8 The Likelihood Principle

Introduction

Consider parametric Bayesian inference. Since

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)d\theta},$$

It follows that if the prior $\pi(\theta)$ is fixed, the posterior $\pi(\theta \mid x)$ will be the same for any models $f_1(x \mid \theta)$ and $f_2(x \mid \theta)$ for which the likelihoods are proportional as functions of θ. In particular (with fixed prior) Bayesian inference will be the same in each case.

Example: First, suppose that $x \sim \text{Bin}(n, \theta)$, then

$$f_1(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},$$

and

$$\begin{aligned}\pi(\theta \mid x) &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\pi(\theta) \\ &\propto \theta^x(1-\theta)^{n-x}\pi(\theta).\end{aligned}$$

Secondly, suppose $n \sim$ Negative Binomial(x, θ).

$$f_2(n \mid \theta) = \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x},$$

and

$$\pi(\theta \mid n) \propto \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x}\pi(\theta).$$

Thus, given a particular prior for θ and the information that x successes have been observed in n trials, Bayesian inference about θ does not depend on whether the Binomial or negative Binomial sampling scheme was used.

Definition: The **Likelihood Principle** is the assertion that the information brought by an observation x about θ is entirely contained in the likelihood function $\ell(\theta \mid x)$ ($\equiv f(x \mid \theta)$ considered as a function of θ), and that, moreover, if x_1 and x_2 are two observations depending on the same parameter θ such that there exists a constant c with

$$\ell_1(\theta \mid x) = c\ell_2(\theta \mid x)$$

for every θ, they bring the same information about θ and must lead to identical inferences.

We have seen that Bayesian inference automatically satisfies the likelihood principle. Many frequentist procedures do not (an exception is maximum likelihood estimation) so that its validity is controversial.

Example: Suppose we observe $(x_1, \ldots, x_n)$=$(0, \ldots, 0, 1)$ with the x's being {0,1} valued quantities. If the sampling is Bin(n, θ) the classical unbiased estimated of θ is $\frac{1}{n}$. If sampling is Geometric(θ), then the only unbiased estimator of θ is

$$\tilde{\theta} = \begin{cases} 1 & n = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Properties like unbiasedness typically violate the likelihood principle because, involving integrals over the sample space, they depend on the value of $f(x \mid \theta)$ for values of x other than that actually observed.

Example: In the setting of the example on page 7.1, suppose we observe 9 heads and 3 tails in 12 tosses of a coin for which we wish to test $H_0 : \theta = \frac{1}{2}$ vs. $H_1 : \theta > \frac{1}{2}$, where θ is the success probability.

a) If the experimenter tossed the coin a fixed number 12, of times, the p-value of the classical Uniformly most powerful test is, for $x \sim$ Bin$(12, 1/2)$,

$$P(x \geq 9) = 0.075.$$

b) If the experimenter continued tossing until 3 tails were recorded, the p-value of the Uniformly most powerful test, for $n \sim$ Negative Binomial$(3, \frac{1}{2})$, is

$$P(n \geq 9) = 0.0325.$$

Thus, (eg. testing at level 5%) in stark contrast to the likelihood principle, different conclusions would be reached for the two sampling models.

Classical significance tests violate the likelihood priniciple because decisions depend on the probabilities of events which haven't occurred. The events whose probabilities are assessed in (a) and (b) above actually belong to different sample spaces.

As Jeffreys put it:

> a hypothesis which may be true may be rejected because it did not predict observable results which have not yet occurred.

In the example above (a binomial or negative binomial) observation larger than 9 is unlikely if $\theta = \frac{1}{2}$, and indeed such values *did not occur*. However, the probabilities of these unpredicted and unobserved observations are included in the classical evidence against the hypothesis.

Example: Suppose x takes values 1, 2, or 3, and that θ is 1 or 0, with mass functions:

x	1	2	3
$f(x \mid \theta = 0)$	0.009	0.001	0.990
$f(x \mid \theta = 1)$	0.001	0.989	0.010

The classical most powerful test of $H_0 : \theta = 0$ vs. $H_1 : \theta = 1$ at level $\alpha = 0.01$ (type I error), rejects H_0 if x = 1 or 2. This test also has type II error probability of 0.01.

A standard frequentist, on observing $x = 1$, would thus reject H_0 and report that the procedure has both error probabilities equal to 0.01.

On the other hand, when $x = 1$, the likelihood ratio (here also the Bayes Factor) for $\theta = 0$ to $\theta = 1$ is 9, indicating reasonable evidence in favor of H_0. The properties of the test depend strongly on behavior when $x = 2$ or $x = 3$, values which did not occur.

7.2 The Conditionality Perspective

Recall that typical frequentist measures of performance are pre-data measures, in the sense that they involve averages over the data which might arise. We have seen that such pre-data measures can be quite misleading once particular data values have been observed, even in the frequentist sense of long term behavior.

Example: (Cox 1958) A scientist needs to measure a physical quantity θ. One possibility is to use an accurate, but frequently busy, machine which provides a measurement $x_1 \sim N(\theta, 1)$. The other is to use a less accurate, but always available machine which gives a measurement $x_2 \sim N(\theta, 100)$. The availability of the accurate machine is beyond the scientists control (and independent of the object to be measured). On any particular occasion, it is available with probability 0.5, and if available the scientist will use it.

On the particular occasion of interest the scientist used the accurate machine. It seems natural (if not overwhelmingly obvious) to condition on the machine used, and for example, report 95% confidence interval for θ of $[x_1 - 1.96, x_1 + 1.96]$.

Note that the standard C.I., of (approximately) $[x_1 - 16.4, x_1 + 16.4]$ is much wider, because of

the possibility that the second machine may have been used.

Example: Pratt 1962 "An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate voltmeter, accurate enough so that measurement error is negligible compared with the variability of the tubes. A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary normal analysis, giving a confidence interval for the true mean. Later he visits the engineer's laboratory, and notices that the voltmeter used reads only as far as 100, so the population appears to be "censored". This necessitates a new analysis, if the statistician is orthodox. However, the engineer says he has another meter, equally accurate and reading to 1000 volts, which he would have used if any voltage had been over 100. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all. But the next day the engineer telephones and says, 'I just discovered my high-range voltmeter was not working the day I did the experiment you analyzed for me.' The statistician ascertains that the engineer would not have held up the experiment until the meter was fixed, and informs him that a new analysis will be required. The engineer is astounded. He says, 'But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope."

In this example, two different sample spaces are being discussed. If the high-range voltmeter had been working, the sample space would have effectively been that of a usual normal distribution. Since the high-range voltmeter was broken, however, the sample space was truncated at 100, and the probability distribution of the observations would have a point mass at 100. Classical analyses (such as the obtaining of confidence intervals) would be considerably affected by this difference. The Likelihood Principle, on the other hand, states that this difference should have no effect on the analysis, since values of x which did not occur (here $x > 100$) have no bearing on inferences or decisions concerning the true mean.

These considerations lead naturally to the:

Conditionality Principle: If two experiments on the parameter θ are available, $\mathcal{E}_1$ and $\mathcal{E}_2$ say, and if one of these two experiments is selected with probability $\frac{1}{2}$, then the resulting inference on θ should only depend on the selected experiment.

This principle seems difficult to reject, as the previous two examples show.

The conditionality principle is an extremely weak version of the suggestion that inference should be conditional on the observed data.

Note that Bayesian inference procedures operate conditionally on the observed data. Thus, in particular they satisfy the Conditionality Principle.

The most common type of partial conditioning advocated in frequentist statistics is conditioning on an ancillary statistic. Recall that an ancillary statistic is one whose distribution is independent of the paramter θ of interest.

For example, in the Cox example above, the indicator of which machine is used is ancillary. Conditioning on the ancillary provides "better" procedures.

A problem with frequentist attempts towards conditional inference is that they are necessarily *ad hoc*. For example, what should be done in the absence of an ancillary statistic?

Example: Suppose $\Theta = [0, \frac{1}{2})$ and

$$x = \begin{cases} \theta & \text{with probability } 1-\theta \\ 0 & \text{with probability } \theta \end{cases}$$

Think of an instrument which measures θ exactly, except that with probability θ it will report the value 0. Now, if $x > 0$, we know $x = \theta$, so it seems natural to condition on $\{x > 0\}$, but there is no ancillary statistic which provides such a conditioning.

Similarly, there is no ancillary statistic for conditioning which overcomes the problems in the example on page 7.3.

7.3 Sufficiency, Conditionality and the Likelihood Principle

In this section we will show that the likelihood principle is equivalent to (the conjuction of) two other very natural, and almost uniformly accepted, principles. The first is the conditionality priniciple. The second is the

Sufficiency Principle: Two observations x and y which lead to the same value $T(x) = T(y)$ of a sufficient statistic for θ must lead to the same inference on θ.

Theorem (Birnbaum, 1962): The Likelihood Principle is equivalent to the conjunction of the Conditionality Principle and the Sufficiency Principle.

The proof is not difficult, see for example Robert, page 18.

Thus, if one violates the Likelihood Principle, one is violating either (or both) of Conditionality

Principle and the Sufficiency Principle. The likelihood principle says that inferences or decisions about θ should be based on the likelihood function. It does not stipulate how this should be done.

The Bayesian approach is one universal method which is consistent with the likelihood principle. Some frequentist procedures are also consistent with it. Other are, in some settings. Note that the likelihood principle states that, for (parametric) inference, it is an error to reason in terms of pre-data measures of performance. One should reason only in terms of the actual sample and likelihood function obtained. This applies to parametric inference. It doesn't apply, for example, to experimental design.