Contents

2
Principles for Statisical Inference

This chapter will be a lot clearer if you have recently read Chapter 1. An extremely compressed version follows. As a working hypothesis, we accept the truth of a statistical model

$$\mathcal{E} := \{\mathcal{X}, \Omega, f\} \tag{2.1}$$

where $\mathcal{X}$ is the realm of a set of random quantities X, θ is a parameter with domain Ω (the 'parameter space'), and f is a probability mass function for which $f(x;\theta)$ is the probability of $X = x$ under parameter value θ.[1] The true value of the parameter is denoted Θ. Statistical inference is learning about Θ from the value of X, described in terms of an algorithm involving $\mathcal{E}$ and x. Although Statistics is really about prediction, inference is a crucial step in prediction, and therefore often taken as a goal in its own right.

Statistical principles guide the way in which we learn about Θ. They are meant to be either self-evident, or logical implications of principles which are self-evident. What is really interesting about Statistics, for both statisticians and philosophers (and real-world decision makers) is that the logical implications of some self-evident principles are not at all self-evident, and have turned out to be inconsistent with prevailing practices. This was a discovery made in the 1960s. Just as interesting, for sociologists (and real-world decision makers) is that the then-prevailing practices have survived the discovery, and continue to be used today.

This chapter is about statistical principles, and their implications for statistical inference. It demonstrates the power of abstract reasoning to shape everyday practice.

2.1 Reasoning about inferences

Statistical inferences can be very varied, as a brief look at the 'Results' sections of the papers in an Applied Statistics journal will reveal. In each paper, the authors have decided on a different interpretation of how to represent the 'evidence' from their dataset. On the surface, it does not seem possible to construct and reason about statistical principles when the notion of 'evidence' is so plastic. It was the inspiration of Allan Birnbaum (Birnbaum, 1962) to see—albeit indistinctly at first—that this issue could be side-stepped.

[1] As is my usual convention, I assume, without loss of generality, that $\mathcal{X}$ is countable, and that Ω is uncountable.

Over the next two decades, his original notion was refined; key papers in this process were Birnbaum (1972), Basu (1975), Dawid (1977), and the book by Berger and Wolpert (1988).

The model $\mathcal{E}$ is accepted as a working hypothesis, and so the existence of the true value Θ is also accepted under the same terms. How the statistician chooses her statements about the true value Θ is entirely down to her and her client: as a point or a set in Ω, as a choice among alternative sets or actions, or maybe as some more complicated, not ruling out visualizations. Dawid (1977) puts this well—his formalism is not excessive, for really understanding this crucial concept. The statistician defines, *a priori*, a set of possible 'inferences about Θ', and her task is to choose an element of this set based on $\mathcal{E}$ and x. Thus the statistician should see herself as a function 'Ev': a mapping from $(\mathcal{E}, x)$ into a predefined set of 'inferences about Θ', or

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

Birnbaum called $\mathcal{E}$ the 'experiment', x the 'outcome', and Ev the 'evidence'.

Birnbaum's formalism, of an experiment, an outcome, and an evidence function, helps us to anticipate how we can construct statistical principles. First, there can be different experiments with the same Θ. Second, under some outcomes, we would agree that it is self-evident that these different experiments provide the same evidence about Θ. Finally, as will be shown, these self-evident principles imply other principles. These principles all have the same form: under such and such conditions, the evidence about Θ should be the same. Thus they serve only to rule out inferences that satisfy the conditions but have different evidences. They do not tell us how to do an inference, only what to avoid.

But if you find the idea of 'Ev' too abstract, then replace it in your mind and your notes with a specific instance of 'Ev', such as the ML estimate or a 95% confidence interval. E.g., everywhere you see 'Ev', read it as 'ML estimate of Θ'.

2.2 *The principle of indifference*

Here is our first example of a statistical principle, using the name conferred by Basu (1975). Recollect that once $f(x;\theta)$ has been defined, $f(x;\bullet)$ is a function of θ, potentially a different function for each x, and $f(\bullet;\theta)$ is a function of x, potentially a different function for each θ.[2]

Definition 2.1 (Weak Indifference Principle, WIP)**.** Let $\mathcal{E} = \{\mathcal{X}, \Omega, f\}$. If $x, x' \in \mathcal{X}$ satisfy $f(x;\bullet) = f(x';\bullet)$, then $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}, x')$.

In my opinion, this is not self-evident, although, at the same time, is it not obviously wrong.[3] But we discover that it is the

[2] I am using '•' instead of '·' in this chapter and subsequent ones, because I like to use '·' to denote scalar multiplication.

[3] Birnbaum (1972) thought it was self-evident.

logical implication of two other principles which I accept as self-evident. These other principles are as follows, using the names conferred by Dawid (1977).

Definition 2.2 (Distribution Principle, DP)**.** If $\mathcal{E} = \mathcal{E}'$, then $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}', x)$.

As Dawid (1977) puts it, any information which is not represented in $\mathcal{E}$ is irrelevant. This seems entirely self-evident to me, once we enter the mathematical realm in which we accept the truth of our statistical model.

Definition 2.3 (Transformation Principle, TP)**.** Let $\mathcal{E} = \{\mathcal{X}, \Omega, f\}$. Let $g : \mathcal{X} \to \mathcal{Y}$ be bijective, and let $\mathcal{E}^g$ be the same experiment as $\mathcal{E}$ but expressed in terms of $Y = g(X)$, rather than X. Then $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}^g, g(x))$.

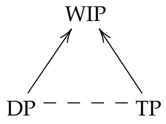
This principle states that inferences should not depend on the way in which the sample space is labelled, which also seems self-evident to me; at least, to violate this principle would be bizarre. But now we have the following result (Basu, 1975; Dawid, 1977).

Theorem 2.1. *($DP \wedge TP$) $\to$ WIP.*

Proof. Fix $\mathcal{E}$, and suppose that $x, x' \in \mathcal{X}$ satisfy $f(x; \bullet) = f(x'; \bullet)$, as in the condition of the WIP. Now consider the transformation $g : \mathcal{X} \to \mathcal{X}$ which switches x for x', but leaves all of the other elements of $\mathcal{X}$ unchanged. In this case $\mathcal{E} = \mathcal{E}^g$. Then

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

which is the WIP. □



So I find, as a matter of logic, I must accept the WIP, or else I must decide which of the two principles DP and TP are, contrary to my initial impression, not self-evident at all. This is the pattern of the next two sections, where either I must accept a principle, or, as a matter of logic, I must reject one of the principles that implies it. From now on, I will treat the WIP as self-evident.

2.3 *The Likelihood Principle*

The new concept in this section is a 'mixture' of two experiments. Suppose I have two experiments,

$$\mathcal{E}_1 = \{\mathcal{X}_1, \Omega, f_1\} \quad \text{and} \quad \mathcal{E}_2 = \{\mathcal{X}_2, \Omega, f_2\},$$

which have the same parameter Θ. Rather than do one experiment or the other, I imagine that I can choose between them randomly, based on known probabilities (p_1, p_2), where $p_2 = 1 - p_1$. The resulting mixture is denoted $\mathcal{E}^* = \{\mathcal{X}^*, \Omega, f^*\}$, where

$$\mathcal{X}^* = (\{1\} \times \mathcal{X}_1) \cup (\{2\} \times \mathcal{X}_2), \tag{2.2a}$$

$$f^*\big((i, x_i); \theta\big) = p_i \cdot f_i(x_i; \theta). \tag{2.2b}$$

$\mathcal{E}^*$ is a mixture experiment.

The famous example of a mixture experiment is the 'two instruments' (see Cox and Hinkley, 1974, sec. 2.3). 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, of course, there is no doubt about which of the two instruments was used. The following principle asserts what must be self-evident to everybody, that inferences should be made according to which instrument was used, and not according to the *a priori* uncertainty. Or, to paraphrase, *don't take into account experiments that were not performed.*

Definition 2.4 (Weak Conditionality Principle, WCP)**.** If $\mathcal{E}^*$ is a mixture experiment, as defined above, then

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

* * *

Another principle does not seem, at first glance, to have anything to do with the WCP. This is the Likelihood Principle.[4]

Definition 2.5 (Likelihood Principle, LP)**.** 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

$$f_1(x_1; \bullet) = c(x_1, x_2) \cdot f_2(x_2; \bullet) \tag{2.3}$$

for some function $c > 0$, then $\text{Ev}(\mathcal{E}_1, x_1) = \text{Ev}(\mathcal{E}_2, x_2)$.

For a given $(\mathcal{E}, x)$, the function $f(x; \bullet)$ is termed the 'likelihood function' for $\theta \in \Omega$. Thus the LP states that if two likelihood functions for the same parameter have the same shape, then the evidence is the same—hence the name. As will be discussed in Section 2.6.3, Frequentist inferences violate the LP. Therefore the following result was something of the bombshell, when it first emerged in the 1960s. The following form is due to Birnbaum (1972) and Basu (1975).[5]

Theorem 2.2 (Birnbaum's Theorem)**.** *(WIP $\wedge$ WCP) $\leftrightarrow$ LP.*

Proof. Both LP $\rightarrow$ WIP and LP $\rightarrow$ WCP are straightforward. The trick is to prove (WIP $\wedge$ WCP) $\rightarrow$ LP. So let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter, and suppose that

[4] The LP is self-attributed to G. Barnard, see his comment to Birnbaum (1962), p. 308. But it is alluded to in the statistical writings of R.A. Fisher, almost appearing in its modern form in Fisher (1956).

[5] Birnbaum's original result (Birnbaum, 1962), used a stronger condition than WIP and a slightly weaker condition than WCP. Theorem 2.2 is clearer.

$x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy $f_2(x_2; \bullet) = c \cdot f_1(x_1; \bullet)$, where $c > 0$ is some constant which may depend on (x_1, x_2), as in the condition of the LP. The value c is known, so consider the mixture experiment with $p_1 = c/(1+c)$ and $p_2 = 1/(1+c)$. Then

$$\begin{aligned} f^*\big((1, x_1); \bullet\big) &= \frac{c}{1+c} \cdot f_1(x_1; \bullet) \\ &= \frac{1}{1+c} \cdot f_2(x_2; \bullet) \\ &= f^*\big((2, x_2); \bullet\big). \end{aligned}$$

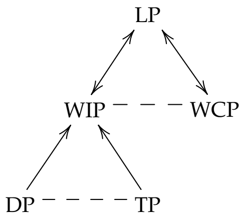
Then the WIP implies that

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

Finally, apply the WCP to each side to infer that

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

which is the LP. □



Again, to be clear about the logic: either I accept the LP, or I explain which of the two principles, WIP and WCP, I refute. To me, the WIP is the implication of two principles that are self-evident, and the WCP is itself self-evident, so I must accept the LP, or else invoke and justify an *ad hoc* abandonment of logic.

A simple way to understand the impact of the LP is to see what it rules out. The following result is used in Section 2.6.3.

Theorem 2.3. *If* Ev *is affected by the allocation of probabilities for outcomes that do not occur, then* Ev *does not satisfy the LP.*

Proof. Because in this case starting from $\mathcal{E} = \{\mathcal{X}, \Omega, f\}$ we could construct another model $\mathcal{E}_1 = \{\mathcal{X}, \Omega, f_1\}$ where $f_1(x; \bullet) = f(x; \bullet)$ but $\operatorname{Ev}(\mathcal{E}_1, x) \neq \operatorname{Ev}(\mathcal{E}, x)$, by manipulating the values of $f_1(x'; \theta)$ for $x' \neq x$. This would violate the LP. □

2.4 *Stronger forms of the Conditionality Principle*

The new concept in this section is 'ancillarity'. This has several different definitions in the Statistics literature; mine is close to that of Cox and Hinkley (1974, sec. 2.2).

Definition 2.6 (Ancillary)**.** X is ancillary in experiment

$$\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega_1 \times \Omega_2, f_{X,Y}\}$$

exactly when $f_{X,Y}$ factorises as

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

In other words, the marginal distribution of X is completely specified. Not all families of distributions will factorise in this way, but when they do, there are new possibilities for inference, based around stronger forms of the WCP, such as the CP immediately below, and the SCP (Definition 2.9).

When X is ancillary, we can consider the conditional experiment

$$\mathcal{E}^{Y|x} = \{\mathcal{Y}, \Omega, f_{Y|x}\}, \tag{2.4}$$

where $f_{Y|x}(y;\theta) := f_{Y|X}(y \mid x;\theta)$. This is an experiment where we condition on $X = x$, i.e. treat X as known, and treat Y as the only random quantity. This is an attractive idea, captured in the following principle.

Definition 2.7 (Conditionality Principle, CP)**.** If X is ancillary in $\mathcal{E}$, then $\mathrm{Ev}\left(\mathcal{E}, (x,y)\right) = \mathrm{Ev}(\mathcal{E}^{Y|x}, y)$.

Clearly the CP implies the WCP, with the experiment indicator $I \in \{1,2\}$ being ancillary, since p is known. It is almost obvious that the CP comes for free with the LP. Another way to put this is that the WIP allows us to 'upgrade' the WCP to the CP.

Theorem 2.4. *LP → CP.*

Proof. Suppose that X is ancillary in $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega, f_{X,Y}\}$. Thus

$$f_{X,Y}(x,y;\bullet) = f_X(x) \cdot f_{Y|X}(y \mid x;\bullet) = c(x) \cdot f_{Y|x}(y;\bullet),$$

where $c > 0$. Then the LP implies that

$$\mathrm{Ev}\left(\mathcal{E}, (x,y)\right) = \mathrm{Ev}(\mathcal{E}^{Y|x}, y),$$

which is the CP. □

I am unsure how useful the CP is in practice. Conditioning on ancillary random quantities is a nice option, but how often do we contemplate an experiment in which X is ancillary? Much more common is the weaker condition that the marginal distribution of X depends on parameters which we are not interested in; such parameters are termed *nuisance parameters*. Hence the following extension.

Definition 2.8 (Auxiliary)**.** X is auxiliary in the experiment

$$\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Psi \times \Omega, f_{X,Y}\}$$

exactly when

$$f_{X,Y}(x,y;\psi,\theta) = f_X(x;\psi) \cdot f_{Y|X}(y \mid x;\theta)$$

and ψ is a nuisance parameter.

In other words, the marginal distribution of X depends on nuisance parameters wich do not occur occur in the conditional distribution of $Y \mid X$. Now this would be a *really* useful principle:

Definition 2.9 (Strong Conditionality Principle, SCP)**.** If X is auxiliary in experiment $\mathcal{E}$, and Ev_θ denotes the evidence about Θ, then $\mathrm{Ev}_\theta\left(\mathcal{E}, (x,y)\right) = \mathrm{Ev}(\mathcal{E}^{Y|x}, y)$.

Here is a example which will be familiar to all statisticians. A regression of Y_i on X_i appears to make a distinction between the 'dependent variable' Y_i and the 'covariates' X_i, with only the former being treated as random. This distinction is insupportable, given that the roles of Y_i and X_i are often interchangeable, and determined by the *hypothèse du jour*. What we are actually doing is asserting that X_i is auxiliary, and then invoking the SCP to treat X_i as known.

Here is another example. We want to find out about the effect of X_i on Y_i. So we set out to collect some values (X_i, Y_i) for a sample of size m, but only $n < m$ people respond. Undoubtedly, the inclination to respond varies in the population, and in a way that depends on (X_i, Y_i). If we choose to simply ignore the non-responders and use the n observations that we have, then what we are actually doing is asserting that (X_i, R_i) is auxiliary in the joint model of (X_i, R_i, Y_i), where $R_i = 0$ for non-response, and $R_i = 1$ for response, and then invoking the SCP to treat (X_i, R_i) as known. Asserting that (X_i, R_i) is auxiliary is a powerful and subtle modelling assumption, termed *non-informative missingness*, and should not be made without careful reflection. *Selection bias* is what happens when this modelling assumption is inappropriate.[6]

There are many other similar examples, to suggest that not only would the SCP be a really useful principle, but in fact it is routinely applied in practice. So it is important to know how the SCP relates to the other principles. The SCP is not deducible from the LP alone. However, it *is* deducibe with an additional and very famous principle, due originally to Savage (1954, sec. 2.7), in a different form.[7]

Definition 2.10 (Sure Thing Principle, STP)**.** Let

$$\mathcal{E}_1 = \{\mathcal{X}_1, \Omega_\theta, f_1\} \quad \text{and} \quad \mathcal{E}_2 = \{\mathcal{X}_2, \Omega_\psi \times \Omega_\theta, f_2\}$$

i.e. where the parameter of $\mathcal{E}_2$ extends that of $\mathcal{E}_1$. Let $\mathrm{Ev}_\theta(\bullet; \psi)$ denote the evidence for Θ, with $\Psi = \psi$. If

$$\mathrm{Ev}_\theta(\mathcal{E}_2, x_2; \psi) = \mathrm{Ev}(\mathcal{E}_1, x_1) \quad \text{for every } \psi,$$

then $\mathrm{Ev}_\theta(\mathcal{E}_2, x_2) = \mathrm{Ev}(\mathcal{E}_1, x_1)$, where Ev_θ is the evidence for Θ.

In words, if our inference about Θ in $\mathcal{E}_2$ were the same as that in $\mathcal{E}_1$ for every possible value of Ψ, then not knowing Ψ would be no impediment to inference about Θ. Most people find this self-evident. This use of the STP to bridge from the CP to the SCP is similar to the Noninformative Nuisance Parameter Principle (NNPP) of Berger and Wolpert (1988, p. 41.5): my point here is that the NNPP is actually the well-known Sure Thing Principle, and does not need a separate name.

Theorem 2.5. $(CP \wedge STP\,) \to SCP.$

Proof. Consider the experiment from Definition 2.8. Treat ψ as known, in which case the parameter is θ, X is ancillary, and the CP

[6] See Harford (2014) for selection bias and Big Data.

[7] See Pearl (2016) for an interesting take on the STP.

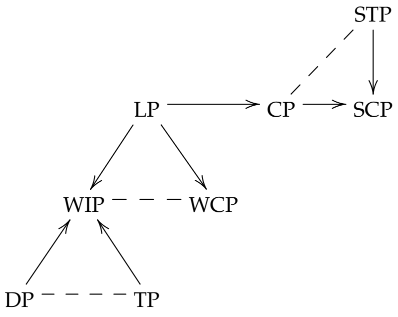
asserts that

$$\mathrm{Ev}_\theta\left(\mathcal{E}, (x,y); \psi\right) = \mathrm{Ev}(\mathcal{E}^{Y|x}, y).$$

As this equality holds for all ψ, the STP implies that

$$\mathrm{Ev}_\theta\left(\mathcal{E}, (x,y)\right) = \mathrm{Ev}(\mathcal{E}^{Y|x}, y),$$

which is the SCP. □



I am happy to accept the STP as self-evident, and since I also accept the LP (which implies the CP), for me to violate the SCP would be illogical. The SCP constrains the way in which I link Ev and Ev_θ.

2.5 *Stopping rules*

Consider a sequence of random quantities $X_1, X_2, \ldots$ with marginal PMFs

$$f_n(x_1, \ldots, x_n; \theta) \qquad n = 1, 2, \ldots,$$

where consistency requires that

$$f_n(x_1, \ldots, x_n; \theta) = \sum_{y_1} \cdots \sum_{y_m} f_{n+m}(x_1, \ldots, x_n, y_1, \ldots y_m; \theta)$$

for each $n, m \in 1, 2, \ldots$.[8] In a sequential experiment, the number of X's that are observed is not fixed in advanced but depends on the values seen so far. That is, at time j, the decision to observe X_{j+1} can be modelled by a probability $p_j(x_1, \ldots, x_j)$. We can assume, resources being finite, that the experiment must stop at specified time m, if it has not stopped already, hence $p_m(x_1, \ldots, x_m) = 0$. Denote the stopping rule as $\tau := (p_1, \ldots, p_m)$.

Definition 2.11 (Stopping Rule Principle, SRP)**.** In a sequential experiment $\mathcal{E}^\tau$, $\mathrm{Ev}\left(\mathcal{E}^\tau, (x_1, \ldots, x_n)\right)$ does not depend on the stopping rule τ.

The SRP is nothing short of revolutionary, if it is accepted. It implies that 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 available. Thus 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. Obviously it would be liberating for the statistician to put aside the experimenter's intentions (since they may not be known and could be highly subjective), but can the SRP possibly be justified? Indeed it can.

Theorem 2.6. *LP → SRP.*

Proof. Let τ be an arbitrary stopping rule, and consider the outcome $(x_1, \ldots, x_n)$, which I will write as $x_{1:n}$ for convenience. The

[8] This is Kolmogorov's consistency condition.

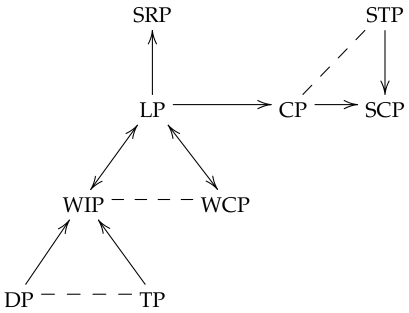
probability of this outcome under τ is

$$\begin{aligned} & f_\tau(x_{1:n};\theta) \\ & \quad = f_1(x_1;\theta)\cdot\prod_{j=1}^{n-1} p_j(x_{1:j})\, f_{j+1}(x_{j+1} \mid x_{1:j};\theta)\cdot\big(1-p_n(x_{1:n})\big) \\ & \quad = \prod_{j=1}^{n-1} p_j(x_{1:j})\cdot\big(1-p_n(x_{1:n})\big)\times f_1(x_1;\theta)\prod_{j=2}^{n} f_j(x_j \mid x_{1:(j-1)};\theta) \\ & \quad = \prod_{j=1}^{n-1} p_j(x_{1:j})\cdot\big(1-p_n(x_{1:n})\big)\times f_n(x_{1:n};\theta). \end{aligned}$$

Now observe that this equation has the form

$$f_\tau(x_{1:n};\bullet) = c(x_{1:n})\cdot f_n(x_{1:n};\bullet) \qquad c>0. \tag{†}$$

Thus the LP implies that $\mathrm{Ev}(\mathcal{E}^\tau, x_{1:n}) = \mathrm{Ev}(\mathcal{E}^n, x_{1:n})$ where $\mathcal{E}^n := \{\mathcal{X}^n, \Omega, f_n\}$. Since the choice of stopping rule was arbitrary, (†) holds for all stopping rules, showing that the choice of stopping rule is irrelevant. □



I think this is one of the most beautiful results in the whole of Theoretical Statistics.

To illustrate the SRP, consider the following example from Basu (1975, p. 42). Four different coin-tossing experiments (with some finite limit on the number of tosses) have the same outcome $x = (\mathrm{T,H,T,T,H,H,T,H,H,H})$:

$\mathcal{E}_1$ 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;

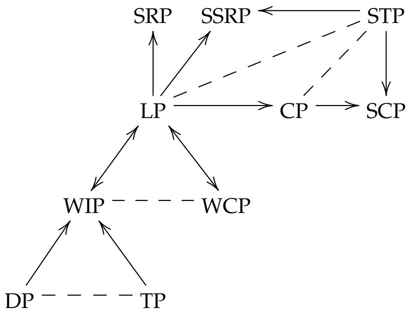
$\mathcal{E}_4$ Continue tossing until the accumulated number of heads exceeds that of tails by exactly 2.

One could easily adduce more sequential experiments which gave the same outcome. According to the SRP, the evidence for 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, and the simplest experiment $\mathcal{E}_1$ can be used for inference.

The SRP can be strengthened to stopping rules which are *unknown* stochastic functions of $(x_1,\dots,x_j)$, as long as the true value of the parameter ψ in $p_j(x_1,\dots,x_j;\psi)$ is unrelated to the true value Θ. This is the *Strong Stopping Rule Principle (SSRP)*.

Theorem 2.7. *(LP $\wedge$ STP) $\to$ SSRP.*

Proof. Repeat the previous proof with a '$;\psi$' inside p_j. Then use the STP to ignore the presence of ψ in c. □



In the absence of any information about the experimenter's intentions, the SSRP is the principle that needs to be invoked.

* * *

The Stopping Rule Principle has become enshrined in our profession's collective memory due to this iconic comment from L.J. Savage, one of the great statisticians of the 20th century:

> 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 et al., 1962, p. 76)

This comment captures the revolutionary and transformative nature of the SRP.

2.6 The Likelihood Principle in practice

Now we should pause for breath, and ask the obvious questions: is the LP vacuuous? Or trivial? In other words, Is there any inferential approach which respects it? Or do all inferential approaches respect it? In this section I consider three approaches: likelihood-based inference, Bayesian inference, and Frequentist inference. The first two satisfy the LP, and the third does not. I also show that the first two also satisfy the SCP, which is the best possible result for conditioning on ancillary random quantities, side-stepping nuisance parameters, and ignoring stopping rules.

2.6.1 Likelihood-based inference (LBI)

The evidence from $(\mathcal{E}, x)$ can be summarised in the *likelihood function*:

$$L : \theta \mapsto f(x; \theta). \tag{2.5}$$

A small but influential group of statisticians have advocated that evidence is not merely summarised by L, but is actually derived entirely from the shape of L; see, for example, Hacking (1965), Edwards (1992), Royall (1997), and Pawitan (2001). Hence:

Definition 2.12 (Likelihood-based inference, LBI)**.** Let $\mathcal{E}$ be an experiment with outcome x. Under LBI,

$$\text{Ev}(\mathcal{E}, x) = \mathcal{I}(L)$$

for some operator $\mathcal{I}$ depending on Ev, for which $\mathcal{I}(L) = \mathcal{I}(c(x) \cdot L)$ for every $c > 0$.

The invariance of $\mathcal{I}$ to c shows that only the shape of L matters: its scale does not matter at all.

The main operators for LBI are the *Maximum Likelihood Estimator (MLE)*

$$\hat{\theta} = \operatorname*{argsup}_{\theta \in \Omega} L(\theta) \tag{2.6}$$

for point estimation, and *Wilks level sets*

$$\widehat{C}_k = \{\theta \in \Omega : \log L(\theta) \geq \log L(\hat{\theta}) - k\} \tag{2.7}$$

for set estimation and hypothesis testing, where k may depend on x. Wilks level sets have the interesting and reassuring property that they are invariant to bijective transformations of the parameter.[9]

Both of these operators satisfy $\mathcal{I}(L) = \mathcal{I}(c \cdot L)$. However, they are not without their difficulties: the MLE is sometimes undefined and often ill-behaved (see, e.g., Le Cam, 1990), and it is far from clear which level set is appropriate, and how this might depend on the dimension of Ω (i.e. how to choose k in eq. 2.7).

LBI satisfies the LP by construction, so it also satisfies the CP. To see whether it satisfies the SCP requires a definition of Ev_θ, the evidence for Θ in the case where the parameter is (ψ, θ) and ψ is a nuisance parameter. The standard definition is based on the *profile likelihood*,

$$L_\theta : \theta \mapsto \sup_\psi L(\psi, \theta), \tag{2.8}$$

from which

$$\mathrm{Ev}_\theta(\mathcal{E}, x) := \mathcal{I}(L_\theta). \tag{2.9}$$

Then we have the following result.

Theorem 2.8. *If profile likelihood is used for* Ev_θ, *then LBI satisfies the SCP.*

Proof. Under the conditions of Definition 2.9 we have, putting '•' where the θ argument goes,

$$\begin{aligned}
\mathrm{Ev}_2\{\mathcal{E}, (x,y)\} &= \mathcal{I}\{\sup_\psi L(\psi, \bullet)\} & \\
&= \mathcal{I}\{\sup_\psi f_X(x;\psi) \cdot f_{Y|X}(y \mid x; \bullet)\} & X \text{ is auxiliary} \\
&= \mathcal{I}\{c(x) \cdot f_{Y|X}(y \mid x; \bullet)\} & \text{where } c > 0 \\
&= \mathcal{I}\{f_{Y|X}(y \mid x; \bullet)\} & \text{property of } \mathcal{I} \\
&= \mathrm{Ev}(\mathcal{E}^{Y|x}, y),
\end{aligned}$$

where $\mathcal{E}^{Y|x}$ was defined in (2.4). □

Therefore, LBI satisfies the SCP and the strong version of the SRP, which is the best possible outcome. But another *caveat*: profile likelihood inherits all of the same difficulties as Maximum Likelihood, and some additional ones as well. LBI has attractive theoretical properties but unattractive practical ones, and for this reason it has been more favoured by philosophers and physicists than by practising statisticians.

2.6.2 *The Bayesian approach*

The Bayesian approach for inference was outlined in Section 1.5. The Bayesian approach augments the experiment $\mathcal{E} := \{\mathcal{X}, \Omega, f\}$

[9] It is insightful to formalize this notion, and prove it.

with a prior probability distribution π on Ω, representing initial beliefs about Θ. The *posterior distribution* for Θ is found by conditioning on the outcome x, to give

$$\pi^*(\theta) \propto f(x;\theta) \cdot \pi(\theta) = L(\theta) \cdot \pi(\theta) \tag{2.10}$$

where L is the Likelihood Function from Section 2.6.1. The missing multiplicative constant can be inferred, if it is required, from the normalisation condition $\int_\Omega \pi^*(\theta)\,\mathrm{d}\theta = 1$. By Bayes's Theorem, it is $1/\,\mathrm{p}(x)$. Usually, Ω is uncountable, and π and π^* are probability density functions (PDFs).

Bayesian statisticians follow exactly one principle.

Definition 2.13 (Bayesian Conditionalization Principle, BCP)**.** Let $\mathcal{E}$ be an experiment with outcome x. Under the BCP

$$\mathrm{Ev}(\mathcal{E}, x) = \mathcal{I}(\pi^*)$$

for some operator $\mathcal{I}$ depending on Ev.

There is a wealth of operators for Bayesian inference. A common one for a point estimator is the *Maxium A Posteriori (MAP)* estimator

$$\hat{\theta}^* = \operatorname*{argsup}_{\theta \in \Omega} \pi^*(\theta). \tag{2.11}$$

The MAP estimator does not require the calculation of the multiplicative constant $1/\,\mathrm{p}(x)$. In a crude sense, it improves on the MLE from Section 2.6.1 by using the prior distribution π to 'regularize' the likelihood function, by downweighting less realistic values. This is the point of view taken in *inverse problems*, where Θ is the signal, x is a set of measurements, f represents the 'forward model' from the signal to the measurements, and π represents beliefs about regularities in Θ. Inverse problems occur throughout science, and this Bayesian approach is ubiquitous where the signal has inherent structure (e.g., the weather, or an image).

A common operator for a Bayesian set estimator is the *High Posterior Density (HPD) region*

$$C_k^* := \big\{\theta \in \Omega : \log \pi^*(\theta) \geq k\big\}. \tag{2.12}$$

The value k is usually set according to the probability content of C_k^*. A level-95% HPD will have k which satisfies

$$\int_{C_k^*} \pi^*(\theta)\,\mathrm{d}\theta = 0.95. \tag{2.13}$$

In contrast to the Wilks level sets in Section 2.6.1, the Bayesian approach 'solves' the problem of how to choose k. HPD regions are not transformation invariant. Instead, an HPD region is the smallest set which contains exactly 95% of the posterior probability. Alternatively, the 'snug' region $\widehat{C}_k$ satisfying $\int_{\widehat{C}_k} \pi^*(\theta)\,\mathrm{d}\theta = 0.95$ *is* transformation-invariant, but it is typically not the smallest set estimator which contains exactly 95% of the posterior probability.[10]

[10] I came across 'snug' regions in the Cambridge lecture notes of Prof. Philip Dawid.

The two estimators often give similar results, for well-understood theoretical reasons (see, e.g., van der Vaart, 1998).

It is straightforward to establish that Bayesian inference satisfies the LP.

Proof. Let $\mathcal{E}_1 := \{\mathcal{X}_1, \Omega, f_1\}$ and $\mathcal{E}_2 := \{\mathcal{X}_2, \Omega, f_2\}$ be two experiments with the same parameter. Because this parameter is the same, the prior distribution is the same; denote it π. Let x_1 and x_2 be two outcomes satisfying $L_1 = c \cdot L_2$, which is the condition of the LP, where L_1 is the likelihood function for $(\mathcal{E}_1, x_1)$, L_2 is the likelihood function for $(\mathcal{E}_2, x_2)$, and $c > 0$ may depend on (x_1, x_2). Then

$$\begin{aligned}
\mathrm{Ev}(\mathcal{E}_1, x_1) &= \mathcal{I}(\pi_1^*) \\
&= \mathcal{I}\left(\frac{L_1(\bullet) \cdot \pi(\bullet)}{\int L_1(\theta) \cdot \pi(\theta)\, \mathrm{d}\theta}\right) \\
&= \mathcal{I}\left(\frac{c \cdot L_1(\bullet) \cdot \pi(\bullet)}{c \cdot \int L_1(\theta) \cdot \pi(\theta)\, \mathrm{d}\theta}\right) \\
&= \mathcal{I}\left(\frac{L_2(\bullet) \cdot \pi(\bullet)}{\int L_2(\theta) \cdot \pi(\theta)\, \mathrm{d}\theta}\right) \\
&= \mathcal{I}(\pi_2^*) = \mathrm{Ev}(\mathcal{E}_2, x_2).
\end{aligned}$$

□

Hence BCP also satisfies the CP. What about the SCP in the case where the parameter is (ψ, θ), ψ is a nuisance parameter, and X is auxiliary? As for LBI in Section 2.6.1, this requires a definition of Ev_θ. In the Bayesian approach there is only one choice, based on the marginal posterior distribution

$$\pi_\theta^* := \theta \mapsto \int_\psi \pi^*(\psi, \theta)\, \mathrm{d}\psi, \tag{2.14}$$

from which

$$\mathrm{Ev}_\theta(\mathcal{E}, x) = \mathcal{I}(\pi_\theta^*). \tag{2.15}$$

Then we have the following result.

Theorem 2.9. *If $\pi(\psi, \theta) = \pi_1(\psi) \cdot \pi_2(\theta)$, then Bayesian inference satisfies the SCP.*

Proof. Under the conditions of Definition 2.9 and the theorem, the posterior distribution satisfies

$$\begin{aligned}
\pi^*(\psi, \theta) &= \frac{f_X(x; \psi) \cdot f_{Y|X}(y \mid x; \theta) \times \pi_1(\psi) \cdot \pi_2(\theta)}{\mathrm{p}(x, y)} \\
&= \frac{f_X(x; \psi) \cdot f_{Y|X}(y \mid x; \theta) \times \pi_1(\psi) \cdot \pi_2(\theta)}{\mathrm{p}(x) \cdot \mathrm{p}(y \mid x)} \\
&= \frac{f_{Y|X}(y \mid x; \theta) \cdot \pi_2(\theta)}{\mathrm{p}(y \mid x)} \times \pi_1^*(\psi),
\end{aligned}$$

where π_1^* is the conditional distribution of Ψ given x. Integrating out ψ then gives

$$\pi_\theta^*(\bullet) = \int_\psi \pi^*(\psi, \bullet)\, \mathrm{d}\psi = \frac{f_{Y|X}(y \mid x; \bullet) \cdot \pi_2(\bullet)}{\mathrm{p}(y \mid x)} =: \pi_{Y|x}^*(\bullet),$$

which is the posterior distribution for θ treating X as given. Thus

$$\begin{aligned}\mathrm{Ev}_\theta\left(\mathcal{E}, (x, y)\right) &= \mathcal{I}\{\pi^*_\theta(\bullet)\} \\ &= \mathcal{I}\{\pi^*_{Y|x}(\bullet))\} \\ &= \mathrm{Ev}(\mathcal{E}^{Y|x}, y). \qquad \square\end{aligned}$$

Therefore, under the mild condition that the prior distribution factorizes, Bayesian inference satisfies the SCP and the strong version of the SRP, which is the best possible outcome.

However ... Bayesian practice is heterogeneous. Two issues are pertinent. First, the Bayesian statistician does not just magic up a model f and a prior distribution π. Instead, she iterates through some different possibilities, modifying her choices using the observations. The decision to replace a model or a prior distribution may depend on probabilities of outcomes which did not occur (see the end of Section 2.3). But this practice *does not* violate the LP, which is about what happens while accepting the model and the prior as true. Statisticians are immune from this criticism while 'inside' their statistical inference. But applied statisticians are obliged to continue the stages in Section 1.1, in order to demonstrate the relevance of their mathematical solution for the real-world problem.

Second, the Bayesian statistician faces the additional challenge of providing a prior distribution. In principle, this prior reflects beliefs about Θ that exist independently of the outcome, and can be an opportunity rather than a threat. In practice, though, is hard to do. Some methods for making default choices for π depend on f_X, notably Jeffreys priors and reference priors (see, e.g., Bernardo and Smith, 2000, sec. 5.4). These methods violate the LP.

2.6.3 Frequentist inference

LBI and Bayesian inference both have simple representations in terms of an operator $\mathcal{I}$. Frequentist inference adopts a different approach, described in Section 1.4, notably Definition 1.1. In a nutshell, algorithms are certified in terms of their sampling distributions, and selected on the basis of their certification. Theorem 2.3 shows that Frequentist inference does not respect the LP, because the sampling distribution of the algorithm depends on values for f other than $f(x; \bullet)$.

The following comments apply to all approaches which violate the LP, including the Bayesian approach using a reference prior distribution. However, I will focus on the Frequentist approach because it is fundamentally opposed to the LP. The two main difficulties are:

1. To reject the LP is to reject at least one of the WIP and the WCP. Yet both of these principles seem self-evident. Therefore the Frequentist statistician is either illogical or obtuse.

2. In their everyday practice, Frequentist statisticians use the (S)CP and the (S)SRP, which are not self-evident, and whose simplest

justfication is via the LP and the STP. To deny the LP requires a different justification.

Alternative justifications for the (S)CP and the (S)SRP have not been forthcoming.

2.7 *Reflections*

The statistician takes delivery of an outcome x. Her standard practice, as mandated by our profession, is to assumes the truth of a statistical model $\mathcal{E}$, and then turn $(\mathcal{E}, x)$ into an inference about the true value of the parameter Θ. As remarked several times already (see Chapter 1), this is *not* the end of her involvement, but it is a key step, which may be repeated several times, under different notions of the outcome and different statistical models. This chapter concerns this key step: how she turns $(\mathcal{E}, x)$ into an inference about Θ.

Whatever inference is required, we assume that the statistician applies an algorithm to $(\mathcal{E}, x)$. In other words, her inference about Θ is not arbitrary, but transparent and reproducible—this is hardly controversial, because anything else would be non-scientific. Following Birnbaum, the algorithm is denoted 'Ev'. The question now becomes: how does she choose her 'Ev'?

This chapter does not explain how to choose 'Ev'; instead it describes some properties that 'Ev' might have. Some of these properties are self-evident, and to violate them would be hard to justify to an auditor. These properties are the DP (Definition 2.2), TP (Definition 2.3), WCP (Definition 2.4), and STP (Definition 2.10). Other properties are not at all self-evident; the most important of these are the LP (Definition 2.5), the SCP (Definition 2.9), and the SSRP (after Definition 2.11). These properties would be extremely convenient, were it possible to justify them. And it turns out that they can all be justified as logical deductions from the properties that are self-evident. This is the essence of Birnbaum's Theorem (Theorem 2.2).

For over a century, statisticians have been proposing methods for selecting algorithms for 'Ev', independently of this strand of research concerning the properties that such algorithms ought to have (remember that Birbaum's Theorem was published in 1962). Crudely, we can label these as 'likelihood-based inference' (LBI, Section 2.6.1), Bayesian inference (Section 2.6.2), and Frequentist inference (Section 2.6.3). The first and second of these approaches are compatible with all of the properties given above, but the third, Frequentist inference, is not. In other words, the practice of certifying every algorithm according to its sampling distribution, and then selecting an algorithm according to its certificate, violates the LP. The two main consequences of this violation are described in Section 2.6.3.

Now it is important to be clear about one thing. Ultimately, an inference is a single element in the space of 'possible inferences about Θ'. An inference cannot be evaluated according to whether or not it satsfies the LP. What is being evaluated in this chapter is the algorithm, the mechanism by which $\mathcal{E}$ and x are turned into an inference. It is quite possible that statisticians of quite different persuasions will produce effectively identical inferences from different algorithms. For example, if asked for a set estimate of Θ, a Bayesian statistician might produce a 95% High Density Region, and a Frequentist statistician a 95% confidence set, but they might be effectively the same set. But it is not the inference that is the primary concern of the auditor: it is the justification for the inference, among the uncountable other inferences that might have been made but weren't. The auditor checks the 'why', before passing the 'what' onto the client.

So the auditor will ask: why do you choose algorithm 'Ev'? The Frequentist statistician will reply, "Because it is a 95% confidence procedure for Θ, and, among the uncountable number of such procedures, this is a good choice [for some reasons that are then given]." The Bayesian statistician will reply "Because it is a 95% High Posterior Density region for Θ for prior distribution π, and among the uncountable number of prior distributions, π is a good choice [for some reasons that are then given]." Let's assume that the reasons are compelling, in both cases. The auditor has a follow-up question for the Frequentist but not for the Bayesian: "Why are you not concerned about violating the Likelihood Principle?" A well-informed auditor will know the theory of the previous sections, and the consequences of violating the LP that are given in Section 2.6.3. For example, violating the LP is either illogical or obtuse—neither of these properties are desirable in an applied statistician.

To be frank I do not have a good answer to this question, which is why I would choose *not* to violate the LP, in the way that I choose 'Ev'. However, in the spirit of fair play I will suggest two possibilities.[11]

First, the Frequentist might reply, "Because this is how we do things in (say) Experimental Psychology", i.e. an appeal to current practice. This answer is contrary to the scientific norm of scepticism, and may upset the client, who thought he was paying for a scientist. The counter-argument is that 'science is what scientists do', which is a naturalistic as opposed to normative view of science (see, e.g., Ziman, 2000). Under the naturalistic view, violating the LP is scientific as long as it is the standard practice among the *soi-disant* scientists in Experimental Psychology. Personally, I don't think this excuses these scientists from having a compelling reason for violating the LP (e.g., explaining why they are neither illogical nor obtuse). But apparently most Experimental Psychologists disagree with me, or else they are ignorant of the LP and its implications.

Second, the Frequentist might reply "Because it is important to me that I control my error rate over the course of my career", which

[11] Another possibility to add to these two might be "I'm not interested in principles, I let the data speak for itself." This person would suit a client who wanted an illogical and unprincipled data analyst; or "reckless and treacherous", according to Alfred Marshall, writing in 1885 (Stigler, 2016, p. 202). If you are this person, you can probably charge a lot of money.

is incompatible with the LP. In other words, the statistician ensures that, by always using a 95% confidence procedure, the true value of Θ will be inside at least 95% of her confidence sets, over her career. This is a very interesting answer, revealing the statistician's egocentricity in putting her career error rate before the needs of her current client. I can just about imagine a client demanding "I want a statistician who is right at least 95% of the time". Personally, though, I would advise a client against this, and favour instead a statistician who is concerned not with her career error rate, but rather with the client's particular problem.

5
Bibliography

Bartlett, M. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44:533–534. 61

Basu, D. (1975). Statistical information and likelihood. *Sankhyā*, 37(1):1–71. With discussion. 14, 15, 16, 21

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York, Inc., NY, USA, second edition. 34

Berger, J. and Boos, D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89:1012–1016. 53

Berger, J. and Wolpert, R. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward CA, USA, second edition. Available online, `http://projecteuclid.org/euclid.lnms/1215466210`. 14, 19

Bernardo, J. and Smith, A. (2000). *Bayesian Theory*. John Wiley & Sons Ltd, Chichester, UK. (paperback edition, first published 1994). 26

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269–306. 13, 16

Birnbaum, A. (1972). More concepts of statistical evidence. *Journal of the American Statistical Association*, 67:858–861. 14, 16

Casella, G. and Berger, R. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition. 1, 3, 50

Çınlar, E. and Vanderbei, R. (2013). *Real and Convex Analysis*. Springer, New York NY, USA. 36

Cormen, T., Leiserson, C., and Rivest, R. (1990). *Introduction to Algorithms*. The MIT Press, Cambridge, MA. 10

Cowles, M. and Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37(5):553–558. 55

Cox, D. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge, UK. 1, 50

Cox, D. and Donnelly, C. (2011). *Principles of Applied Statistics*. Cambridge University Press, Cambridge, UK. 1

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London, UK. 16, 17, 35, 46

Davison, A. (2003). *Statistical Models*. Cambridge University Press, Cambridge, UK. 3

Dawid, A. (1977). Conformity of inference patterns. In Barra, J. et al., editors, *Recent Developments in Statistcs*. North-Holland Publishing Company, Amsterdam. 14, 15

DiCiccio, T. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3):189–212. with discussion and rejoinder, 212–228. 51

Draper, N. and Smith, H. (1998). *Applied Regression Analysis*. New York: John Wiley & Sons, 3rd edition. 48

Edwards, A. (1992). *Likelihood*. The Johns Hopkins University Press, Baltimore, USA, expanded edition. 22

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119–127. Available at `http://statweb.stanford.edu/~ckirby/brad/other/Article1977.pdf`. 38

Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Edinburgh and London: Oliver and Boyd. 16, 55

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton FL, USA, 3rd edition. Online resources at `http://www.stat.columbia.edu/~gelman/book/`. 9

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London, UK. 34

Greenland, S. and Poole, C. (2013). Living with *P* values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, 24(1):62–68. With discussion and rejoinder, pp. 69–78. 56

Hacking, I. (1965). *The Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK. 22

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge University Press, Cambridge, UK. 1

Hacking, I. (2014). *Why is there a Philosophy of Mathematics at all?* Cambridge University Press, Cambridge, UK. 2

Harford, T. (2014). Big data: Are we making a big mistake? *Financial Times Magazine*. Published online Mar 28, 2014. Available at `http://on.ft.com/P0PVBF`. 11, 19

Lad, F. (1996). *Operational Subjective Statistical Methods*. New York: John Wiley & Sons. 2

Le Cam, L. (1990). Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171. 5, 23

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192. See also Bartlett (1957). 56

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The BUGS Book: A Practical introduction to Bayesian Analysis*. CRC Press, Boca Raton FL, USA. 9

MacKay, D. (2009). *Sustainable Energy – Without the Hot Air*. UIT Cambridge Ltd, Cambridge, UK. available online, at `http://www.withouthotair.com/`. 2

Madigan, D., Strang, P., Berlin, J., Schuemie, M., Overhage, J., Suchard, M., Dumouchel, B., Hartzema, A., and Ryan, P. (2014). A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1:11–39. 8

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Harcourt Brace & Co., London, UK. 47, 49

Morey, R., Hoekstra, R., Rouder, J., Lee, M., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bullentin & Review*, 23(1):103–123. 44

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. New York: Springer, 2nd edition. 4

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon Press. 22

Pearl, J. (2016). The Sure-Thing Principle. *Journal of Causal Inference*, 4(1):81–86. 19

Rougier, J., Sparks, R., and Cashman, K. (2016). Global recording rates for large eruptions. *Journal of Applied Volcanology*, forthcoming. 51

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall/CRC Press, Boca Raton FL, USA. 22

Samworth, R. (2012). Stein's paradox. *Eureka*, 62:38–41. Available online at `http://www.statslab.cam.ac.uk/~rjs57/SteinParadox.pdf`. Careful readers will spot a typo in the maths. 38

Savage, L. (1954). *The Foundations of Statistics*. Dover, New York, revised 1972 edition. 19

Savage, L. et al. (1962). *The Foundations of Statistical Inference*. Methuen, London, UK. 1, 22

Schervish, M. (1995). *Theory of Statistics*. Springer, New York NY, USA. Corrected 2nd printing, 1997. 1, 5, 9, 34

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64(4):583–616. With discussion, pp. 616–639. 9

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, 76(3):485–493. 9

Stigler, S. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge MA, USA. 1, 3, 28

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK. 25, 50

Wood, S. (2015). *Core Statistics*. Cambridge University Press, Cambridge, UK. 47

Ziman, J. (2000). *Real Science: What it is, and what it means*. Cambridge University Press, Cambridge, UK. 28