Statistical inference
Part 2
The Likelihood Principle.

Michael Goldstein
Durham University
APTS December 2023

Outline.

In this section, we will introduce the notion of the evidence that is provided by data in a statistical model.

We will then consider what (arguably) self-evident properties such evidence should satisfy.

We will then show that these “self-evident properties” imply the likelihood principle - informally, that our inference should only depend on the likelihood function.

We will then explore some important implications of the likelihood principle.

As many standard statistical procedures do not obey the likelihood principle, this will raise some interesting issues.

Notation

Our notation for the general class of models, $\mathcal{E}$ that we shall consider is

$$\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$$

where $x \in \mathcal{X}$ (typically vector valued) is the collection of possible observations that we may make

The probability distribution of x depends on the model parameter $\theta \in \Theta$

and is of form $f_X(x \mid \theta)$

We assume that the model is true, so that only $\theta \in \Theta$ is unknown.

We are concerned with inferences within the constraints of the model.

We wish to learn about θ from observations x so that $\mathcal{E}$ represents a model for this ***experiment***.

In this section, we shall assume that $\mathcal{X}$ is finite. [This is to simplify our proofs - and real observations have finite outcome spaces anyway.]

Reasoning about inferences

We consider a series of statistical principles to guide the way to learn about θ.

The principles are meant to be either self-evident or logical implications of principles which are self-evident.

The model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ is accepted as a working hypothesis.

How the statistician chooses the inference statements about the true value θ is entirely down to the requirements of the problem, for example:

- as a point or a set in Θ;
- as a choice among alternative sets or actions;
- or maybe as something more complicated.

We suppose that the statistician defines, a priori, a set of possible inferences about θ. The task is to choose an element of this set based on $\mathcal{E}$ and x.

In this way, the inference can be viewed as part of the wider model specification.

Inference

In this formulation, the statistician becomes a function Ev: a mapping from $(\mathcal{E}, x)$ into a predefined set of inferences about θ.

$$(\mathcal{E}, x) \xmapsto{\text{statistician, Ev}} \text{Inference about } \theta.$$

For example, Ev$(\mathcal{E}, x)$ might be:

- the maximum likelihood estimator of θ
- a 95% confidence interval for θ

The approach draws on the work of Birnbaum who called
$\mathcal{E}$ the experiment,
x the outcome, and
Ev the evidence.

Birnbaum's formulation

It was the inspiration of Allan Birnbaum (1923-1976) to see how to construct and reason about statistical principles given evidence from data.

This reasoning works at a very general level, and so applies to any form of evidence.

Our aim is not to identify a particular form of inference for any particular problem, but rather to rule out certain kinds of inference for such problems.

Two particular features of interest are

(i) our attention is drawn to certain types of inference, based around properties of the likelihood function,

(ii) this formulation appears to rule out many statistical procedures in current practice.

Equivalence of evidence

There can be different experiments with the same parameter θ.

Consider two experiments
$\mathcal{E}_1 = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \mid \theta)\}$ and
$\mathcal{E}_2 = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \mid \theta)\}$.

Under some outcomes, we would agree that it is self-evident that these different experiments provide the same evidence about θ.

The equality or equivalence of

$\mathsf{Ev}(\mathcal{E}_1, x_1)$ and $\mathsf{Ev}(\mathcal{E}_2, x_2)$ means that:

1. $\mathcal{E}_1$ and $\mathcal{E}_2$ are related to the same parameter θ.
2. Everything else being equal, the outcome x_1 from $\mathcal{E}_1$ warrants the same inference about θ as does the outcome x_2 from $\mathcal{E}_2$.

Statistical principles

We now consider constructing statistical principles and demonstrate how these principles imply other principles.

These principles all have the same form:

under such and such conditions, the evidence about θ should be the same in two separate cases.

Thus they serve only to rule out inferences that satisfy the conditions but have different evidences.

They do not tell us how to make an inference, only what to avoid.

Distribution principle (DP)

Our first principle, **the distribution principle**, sets up the constraints of our inference.

The distribution principle is as follows.

Suppose that $\mathcal{E} = \mathcal{E}' = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$.

Then $\mathsf{Ev}(\mathcal{E}, x) = \mathsf{Ev}(\mathcal{E}', x)$.

Interpretation The only aspects of an experiment which are relevant to our inference are the sample space and the family of distributions over it.

[Therefore, any differences between the inferences for the two experiments must be due to features which are not contained in the model description.]

Transformation Principle (TP)

The **Transformation Principle** is as follows.

Let $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$.

For the bijective $g : \mathcal{X} \to \mathcal{Y}$,

let $\mathcal{E}^g = \{\mathcal{Y}, \Theta, f_Y(y \mid \theta)\}$,

the same experiment as $\mathcal{E}$ but expressed in terms of $Y = g(X)$, rather than X.

Then $\mathsf{Ev}(\mathcal{E}, x) = \mathsf{Ev}(\mathcal{E}^g, g(x))$.

Interpretation Inferences should not depend on the way in which the sample space is labelled.

For example, we should form the same evidence if we recorded X or X^{-1}.

Next steps

Assuming that we have accepted the distribution principle and the transformation principle,

we will show that in combination they imply a further principle, the weak indifference principle.

We will then add one further principle, the weak conditionality principle, which again seems reasonably self evident.

Weak indifference and weak conditionality together lead us to our main objective, namely the strong likelihood principle, (informally, that outcomes from two experiments with proportional likelihoods should lead to the same inference) which therefore we will show is a consequence of the three reasonably self evident principles.

Weak indifference principle (WIP)

The Weak Indifference Principle is as follows.

Let $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$.

If $f_X(x \mid \theta) = f_X(x' \mid \theta)$ for all $\theta \in \Theta$

then $\mathsf{Ev}(\mathcal{E}, x) = \mathsf{Ev}(\mathcal{E}, x')$.

Example Suppose $X = (X_1, \ldots, X_n)$ where the X_is are a series of independent Bernoulli trials with parameter θ.

$f_X(x \mid \theta) = f_X(x' \mid \theta)$ for all $\theta \in \Theta$ if x and x' contain the same number of successes.

Therefore, the two observations should lead to the same evidence.

Interpretation We are indifferent between two models of evidence if they differ only in the manner of the labelling of sample points.

THEOREM: DP and TP imply WIP

THEOREM

$$\big(\text{DP} \wedge \text{TP}\big) \rightarrow \text{WIP}.$$

Proof Fix $\mathcal{E}$, and suppose that $x, x' \in \mathcal{X}$ satisfy

$$f_X(x \,|\, \theta) = f_X(x' \,|\, \theta)$$

for all $\theta \in \Theta$, as in the condition of the WIP.

Let $g : \mathcal{X} \rightarrow \mathcal{X}$ be the function which switches x for x', but leaves all of the other elements of $\mathcal{X}$ unchanged. Then $\mathcal{E} = \mathcal{E}^g$ and

$$\begin{aligned} \text{Ev}(\mathcal{E}, x') &= \text{Ev}(\mathcal{E}^g, x') \quad \text{[by the DP]} \\ &= \text{Ev}(\mathcal{E}^g, g(x)) \\ &= \text{Ev}(\mathcal{E}, x), \quad \text{[by the TP]} \end{aligned}$$

which gives the WIP. □

Mixture experiments

- Consider experiments $\mathcal{E}_i = \{\mathcal{X}_i, \Theta, f_{X_i}(x_i \mid \theta)\}$, $i = 1, 2, \ldots$, where the parameter space Θ is the same for each experiment.
- Let $p_1, p_2, \ldots$ be a set of known probabilities so that $p_i \geq 0$ and $\sum_i p_i = 1$.

Mixture experiment The mixture $\mathcal{E}^*$ of the experiments $\mathcal{E}_1, \mathcal{E}_2, \ldots$ according to mixture probabilities $p_1, p_2, \ldots$ is the two-stage experiment

1. A random selection of one of the experiments: $\mathcal{E}_i$ is selected with probability p_i.
2. The experiment selected in stage 1. is performed.

Thus, each outcome of the experiment $\mathcal{E}^*$ is a pair (i, x_i), where $i = 1, 2, \ldots$ and $x_i \in \mathcal{X}_i$, and family of distributions

$$f^*((i, x_i) \mid \theta) \;=\; p_i f_{X_i}(x_i \mid \theta).$$

Mixture experiments

A famous example of a mixture experiment is the ‘two instruments’ (see Section 2.3 of Cox and Hinkley (1974)).

There are two instruments in a laboratory, and one is accurate, the other less so.

The accurate one is more in demand, and typically it is busy 80% of the time. The inaccurate one is usually free.

So, a priori, there is a probability of $p_1 = 0.2$ of getting the accurate instrument, and $p_2 = 0.8$ of getting the inaccurate one.

Once a measurement is made, there is no doubt about which of the two instruments was used.

The following principle asserts what should be self-evident, namely that inferences should be made according to which instrument was used and not according to the a priori uncertainty.

Weak Conditionality Principle (WCP)

The Weak Conditionality Principle is as follows.

Let $\mathcal{E}^*$ be the mixture of the experiments $\mathcal{E}_1, \mathcal{E}_2$ according to mixture probabilities p_1, $p_2 = 1 - p_1$. Then

$$\mathsf{Ev}\left(\mathcal{E}^*, (i, x_i)\right) = \mathsf{Ev}(\mathcal{E}_i, x_i).$$

Interpretation The WCP says that inferences for θ depend only on the experiment performed and not which experiments could have been performed.

Suppose that $\mathcal{E}_i$ is randomly chosen with probability p_i and x_i is observed.

The WCP states that the same evidence about θ would have been obtained if it was decided non-randomly to perform $\mathcal{E}_i$ from the beginning and x_i is observed.

As Casella and Berger (2002, p293) state "the fact that this experiment was performed rather than some other, has not increased, decreased, or changed knowledge of θ."

Likelihood notation

We have an experiment

$$\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$$

where $x \in \mathcal{X}$, is the collection of possible observations that we may make

The probability distribution of x depends on the model parameter $\theta \in \Theta$

and is of form $f_X(x \mid \theta)$

Notation If we observe x, then the likelihood function is

$$L(\theta) = L_X(\theta; x) = f(x \mid \theta)$$

considered as a function of θ.

Strong Likelihood Principle (SLP)

The Strong Likelihood Principle, SLP is as follows.

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter θ.

If $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy

$$f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta), \ \forall \theta \in \Theta,$$

or, equivalently,

$$L_{X_1}(\theta; x_1) \ = \ c(x_1, x_2) L_{X_2}(\theta; x_2)$$

for some function $c > 0$ for all $\theta \in \Theta$ then

$$\mathsf{Ev}(\mathcal{E}_1, x_1) = \mathsf{Ev}(\mathcal{E}_2, x_2).$$

Interpretation The SLP states that if two likelihood functions for the same parameter have the same shape, then the evidence is the same.

Likelihood inference

A fundamental corollary of the SLP is obtained by setting

$$\mathcal{E}_1 = \mathcal{E}_2 = \mathcal{E}$$

We have that

$$\mathsf{Ev}(\mathcal{E}, x)$$

should depend on $\mathcal{E}$ and x **only** through $L_X(\theta; x)$.

Birnbaum's theorem

Many classical statistical procedures violate the SLP and the following result was something of the bombshell, when it first emerged in the 1960s.

The following form is due to:
Birnbaum, A. (1972). More concepts of statistical evidence. Journal of the American Statistical Association 67, 858–861.
Basu, D. (1975). Statistical information and likelihood. Sankhya 37(1), 1–71.

Birnbaum's Theorem

$$\big(\text{WIP} \wedge \text{WCP}\big) \leftrightarrow \text{SLP}.$$

Proof

Both SLP $\rightarrow$ WIP and SLP $\rightarrow$ WCP are straightforward.

The trick is to prove $\big(\text{WIP} \wedge \text{WCP}\big) \rightarrow \text{SLP}$.

Birnbaum's theorem: Proof

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter.

Suppose that $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy, for all θ,

$$f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$$

where the function $c > 0$.

As the value c is known, consider the mixture experiment with

$\mathcal{E}_1$ with probability $p_1 = 1/(1+c)$

and $\mathcal{E}_2$ with probability $p_2 = c/(1+c)$.

Birnbaum's theorem: Proof continued

$$f^*((1,x_1)\,|\,\theta) \;=\; \frac{1}{1+c} f_{X_1}(x_1\,|\,\theta) \;=\; \frac{c}{1+c} f_{X_2}(x_2\,|\,\theta) \;=\; f^*((2,x_2)\,|\,\theta)$$

Then the WIP implies that

$$\mathsf{Ev}\,(\mathcal{E}^*,(1,x_1)) \quad = \quad \mathsf{Ev}\,(\mathcal{E}^*,(2,x_2)).$$

Applying the WCP to each side we infer that

$$\mathsf{Ev}(\mathcal{E}_1,x_1) \quad = \quad \mathsf{Ev}(\mathcal{E}_2,x_2),$$

as required. □

Comment Either I accept the SLP, or I explain which of the two principles, WIP and WCP, I refute.

Methods, which include many classical procedures, which violate the SLP face exactly this challenge.

Coming up.

So far, we have derived the likelihood principle from some weak inferential conditions.

We will now look at various implications of the likelihood pronciple.

Firstly, we consider sufficient statistics and whether it is appropriate to construct inferences solely on their values.

Secondly, we consider the implications of the likelihood principle for stopping rules in sequential sampling.

Finally, we consider the appropriate treatment of ancillary quantities in inference.

Then we will consider further issues as to how the likelihood principle applies in practice.

Recap: the strong likelihood principle

We have shown that, under some weak assumptions, we may deduce **The Strong Likelihood Principle, SLP** which is as folows.

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter θ. If $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy $f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$, that is

$$L_{X_1}(\theta; x_1) \;=\; c(x_1, x_2) L_{X_2}(\theta; x_2)$$

for some function $c > 0$ for all $\theta \in \Theta$ then

$$\mathsf{Ev}(\mathcal{E}_1, x_1) = \mathsf{Ev}(\mathcal{E}_2, x_2).$$

The SLP states that if two likelihood functions for the same parameter have the same shape, then the evidence is the same.

In particular, $\mathsf{Ev}(\mathcal{E}, x)$ should depend on $\mathcal{E}$ and x only through $L_X(\theta; x)$.

We now explore some implications of SLP.

Sufficiency

Recall the idea of sufficiency:

$S = s(X)$ is sufficient for θ if and only if

$$f_X(x \mid \theta) \quad = \quad f_{X|S}(x \mid s, \theta) f_S(s \mid \theta)$$

where $f_{X|S}(x \mid s, \theta)$ does not depend upon θ.

A natural consequence of sufficiency is that two samples with the same value for the sufficient statistic should result in the same inference.

The Sufficiency Principle

This suggests the following principle.

Weak Sufficiency Principle, WSP

If $S = s(X)$ is a sufficient statistic for $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ and

$$s(x) = s(x')$$

then

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

Strong Sufficiency

A stronger version of the sufficiency principle is as follows.

Suppose that we record only the value of the sufficient statistic for an experiment.

This results in a modified experiment

$$\mathcal{E}^S = \{s(\mathcal{X}), \Theta, f_S(s \mid \theta)\}.$$

Is our reduced experiment the same as the original experiment?

The Strong Sufficiency Principle

The strong likelihood principle says that it is.

Strong Sufficiency Principle, SSP

If $S = s(X)$ is a sufficient statistic for $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ then

$$\mathsf{Ev}(\mathcal{E}, x) = \mathsf{Ev}(\mathcal{E}^S, s(x)).$$

What is the justification for these two principles?

Likelihood principle implies sufficiency principle

The strong likelihood principle implies the sufficiency principles as follows.

Theorem SLP $\rightarrow$ SSP $\rightarrow$ WSP.

Proof As s is sufficient,

$$f_X(x \mid \theta) = c f_S(s \mid \theta)$$

where $c = f_{X|S}(x \mid s, \theta)$ does not depend on θ.

Applying the SLP,

$$\mathsf{Ev}(\mathcal{E}, x) = \mathsf{Ev}(\mathcal{E}^S, s(x))$$

which is the SSP.

Proof continued

Note, that from the SSP, if $s(x) = s(x')$, then

$$\begin{aligned} \mathsf{Ev}(\mathcal{E}, x) &= \mathsf{Ev}(\mathcal{E}^S, s(x)) && \text{(by the SSP)} \\ &= \mathsf{Ev}(\mathcal{E}^S, s(x')) && \text{(by the SLP)} \\ &= \mathsf{Ev}(\mathcal{E}, x') && \text{(by the SSP)} \end{aligned}$$

We thus have the WSP.

□

Discussion

In practice, sufficiency principles are widely accepted and used, even within the classical framework.

However, the basic rationale for sufficiency - that inference should not depend on irrelevant randomisations - is similar in spirit to the likelihood principle itself. Hence the formal link we have established.

Note that there are important considerations that we are excluding from the strong sufficiency principle, such as testing goodness of fit of the model.

Like all of our principles in this section, these inferences take place within the world in which the model is assumed true.

Stopping rules

Consider observing a sequence of random variables $X_1, X_2, \ldots$ where the number of observations is not fixed in advance but depends on the values seen so far.

- At time j, the decision to observe X_{j+1} can be modelled by a probability $p_j(x_1, \ldots, x_j)$.
- We assume, resources being finite, that the experiment must stop at specified time m, if it has not stopped already, hence $p_m(x_1, \ldots, x_m) = 0$.

The stopping rule may then be denoted as $\tau = (p_1, \ldots, p_m)$. This gives an experiment $\mathcal{E}^\tau$ with distribution $f_n(x_1, \ldots, x_n \,|\, \theta)$ for $n = 1, 2, \ldots$, where consistency requires that

$$f_n(x_1, \ldots, x_n \,|\, \theta) \;=\; \sum_{x_{n+1}} \cdots \sum_{x_m} f_m(x_1, \ldots, x_n, x_{n+1}, \ldots x_m \,|\, \theta).$$

Motivation for the stopping rule principle (Basu, 1975)

Consider four different coin-tossing experiments (with some finite limit on the number of tosses).

$\mathcal{E}_1$ Toss the coin exactly 10 times;
$\mathcal{E}_2$ Continue tossing until 6 heads appear;
$\mathcal{E}_3$ Continue tossing until 3 consecutive heads appear;
$\mathcal{E}_4$ Continue tossing until the accumulated number of heads exceeds that of tails by exactly 2.

Suppose that all four experiments have the same outcome
$x = (\mathsf{T},\mathsf{H},\mathsf{T},\mathsf{T},\mathsf{H},\mathsf{H},\mathsf{T},\mathsf{H},\mathsf{H},\mathsf{H})$.
We may feel that the evidence for θ, the probability of heads, is the same in every case.
Once the sequence of heads and tails is known, the intentions of the original experimenter (i.e. the experiment she was doing) are immaterial to inference about the probability of heads.
The simplest experiment $\mathcal{E}_1$ can be used for inference.

Stopping Rule Principle, SRP

The SRP is as follows.

In a sequential experiment $\mathcal{E}^\tau$, Ev $(\mathcal{E}^\tau, (x_1, \ldots, x_n))$ does not depend on the stopping rule τ.

Comment Basu (1975) claims the SRP is due to George Barnard.

If it is accepted, the SRP is revolutionary.

It implies that the intentions of the experimenter, represented by τ, are irrelevant for making inferences about θ, once the observations $(x_1, \ldots, x_n)$ are known. Once the data is observed, we can ignore the sampling plan.

The statistician could proceed as though the simplest possible stopping rule were in effect, which is $p_1 = \cdots = p_{n-1} = 1$ and $p_n = 0$, an experiment with n fixed in advance, $\mathcal{E}^n = \{\mathcal{X}_{1:n}, \Theta, f_n(x_{1:n} \mid \theta)\}$.

Can the SRP be justified?

The likelihood principle implies the stopping rule principle

Theorem

$$\text{SLP} \rightarrow \text{SRP}$$

Proof Let τ be an arbitrary stopping rule, and consider the outcome $(x_1, \ldots, x_n)$, which we will denote as $x_{1:n}$.
We take the first observation with probability one.
For $j = 1, \ldots, n-1$, the $(j+1)$th observation is taken with probability $p_j(x_{1:j})$.
We stop after the nth observation with probability $1 - p_n(x_{1:n})$.

Consequently, the probability of this outcome under τ is

$$f_\tau(x_{1:n} \,|\, \theta) \quad = \quad f_1(x_1 \,|\, \theta) \left\{ \prod_{j=1}^{n-1} p_j(x_{1:j})\, f_{j+1}(x_{j+1} \,|\, x_{1:j}, \theta) \right\} (1 - p_n(x_{1:n}))$$

Proof continued

$$
\begin{aligned}
f_\tau(x_{1:n}\,|\,\theta) &= \left\{\prod_{j=1}^{n-1} p_j(x_{1:j})\right\}(1-p_n(x_{1:n}))\, f_1(x_1\,|\,\theta)\prod_{j=2}^{n} f_j(x_j\,|\,x_{1:(j-1)},\theta) \\
&= \left\{\prod_{j=1}^{n-1} p_j(x_{1:j})\right\}(1-p_n(x_{1:n}))\, f_n(x_{1:n}\,|\,\theta).
\end{aligned}
$$

Now observe that this equation has the form

$$
f_\tau(x_{1:n}\,|\,\theta) = c(x_{1:n}) f_n(x_{1:n}\,|\,\theta) \tag{1}
$$

where $c(x_{1:n}) > 0$.
Thus the SLP implies that $\mathsf{Ev}(\mathcal{E}^\tau, x_{1:n}) = \mathsf{Ev}(\mathcal{E}^n, x_{1:n})$ where $\mathcal{E}^n = \{\mathcal{X}_{1:n}, \Theta, f_n(x_{1:n}\,|\,\theta)\}$. Since the choice of stopping rule was arbitrary, equation (1) holds for all stopping rules, showing that the choice of stopping rule is irrelevant. □

Discussion

A comment from Leonard Jimmie Savage, one of the great statisticians of the Twentieth Century, captured the revolutionary and transformative nature of the SRP.

> May I digress to say publicly that I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right.
>
> (Savage, 1962, Foundations of statistical inference)

Ancillarity

Consider the concept of **ancillarity**.

This has several different definitions in the Statistics literature; the one we use is close to that of Cox and Hinkley (1974, Section 2.2) Theoretical Statistics.

Definition (Ancillarity)

Y is ancillary in the experiment $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \mid \theta)\}$ exactly when $f_{X,Y}$ factorises as

$$f_{X,Y}(x, y \mid \theta) \;=\; f_Y(y) f_{X|Y}(x \mid y, \theta).$$

The marginal distribution of Y is completely specified: it does not depend on θ.

Ancillarity examples

We have already considered versions of this.

A familiar example is that of a random sample size:

in a sample $\underline{x} = (x_1, \ldots, x_n)$,

n may be the outcome of a random variable N.

We seldom concern ourselves with the distribution of N when we evaluate $\underline{x}$; instead we treat N as known.

Equivalently, we treat N as ancillary and condition on $N = n$.

Therefore, we consider that inferences drawn from observing (n, x) should be the same as those for x conditioned on $N = n$.

Ancillarity in regression

Another common example arises when we perform a regression of Z on U, say to estimate the parameters in

$$z = \beta u + \alpha + \epsilon$$

from a sample of (U, Z) values.

Both U and Z are random, but U is treated as ancillary for the parameters in $f_{Z|U}$.

We model Z conditionally on U, treating U as known.

Therefore, we estimate the regression parameters, for example by least squares, and then assess the estimation error treating U values as fixed.

Strong conditionality

When Y is ancillary, we can consider the conditional experiment

$$\mathcal{E}^{X\,|\,y} = \{\mathcal{X}, \Theta, f_{X\,|\,Y}(x\,|\,y, \theta)\}.$$

That is, we treat Y as known, and treat X (conditional on $Y = y$) as the only random variable.

Strong Conditionality Principle, SCP

If Y is ancillary in $\mathcal{E}$, then $\mathsf{Ev}\,(\mathcal{E}, (x, y)) = \mathsf{Ev}(\mathcal{E}^{X|y}, x)$.

[The SCP implies the WCP, with the experiment indicator $I \in \{1, 2\}$ being ancillary.]

Ancillarity theorem

We have the following result.

Theorem SLP $\rightarrow$ SCP.

Proof Suppose that Y is ancillary in $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \,|\, \theta)\}$. Thus, for all $\theta \in \Theta$,

$$\begin{aligned} f_{X,Y}(x, y \,|\, \theta) &= f_Y(y) f_{X|Y}(x \,|\, y, \theta) \\ &= c(y) f_{X|Y}(x \,|\, y, \theta) \end{aligned}$$

Then the SLP implies that

$$\mathsf{Ev}\left(\mathcal{E}, (x, y)\right) \;=\; \mathsf{Ev}(\mathcal{E}^{X|y}, x),$$

as required. $\square$

The Likelihood Principle in practice

Which inferential approaches respect the SLP?

A **Bayesian statistical model** adds prior $\pi(\theta)$ giving the collection

$$\mathcal{E}_B \;=\; \{\mathcal{X}, \Theta, f_X(x \,|\, \theta), \pi(\theta)\}.$$

The **posterior distribution** is $\pi(\theta \,|\, x) = c(x) f_X(x \,|\, \theta)\pi(\theta)$ where $c(x)$ is the normalising constant,

$$c(x) \;\;=\;\; \left\{ \int_\Theta f_X(x \,|\, \theta)\pi(\theta)\, d\theta \right\}^{-1}.$$

All knowledge about θ given the data x is represented by $\pi(\theta \,|\, x)$.

Any inferences made about θ are derived from this distribution.

Bayesian models

Consider two Bayesian models with the same prior distribution, $\pi(\theta)$,

$$\mathcal{E}_{B,1} = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \,|\, \theta), \pi(\theta)\}$$

$$\mathcal{E}_{B,2} = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \,|\, \theta), \pi(\theta)\}$$

Suppose that $f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$. Then

$$\begin{aligned} \pi_1(\theta \,|\, x_1) \;=\; c(x_1) f_{X_1}(x_1 \,|\, \theta)\pi(\theta) \quad &= \quad c(x_1)c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)\pi(\theta) \\ &= \quad \pi_2(\theta \,|\, x_2) \end{aligned}$$

Hence, the posterior distributions are the same.

Consequently, the same inferences are drawn from either model and so the Bayesian approach satisfies the SLP.

Comment

This argument assumes that $\pi(\theta)$ does not depend upon the form of the data.

Some methods for making ”default” choices for $\pi(\theta)$ depend on $f_X(x \mid \theta)$.

For example, Jeffreys priors and reference priors.

These methods violate the SLP.

Violating the LP

Maximum likelihood estimation clearly satisfies the SLP and methods, such as penalised likelihood theory, have been generated to satisfy the SLP.

However, inference tools used in the classical approach typically violate the SLP.

These inference techniques depend upon the sampling distribution and therefore depend on the whole sample space $\mathcal{X}$ and not just the observed $x \in \mathcal{X}$.

The sampling distribution depends on values of f_X other than $L(\theta; x) = f_X(x \mid \theta)$.

For example, a statistic $T(X)$, chosen on the basis of a criterion such as

$$MSE(T \mid \theta) = Var(T \mid \theta) + \text{bias}(T \mid \theta)^2$$

depends upon the first and second moments of the distribution of $T \mid \theta$.

Binomial and Negative Binomial example

Here's a simple example of the ways in which traditional inferences can contradict the likelihood principle.

We wish to test the hypothesis $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta < \frac{1}{2}$ where θ is the probability that a single random trial is a success.

Experiment one: n independent trials. Count the number of successes X.

Experiment two: count the number of independent trials Y until the number of successes is r.

Binomial and Negative Binomial example

Let $\mathcal{E}_1 = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$, where $X \mid \theta \sim \text{Bin}(n, \theta)$ so that

$$f_X(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n.$$

Let $\mathcal{E}_2 = \{\mathcal{Y}, \Theta, f_Y(y \mid \theta)\}$, where $Y \mid \theta \sim \text{Nbin}(r, \theta)$, so that

$$f_Y(y \mid \theta) = \binom{y-1}{r-1} \theta^r (1-\theta)^{y-r}, \quad y = r, r+1, \ldots.$$

Example ctd

Suppose we observe $x = r = 3$ and $y = n = 12$, so that in each experiment there were 12 trials and 3 successes. Therefore

$$f_X(3 \mid \theta) = \binom{12}{3}\theta^3(1-\theta)^9$$

$$f_Y(12 \mid \theta) = \binom{11}{2}\theta^3(1-\theta)^9$$

Thus, $f_X(3 \mid \theta) \propto f_Y(12 \mid \theta)$.

SLP implies each experiment should reach the same conclusion, as we have seen in our discussion of stopping rules.

Suppose we assess the hypothesis test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta < \frac{1}{2}$ by carrying out a significance test, at conventional significance level 5%, for each experiment.

Example ctd

Let $\mathsf{Ev}(\mathcal{E}_1, 3)$ be the result of the hypothesis test for the Binomial model where small values of X support H_1

$$\mathbb{P}(X \leq 3 \,|\, \theta = 1/2) \;=\; \sum_{x=0}^{3} f_X(x \,|\, \theta = 1/2) \;=\; 0.0730.$$

Thus, $\mathsf{Ev}(\mathcal{E}_1, 3)$ is to not reject H_0, using conventional 5% significance.
Let $\mathsf{Ev}(\mathcal{E}_2, 12)$ be the result of the hypothesis test for the Negative Binomial model where large values of Y support H_1

$$\mathbb{P}(Y \geq 12 \,|\, \theta = 1/2) \;=\; \sum_{y=12}^{\infty} f_Y(y \,|\, \theta = 1/2) \;=\; 0.0327.$$

Thus, $\mathsf{Ev}(\mathcal{E}_2, 12)$ is to reject H_0. This inference method does not respect the SLP: the choice of the model is relevant to the inference.

Discussion 1

Any inferential method which relies on values of $f(x'|\theta)$ other than the observed value of x (like significance tests) will violate the likelihood principle.

Among the difficulties with violating the SLP are the following:

To reject the SLP is to reject at least one of the WIP and the WCP.
Yet both of these principles seem self-evident.
Therefore it is hard to justify violating the SLP.

In their everyday practice, statisticians use the SRP (ignoring the intentions of the experimenter) and the SCP (conditioning on ancillary statistics) .
These are not self-evident, but they are implied by the SLP.
If the SLP is violated, they need an alternative justification which has not yet been forthcoming.

Discussion 2

Our framework assumes the truth of the model and that the evidence that we shall produce has been decided before we conduct the experiment.

All that we don't know is the experimental outcome.

In this formulation, the statistician sets up the framework and then leaves.

Issues around the likelihood principle (and many other statistical principles) are more complicated when the statistician stays around!