistent with

Jeffreys's rule revisited II

Bayes's theorem although it violates the generalized likelihood principle. In defense of Jeffreys's rule and related ideas such as priors based on data-translated likelihoods, Box and Tiao (1973, p. 44) argue that the aim in specifying a noninformative prior is to "represent not total ignorance but an amount of prior information which is small *relative* to what the particular projected experiment can be expected to provide." Hence the prior density may well depend on the experimental design.

A good example of how Jeffreys's rule can generate noninformative priors tailored to specific experimental designs is provided by comparing priors chosen by Jeffreys's rule in the cases of binomial and negative binomial (Pascal) distributions. The binomial distribution, as we know, involves a fixed number of trials with a random number of successes. In this case, Jeffreys's rule leads, as we saw in our fifth class, to the prior density

$p(\pi) \propto \pi^{-1/2}(1 - \pi)^{-1/2}$, where the unknown parameter π

Jeffreys's rule revisited III

represents the propensity of a trial to end in success—that is, the relative frequency of success in repeated trials. In contrast, a negative binomial distribution involves a fixed number of successes and random number of failures or total number of trials. In this case, Jeffreys's rule leads to a slightly different prior density: $p(\pi) \propto \pi^{-1}(1 - \pi)^{-1/2}$. To see how the different priors affect posteriors in a simple context, imagine an experiment with just one trial, which is a success. The posterior density in the binomial case is proportional to $\pi^{1/2}(1 - \pi)^{-1/2}$ and thus represents a $\text{Be}(1.5, 0.5)$ distribution. The posterior density in the negative binomial case is proportional to $\pi^0(1 - \pi)^{-1/2}$ and hence corresponds to a $\text{Be}(1, 0.5)$ distribution. Recalling that a variable with a $B(\alpha, \beta)$ distribution has a mean $\alpha/(\alpha + \beta)$, we see that the mean of the posterior distribution of π is $3/4$ in the binomial case and $2/3$ in the negative binomial case. The downward adjustment of the posterior mean from the first to second case makes intuitive sense:

Jeffreys's rule revisited IV

The last trial can be either a success or a failure in the binomial experiment but has a guaranteed outcome (in our case a guaranteed success) in the negative binomial experiment. It should not be surprising that a successful trial suggests a higher π in an experiment that could have ended without such a trial than one guaranteed to end with one. A disagreement between posteriors based on such different experiments is “is much less surprising than the claim that they ought to agree” (Box and Tiao 1973, p. 46).

Thus, despite concerns raised by proponents of the generalized likelihood principle, “most Bayesians” who need noninformative priors favor those based on Jeffreys's rule (Marin and Robert 2007, p. 24).

References I

- Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society*, 11(2):115–39.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition.
- Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*, volume 6 of *Lecture Notes-Monograph Series*. Institute of Mathematical Statistics, Hayward, California, second edition.
- Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.
- Birnbaum, A. (1970). Statistical methods in scientific inference. *Nature*.
- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

References II

- Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury, Pacific Grove, CA, second edition.
- Evans, M. J., Fraser, D. A. S., and Monette, G. (1986). On principles and arguments to likelihood. *The Canadian Journal of Statistics*, 14(3):181–99.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall, London, second edition.
- Hill, B. M. (1987). The validity of the likelihood principle. *The American Statistician*, 41(2):95–100.
- Lee, P. M. (2004). *Bayesian Statistics: An Introduction*. Arnold, London, third edition.
- Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, New York.

References III

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, New York.