Jonathan Rougier

Department of Mathematics
University of Bristol

APTS Lecture notes on Statistical Inference

Our mission: To help people make better choices under uncertainty.

Version 5, compiled on December 5, 2016.

1
Statistics: another short introduction

From *APTS Lecture Notes on Statistical Inference*, Jonathan Rougier, Copyright © University of Bristol 2016.

In Statistics we quantify our beliefs about things which we would like to know in the light of other things which we have measured, or will measure. This programme is not unique to Statistics: one distinguishing feature of Statistics is the use of *probability* to quantify the uncertainty in our beliefs. Within Statistics we tend to separate Theoretical Statistics, which is the study of algorithms and their properties, from Applied Statistics, which is the use of carefully-selected algorithms to quantify beliefs about the real world. This chapter is about Theoretical Statistics.

If I had to recommend one introductory book about Theoretical Statistics, it would be Hacking (2001). The two textbooks I find myself using most regularly are Casella and Berger (2002) and Schervish (1995). For travelling, Cox (2006) and Cox and Donnelly (2011) are slim and full of insights. If you can find it, Savage et al. (1962) is a short and gripping account of the state of Statistics at a critical transition, in the late 1950s and early 1960s.[1]

1.1 Statistical models

This section covers the nature of a statistical model, and some of the basic conventions for notation.

A *statistical model* is an artefact to link our beliefs about things which we can measure to things we would like to know. Denote the values of the things we can measure as Y, and the values of the things we would like to know as X. These are *random quantities*, indicating that their values, ahead of taking the measurements, are unknown to us.

The convention in Statistics is that random quantities are denoted with capital letters, and particular values of those random quantities with small letters; e.g., x is a particular value that X could take. This sometimes clashes with another convention that matrices are shown with capital letters and scalars with small letters. A partial resolution is to use normal letters for scalars, and bold-face letters for vectors and matrices. However, I have stopped adhering to this convention, as it it usually clear what X is from the context. Therefore both X and Y may be collections of quantities.

I term the set of possible (numerical) values for X the *realm* of

[1] And contains the funniest sentence ever written in Statistics, contributed by L.J. Savage.

X, after Lad (1996), and denote it $\mathcal{X}$. This illustrates another convention, common throughout Mathematics, that sets are denoted with ornate letters. The realm of (X, Y) is denoted $\mathcal{X} \times \mathcal{Y}$. Where the realm is a product, then the margins are denoted with subscripts. So if $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, then $Z_1 = X$ and $Z_2 = Y$. The most common example is where $X = (X_1, \dots, X_m)$, and the realm of each X_i is $\mathcal{X}$, so that the realm of X is $\mathcal{X}^m$.

In the definition of a statistical model, 'artefact' denotes an object made by a human, e.g. you or me. There are no statistical models that don't originate inside our minds. So there is no arbiter to determine the 'true' statistical model for (X, Y)—we may expect to disagree about the statistical model for (X, Y), between ourselves, and even within ourselves from one time-point to another.[2] In common with all other scientists, statisticians do not require their models to be true. Statistical models exist to make prediction feasible (see Section 1.3).

Maybe it would be helpful to say a little more about this. Here is the usual procedure in 'public' Science, sanitised and compressed:

1. Given an interesting question, formulate it as a problem with a solution.

2. Using experience, imagination, and technical skill, make some simplifying assumptions to move the problem into the mathematical domain, and solve it.

3. Contemplate the simplified solution in the light of the assumptions, e.g. in terms of robustness. Maybe iterate a few times.

4. Publish your simplified solution (including, of course, all of your assumptions), and your recommendation for the original question, if you have one. Prepare for criticism.

MacKay (2009) provides a masterclass in this procedure.[3] The statistical model represents a statistician's 'simplifying assumptions'.

A statistical model takes the form of a *family of probability distributions* over $\mathcal{X} \times \mathcal{Y}$. I will assume, for notational convenience, that $\mathcal{X} \times \mathcal{Y}$ is countable.[4] Dropping Y for a moment, let $\mathcal{X} = \{x^{(1)}, x^{(2)}, \dots\}$. The complete set of probability distributions for X is

$$\mathcal{P} = \left\{ p \in \mathbb{R}^k : \forall i \; p_i \geq 0, \sum_{i=1}^{k} p_i = 1 \right\}, \tag{1.1}$$

where $p_i = \Pr(X = x^{(i)})$, and $k = |\mathcal{X}|$, the number of elements of $\mathcal{X}$. A family of distributions is a subset of $\mathcal{P}$, say $\mathcal{F}$. In other words, a statistician creates a statistical model by ruling out many possible probability distributions. The family is usually denoted by a *probability mass function (PMF)* f_X, a *parameter* θ, and a *parameter space* Ω, such that

$$\mathcal{F} = \left\{ p \in \mathcal{P} : \forall i \; p_i = f_X(x^{(i)}; \theta) \text{ for some } \theta \in \Omega \right\}. \tag{1.2}$$

[2] Some people refer to the unknown *data generating process (DGP)* for (X, Y), but I have never found this to be a useful concept.

[3] Many people have discussed the "unreasonable effectiveness of mathematics", to use the phrase of Eugene Wigner; see `https://en.wikipedia.org/wiki/The_Unreasonable_Effectiveness_of_Mathematics_in_the_Natural_Sciences`. Or, for a more nuanced view, Hacking (2014).

[4] Everything in this chapter generalizes to the case where the realm is uncountable.

For obvious reasons, we require that if $\theta' \neq \theta''$, then

$$f_X(\cdot\,;\theta') \neq f_X(\cdot\,;\theta''); \qquad (1.3)$$

such models are termed *identifiable*.[5] Taken all together, it is convenient to denote a statistical model for X as the triple

$$\mathcal{E} = \{\mathcal{X}, \Omega, f_X\}. \qquad (1.4)$$

I will occasionally distinguish between the family $\mathcal{F}$ and the statistical model $\mathcal{E}$. This is because the model is just one of uncountably many different instantiations of the same family. That is to say, two statisticians may agree on the family $\mathcal{F}$, but choose different models $\mathcal{E}_1$ and $\mathcal{E}_2$.[6]

Most statistical procedures start with the specification of a statistical model for (X, Y),

$$\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega, f_{X,Y}\}. \qquad (1.5)$$

The method by which a statistician chooses $\mathcal{F}$ and then $\mathcal{E}$ is hard to codify, although experience and precedent are obviously relevant. See Davison (2003) for a book-length treatment with many useful examples.

1.2 *Hierarchies of models*

The concept of a statistical model was crystalized in the early part of the 20th century. At that time, when the notion of a digital computer was no more than a twinkle in John von Neumann's eye, the 'f_Y' in the model $\{\mathcal{Y}, \Omega, f_Y\}$ was assumed to be a known analytic function of y for each θ.[7] As such, all sorts of other useful operations are possible, such as differentiating with respect to θ. Expressions for the PMFs of specified functions of set of random quantities are also known analytic functions: sums, differences, and more general transformations.

This was computationally convenient—in fact it was critical given the resources of the time—but it severely restricted the models which could be used in practice, more-or-less to the models found today at the back of every textbook in Statistics (e.g. Casella and Berger, 2002), or simple combinations thereof. Since about the 1950s—the start of the computer age—we have had the ability to evaluate a much wider set of functions, and to simulate random quantities on digital computers. As a result, the set of usable statistical models has dramatically increased. In modern Statistics, we now have the freedom to specify the model that most effectively represents our beliefs about the set of random quantities of interest. Therefore we need to update our notion of statistical model,

[5] Some more notation. f_X is a function; formally, $f_X : \mathcal{X} \times \Omega \to [0,1]$. Two functions can be compared for equality: as functions are sets of tuples, the comparison is for the equality of two sets. $f_X(\cdot\,;\theta)$ is also a function, $f_X(\cdot\,;\theta) : \mathcal{X} \to [0,1]$ but different for each value of θ. It is a convention in Statistics to separate the argument x from the parameter θ using a semicolon.

[6] Some algorithms, such as the MLE (see eq. 1.6), are model-invariant in the sense that their results translate from one model to another, within the same family. But many are not. It's a moot question whether we should value algorithms that are model-invariant. My feeling is that we should, but the topic does not get a lot of attention in textbooks.

[7] That is, a function which can be evaluated to any specified precision using a finite number of operations, like the Poisson PMF or the Normal probability density function (PDF).

according to the following hierarchy.

A. Models where f_Y has a known analytic form.

B. Models where $f_Y(y;\theta)$ can be evaluated.

C. Models where Y can be simulated from $f_Y(\cdot;\theta)$.

Between (B) and (C) exist models where $f_Y(y;\theta)$ can be evaluated up to an unknown constant, which may or may not depend on θ.

To illustrate the difference, consider the Maximum Likelihood Estimator (MLE) of the 'true' value of θ based on Y, defined as

$$\hat{\theta}(y) := \sup_{\theta \in \Omega} f_Y(y;\theta). \tag{1.6}$$

Eq. (1.6) is just a sequence of mathematical symbols, waiting to be instantiated into an algorithm. If f_Y has a known analytic form, i.e. level (A) of the hierarchy, then it may be possible to solve the first-order conditions,[8]

$$\frac{\partial}{\partial \theta} f_Y(y;\theta) = 0, \tag{1.7}$$

uniquely for θ as a function of y (assuming, for simplicity, that Ω is a convex subset of $\mathbb{R}$) and to show that $\frac{\partial^2}{\partial \theta^2} f_Y(y;\theta)$ is negative at this solution. In this case we are able to derive an analytic expression for $\hat{\theta}$. Even if we cannot solve the first order conditions, we might be able to prove that $f_Y(y;\cdot)$ is strictly concave, so that we know there is a unique maximum. This means that any numerical maximization of $f_Y(y;\cdot)$ is guaranteed to converge to $\hat{\theta}(y)$.

But what if we can evaluate $f_Y(y;\theta)$, but do not know its form, i.e. level (B) of the hierarchy? In this case we can still numerically maximize $f_Y(y;\cdot)$, but we cannot be sure that the maximizer will converge to $\hat{\theta}(y)$: it may converge to a local maximum. So the algorithm for finding $\hat{\theta}(y)$ must have some additional procedures to ensure that all local maxima are ignored: this is very complicated in practice, very resource intensive, and there are no guarantees.[9] So in practice the Maximum Likelihood algorithm does not necessarily give the MLE. We must recognise this distinction, and not make claims for the MLE algorithm which we implement, that are based on theoretical properties of the MLE.

And what about level (C) of the hierarchy? It is very tricky indeed to find the MLE in this case, and any algorithm that tries will be very imperfect. Other estimators of θ would usually be preferred. This example illustrates that in Statistics it is the choice of algorithm that matters. The MLE is a good choice only if (i) you can prove that it has good properties for your statistical model,[10] and (ii) you can prove that your algorithm for finding the MLE is in fact guaranteed to find the MLE for your statistical model. If you have used an algorithm to find the MLE without checking both (i) and (ii), then your results bear the same relation to Statistics as Astrology does to Astronomy. Doing Astrology is fine, but not if your client has paid you to do Astronomy.

[8] For simplicity and numerical stability, these would usually be applied to $\log f_Y$ not f_Y.

[9] See, e.g., Nocedal and Wright (2006). Do not be tempted to make up your own numerical maximization algorithm.

[10] Which is often very unclear; see Le Cam (1990).

1.3 Prediction and inference

The task in Applied Statistics is to predict X using y^{obs}, the measured value of Y. It is convenient to term Y the *observables* and y^{obs} the *observations*. X is the *predictand*.

The applied statistician proposes a statistical model for (X, Y),

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

She then turns $\mathcal{E}$ and y^{obs} into a prediction for X. Ideally she uses an algorithm, in the sense that were she given the same statistical model and same observations again, she would produce the same prediction.

A statistical prediction is always a probability distribution for X, although it might be summarised, for example as the expectation of some specified function of X. From the starting point of the statistical model $\mathcal{E}$ and the value of an observable Y we derive the *predictive model*

$$\mathcal{E}^* = \{\mathcal{X}, \Omega, f_X^*\} \tag{1.8a}$$

where

$$f_X^*(\cdot; \theta) = \frac{f_{X,Y}(\cdot, y; \theta)}{f_Y(y; \theta)} \tag{1.8b}$$

$$\text{and } f_Y(y; \theta) = \sum_x f_{X,Y}(x, y; \theta); \tag{1.8c}$$

I often write '$*$' to indicate a suppressed y argument. Here f_X^* is the conditional PMF of X given that $Y = y$, and f_Y is the marginal PMF of Y. Both of these depend on the parameter θ. The challenge for prediction is to reduce the family of distributions $\mathcal{E}^*$ down to a single distribution; effectively, to 'get rid of' θ.

There are two approaches to getting rid of θ: *plug in* and *integrate out*, found in the Frequentist and Bayesian paradigms respectively, for reasons that will be made clear below. We accept, as our working hypothesis, that one of the elements of the family $\mathcal{F}$ is true. For a specified statistical model $\mathcal{E}$, this is equivalent to stating that exactly one element in Ω is true: denote this element as Θ.[11,12] Then $f_X^*(\cdot; \Theta)$ is the true predictive PMF for X.

For the plug-in approach we replace Θ with an estimate based on y, for example the MLE $\hat{\theta}$. In other words, we have an algorithm

$$y \mapsto f_X^*(\cdot\,; \hat{\theta}(y)) \tag{1.9}$$

to derive the predictive distribution for X for any y. The estimator does not have to be the MLE: different estimators of Θ produce different algorithms.

For the integrate-out approach we provide a *prior distribution* over Ω, denoted π.[13] This produces a *posterior distribution*

$$\pi^*(\cdot) = \frac{f_Y(y; \cdot)\, \pi(\cdot)}{\mathrm{p}(y)} \tag{1.10a}$$

[11] Note that I do not feel the need to write 'true' in scare-quotes. Clearly there is no such thing as a true value for θ, because the model is an artefact (i.e. not true in any defensible sense). But once we accept, as a working hypothesis, that one of the elements of $\mathcal{F}$ is true, we do not have to belabour the point.

[12] I am following Schervish (1995) and using Θ for the true value of θ, although it is a bit clunky as notation.

[13] For simplicity, and almost always in practice, π is a probability density function (PDF), given that Ω is almost always a convex subset of Euclidean space.

where

$$\mathrm{p}(y) = \int_{\Omega} f_Y(y;\theta)\, \pi(\theta)\, \mathrm{d}\theta \tag{1.10b}$$

(Bayes's theorem, of course). Here $\mathrm{p}(y)$ is termed the *marginal likelihood* of y. Then we integrate out θ according to the posterior distribution—another algorithm:

$$y \mapsto \int_{\Omega} f_X^*(\cdot\,;\theta)\, \pi^*(\theta)\, \mathrm{d}\theta. \tag{1.11}$$

Different prior distributions produce different algorithms.

That is prediction in a nutshell. In the plug-in approach, each estimator for Θ produces a different algorithm. In the integrate-out approach each prior distribution for Θ produces a different algorithm. Neither approach works on y alone: both need the statistician to provide an additional input: a point estimator, or a prior distribution. Frequentists dislike specifying prior distributions, and therefore favour the plug-in approach. Bayesians like specifying prior distributions, and therefore favour the integrate-out approach.[14]

* * *

This outline of prediction illustrates exactly how Statistics has become so concerned with *inference*. Inference is learning about Θ, which is a key part of either approach to prediction: either we need a point estimator for Θ (plug-in), or we need a posterior distribution for Θ (integrate-out). It often seems as though Statistics is mainly about inference, but this is misleading. It is about inference only insofar as inference is the first part of prediction.

Ideally, algorithms for inference should only be evaluated in terms of their performance as components of algorithms for prediction. This does not happen in practice: partly because it is much easier to assess algorithms for inference than for prediction; partly because of the fairly well-justified belief that algorithms that perform well for inference will produce algorithms that perform well for prediction. I will adhere to this practice, and focus mainly on inference. *But not forgetting that Statistics is mainly about prediction.*

1.4 Frequentist procedures

As explained immediately above, I will focus on inference. So consider a specified statistical model $\mathcal{E} = \{\mathcal{Y}, \Omega, f_Y\}$, where the objective is to learn about the true value $\Theta \in \Omega$ based on the value of the observables Y.

We have already come across the notion of an *algorithm*, which is represented as a function of the value of the observables; in this section I will denote the algorithm as 'g'. Thus the domain of g is always $\mathcal{Y}$. The co-domain of g depends on the type of inference (see below for examples). The key feature of the Frequentist paradigm is the following principle.

[14] We often write 'Frequentists' and 'Bayesians', and most applied statisticians will tend to favour one approach or the other. But applied statisticians are also pragmatic. Although a 'mostly Bayesian' myself, I occasionally produce confidence sets.

Definition 1.1 (Certification)**.** For a specified model $\mathcal{E}$ and algorithm g, the *sampling distribution* of g is

$$f_G(v;\theta) = \sum_{y:g(y)=v} f_Y(y;\theta). \tag{1.12}$$

Then:

1. Every algorithm is certified by its sampling distribution, and
2. The choice of algorithm depends on this certification.

This rather abstract principle may not be what you were expecting, based on your previous courses in Statistics, but if you reflect on the following outline you will see that is the common principle underlying what you have previously been taught.

Different algorithms are certified in different ways, depending on their nature. Briefly, point estimators of Θ may be certified by their *Mean Squared Error function*. Set estimators of Θ may be certified by their *coverage function*. Hypothesis tests for Θ may be certified by their *power function*. The definition of each of these certifications is not important here, although they are easy to look up. What is important to understand is that in each case an algorithm g is proposed, f_G is inspected, and then a certificate is issued.

Individuals and user communities develop conventions about what certificates they like their algorithms to possess, and thus they choose an algorithm according to its certification. They report both $g(y^{\text{obs}})$ and the certification of g. For example, "$(0.73, 0.88)$ is a 95% confidence interval for Θ". In this case g is a set estimator for Θ, it is certified as 'level 95%', and its value is $g(y^{\text{obs}}) = (0.73, 0.88)$.

* * *

Certification is extremely challenging. Suppose I possess an algorithm $g : \mathcal{Y} \to 2^{\Omega}$ for set estimation.[15] In order to certify it as a confidence procedure for my model $\mathcal{E}$ I need to compute its coverage for every $\theta \in \Omega$, defined as

$$\text{coverage}(\theta; \mathcal{E}) = \Pr\{\theta \in g(Y); \theta\} = \sum_v \mathbb{1}_{\theta \in v} f_G(v;\theta), \tag{1.13}$$

where '$\mathbb{1}_a$' is the indicator function of the proposition a, which is 0 when a is false, and 1 when a is true. Except in special cases, computing the coverage for every $\theta \in \Omega$ is impossible, given that Ω is uncountable.[16]

So, in general, I cannot know the coverage function of my algorithm g for my model $\mathcal{E}$, and thus I cannot certify it accurately, but only approximately. Unfortunately, then I have a second challenge. After much effort, I might (approximately) certify g for my model $\mathcal{E}$ as, say, 'level 83%'; this means that the coverage is at least 83% for every $\theta \in \Omega$. Unfortunately, the convention in my user community is that confidence procedures should be certified as 'level 95%'. So it turns out that my community will not accept g. I have to find a

[15] Notation. 2^{Ω} is the set of all subsets of Ω, termed the 'power set' of Ω.

[16] The special cases are a small subset of models from (A) in the model hierarchy in Section 1.2, where, for a particular choice of g, the sampling distribution of g and the coverage of g can be expressed as an analytic function of θ. If you ever wondered why the Normal linear model is so common in applied statistics (linear regression, z-scores, t-tests, and F-statistics, ANOVA, etc.), then wonder no more. Effectively, this family makes up most of the special cases.

way to work backwards, *from* the required certificate, *to* the choice of algorithm.

So Frequentist procedures require the solution of an intractable inverse problem: for specified model $\mathcal{E}$, produce an algorithm g with the required certificate. Actually, it is even harder than this, because it turns out that there are an uncountable number of algorithms with the right certificate, but most of them are useless. Most applied statisticians do not have the expertise or the computing resources to solve this problem to find a good algorithm with the required certificate, for their model $\mathcal{E}$. And so Frequentist procedures, when they are used by applied statisticians, tend to rely on a few special cases. Where these special cases are not appropriate, applied statisticians tend to reach for an off-the-shelf algorithm justified using a theoretical approximation, plus hope.

The empirical evidence collected over the last decade suggests that the hope has been in vain. Most algorithms (including those based on the special cases) did not, in fact, have the certificate that was claimed for them.[17] Opinion is divided about whether this is fraud or merely ignorance. Practically speaking, though, there is no doubt that Frequentist procedures are not being successfullly implemented by applied statisticians.

1.5 Bayesian procedures

We continue to treat the model $\mathcal{E}$ as given. As explained in the previous section, Frequentist procedures select algorithms according to their certificates. By contrast, Bayesian procedures select algorithms mainly according to the prior distribution π (see Section 1.3), without regard for the algorithm's certificate.

A Bayesian inference is synonymous with the posterior distribution π^*, see (1.10). This posterior distribution may be summarized according to some method, for example to give a point estimate, a set estimate, do a hypothesis test, and so on. These summary methods are fairly standard, and do not represent an additional source of choice for the statistician. For example, a Bayesian algorithm for choosing a set estimator for Θ would be (i) choose a prior distribution π, (ii) compute the posterior distribution π^*, and (iii) extract the 95% High Density Region (HDR).

In principle, we could compute the coverage function of this algorithm, and certify it as a confidence procedure. It is very unlikely that it would be certified as a 'level 95%' confidence procedure, because of the influence of the prior distribution.[18] A Bayesian statistician would not care, though, because she does not concern herself with the certificate of her algorithm. When the model is given, the only thing the Bayesian has to worry about is her prior distribution.

Bayesians see the prior distribution as an opportunity to construct a richer model for (X, Y) than is possible for Frequentists. This is most easily illustrated with a hierarchical model, for a

[17] See Madigan et al. (2014) for one such study or, if you want to delve, google "crisis reproducibility science". There is even a wikipedia page, `https://en.wikipedia.org/wiki/Replication_crisis`, which dates from Jan 2015.

[18] Nevertheless, there are theorems that give conditions on the model and the prior distribution such that the posterior 95% HDR is approximately a level 95% confidence procedure; see, e.g., Schervish (1995, ch. 7).

population of quantities that are similar, and a sample from that population. Hierarchical models have a standard notation:[19]

$$Y_i \mid X_i, \sigma^2 \sim f_{\epsilon_i}(X_i, \sigma^2) \qquad i = 1, \dots, n \tag{1.14a}$$

$$X_i \mid \theta_i \sim f_{X_i}(\theta_i) \qquad i = 1, \dots, m \tag{1.14b}$$

$$\theta_i \mid \psi \sim f_\theta(\psi) \qquad i = 1, \dots, m \tag{1.14c}$$

$$(\sigma^2, \psi) \sim f_0 \, . \tag{1.14d}$$

At the top (first) level is the measurement model for the sample $(Y_1, \dots, Y_n)$, where f_{ϵ_i} describes the measurement error and σ^2 would usually be a scale parameter. At the second level is the model for the population $(X_1, \dots, X_m)$, where $n \leq m$, showing how each element X_i is 'summarised' by its own parameter θ_i. At the third level is the parameter model, in which the parameters are allowed to be different from each other. At the bottom (fourth) level is the 'hyper-parameter' model, which describes how much the parameters can differ, and also provides a PDF for the scale parameter σ^2.

Frequentists would specify their statistical model using just the top two levels, in terms of the parameter $(\sigma^2, \theta_1, \dots, \theta_m)$, or, if this is too many parameters for the n observables, as it usually is, they will insist that $\theta_1 = \cdots = \theta_m = \theta$, and have just (σ^2, θ). The bottom two levels are the Bayesian's prior distribution. By adding these two levels, Bayesians can allow the θ_i's to vary, but in a limited way that can be controlled by their choices for f_θ and f_0. Usually, f_0 is a 'vague' PDF selected according to some simple rules.

In a Frequentist model we can count the number of parameters, namely $1 + m \cdot \dim \Omega$, or just $1 + \dim \Omega$ if the θ_i's are all the same. We can do that in a Bayesian model too, to give $1 + m \cdot \dim \Omega + \dim \Psi$, if Ψ is the realm of ψ. Bayesian models tend to have many more parameters, which makes them more flexible. But there is a second concept in a Bayesian model, which is the *effective* number of parameters. This can be a lot lower than the actual number of parameters, if it turns out that the observations indicate that the θ_i's are all very similar. So in a Bayesian model the effective number of parameters can depend on the observations. In this sense, a Bayesian model is more adaptive than a Frequentist model.[20]

1.6 *So who's right?*

We return to the problem of inference, based on the model $\mathcal{E} = \{\mathcal{Y}, \Omega, f_Y\}$. Here is the pressing question, from the previous two sections: should we concern ourselves with the certificate of the algorithm, or with the choice of the prior distribution?

A Frequentist would say "Don't you want to know that you will be right 'on average' according to some specified rate?" (like 95%). And a Bayesian will reply "Why should my rate 'on average' matter to me right now, when I am thinking only of Θ?"[21] The Bayesian

[19] See, e.g., Lunn et al. (2013) or Gelman et al. (2014). Each of the f functions is a PMF or PDF, and the first argument is suppressed. The i index in the first three rows indicates that the components are mutually independent, and then the f function shows the marginal distribution for each i, which may depend on i. In the third row f does not depend on i, so that the θ_i's are mutually independent and identically distributed, or 'IID'.

[20] The issue of how to quantify the effective number of parameters is quite complicated. Spiegelhalter et al. (2002) was a controversial suggestion, and there have been several developments since then, summarised in Spiegelhalter et al. (2014).

[21] And if she really wants to twist the knife she will also mention the overwhelming evidence that Frequentist statisticians have apparently not been able to achieve their target rates, mentioned at the end of Section 1.4.

will point out the advantage of being able to construct hierarchical models with richer structure. Then the Frequentist will criticise the 'subjectivity' of the Bayesian's prior distribution. The Bayesian will reply that the model is also subjective, and so 'subjectivity' of itself cannot be used to criticise only Bayesian procedures. And she will go on to point out that there is just as much subjectivity in the Frequentist's choice of algorithm as there is in the Bayesian's choice of prior.

There is no clear winner when two paradigms butt heads. However, momentum is now on the side of the Bayesians. Back in the 1920s and 1930s, at the dawn of modern Statistics, the Frequentist paradigm seemed to provide the 'objectivity' that was then prized in science. And computation was so rudimentary that no one thought beyond the simplest possible models, and their natural algorithms. But then the Frequentist paradigm took a couple of hard knocks: from Wald's Complete Class Theorem in 1950 (covered in Chapter 3), and from Birnbaum's Theorem and the Likelihood Principle in the 1960s (covered in Chapter 2). Significance testing was challenged by Lindley's paradox; estimator theory by Stein's paradox and the Neyman-Scott paradox. Bayesian methods were much less troubled by these results, and were developed in the 1950s and 1960s by two very influential champions, L.J. Savage and Dennis Lindey, building on the work of Harold Jeffreys.[22]

And then in the 1980s, the exponential growth in computer power and new Monte Carlo methods combined to make the Bayesian approach much more practical. Additionally, datasets have got larger and more complicated, favouring the Bayesian approach with its richer model structure, when incorporating the prior distribution. Finally, there is now much more interest in uncertainty in predictions, something that the Bayesian integrate-out approach handles much better than the Frequentist plug-in approach (Section 1.3).

However, I would not rule out a partial reversal in due course, under pressure from Machine Learning (ML). ML is all about algorithms, which are often developed quite independently of any statistical model. With modern Big Data (BD), the primary concern of an algorithm is that it executes in a reasonable amount of time (see, e.g., Cormen et al., 1990). But it would be natural, when an ML algorithm might be applied by the same agent thousands of times in quite similar situations, to be concerned about its sampling distribution.[23] With BD the certificate can be assessed from a held-out subset of the data, without any need for a statistical model—no need for statisticians at all then! Luckily for us statisticians, there will always be plenty of applications where ML techniques are less effective, because the datasets are smaller, or more complicated. In these applications, I expect Bayesian procedures will come to dominate.

[22] With a strong assist from the maverick statistician I.J. Good. The intellectual forebears of the 20th century Bayesian revival included J.M. Keynes, F.P. Ramsey, Bruno de Finetti, and R.T. Cox.

[23] For example, if an algorithm is a binary classifier, to want to know its 'false positive' and 'false negative' rates.

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'.[2]

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.

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 $f(x; \bullet) = f(x'; \bullet)$ 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 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 $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}', x)$.

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

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

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

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

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

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 $\mathcal{E}^*$, and it has outcomes of the form (i, x_i), and a statistical model of the form $f^*\big((i, x_i); \bullet\big) = p_i \cdot f_i(x_i; \bullet)$.

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.

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

$$f_1(x_1; \bullet) = c(x_1, x_2) \cdot f_2(x_2; \bullet) \tag{2.2}$$

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. 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 $x_1 \in \mathcal{X}_1$ 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

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

Then the WIP implies that

$$\text{Ev}\left(\mathcal{E}^*, (1, x_1)\right) = \text{Ev}\left(\mathcal{E}^*, (2, x_2)\right).$$

Finally, apply the WCP to each side to infer that

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

as required. □

[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.

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. Let the experiment $\mathcal{E}$ and the outcome x be fixed. Let $\mathcal{E}_2 := \{\mathcal{X}, \Omega, f_2\}$ be another experiment, where $f_2(x; \bullet) = f(x; \bullet)$, but $f_2(x'; \theta) \neq f(x'; \theta)$ for at least one $x' \in \mathcal{X} \setminus \{x\}$ and at least one $\theta \in \Omega$.[6] If Ev is affected by the allocation of probabilities for outcomes that do not occur, then we can be sure that $\text{Ev}(\mathcal{E}_2, x) \neq \text{Ev}(\mathcal{E}, x)$ for some choice of f_2. This contradicts the LP, which would imply that $\text{Ev}(\mathcal{E}_2, x) = \text{Ev}(\mathcal{E}, x)$ because $f_2(x; \bullet) = f(x; \bullet)$. □

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 (Ancillarity)**.** X is ancillary for θ_2 in experiment $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega_1 \times \Omega_2, f_{X,Y}\}$ exactly when $f_{X,Y}$ factorises as

$$f_{X,Y}(x, y; \theta) = f_X(x; \theta_1) \cdot f_{Y|X}(y \mid x; \theta_2).$$

X is ancillary in $\{\mathcal{X} \times \mathcal{Y}, \Omega, f_{X,Y}\}$ exactly when f_X does not depend on θ.

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

$$\mathcal{E}^{Y|x} = \{\mathcal{Y}, \Omega, f_{Y|x}\}, \tag{2.3}$$

where $f_{Y|x}(\bullet; \theta) := f_{Y|X}(\bullet \mid x; \theta)$. 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 $\text{Ev}\,(\mathcal{E}, (x, y)) = \text{Ev}(\mathcal{E}^{Y|x}, y)$.

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.

[6] Actually, $f_2(x'; \theta)$ must vary at at least two values in $\mathcal{X} \setminus \{x\}$, due to the constraint that $\sum_x f_2(x; \theta) = 1$.

Theorem 2.4. *LP → CP.*

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

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

Then the LP implies that

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

as required. □

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 X is ancillary for θ_2, where Ω_1 is not empty. In other words, the distribution for X is incompletely specified, but its parameters (i.e. θ_1) are distinct from the parameters of $f_{Y|X}$ (i.e. θ_2).

Definition 2.8 (Auxiliary parameter). θ_1 is auxiliary in the experiment defined in Definition 2.6 exactly when X is ancillary for θ_2, and Θ_2 is of interest.

Now this would be a *really* useful principle:

Definition 2.9 (Strong Conditionality Principle, SCP). If θ_1 is auxiliary in experiment $\mathcal{E}$, and Ev_2 denotes the evidence about Θ_2, then $\text{Ev}_2\left(\mathcal{E}, (x, y)\right) = \text{Ev}(\mathcal{E}^{Y|x}, y)$.

The SCP would allow us to treat all ancillary quantities whose parameters were uninteresting to us as though they were known, and condition on them, thus removing all reference to their unknown marginal distribution and their parameters.

For example, we have a sample of size n, but we are unsure about all the circumstances under which the sample was collected, and suspect that n itself is the outcome of an experiment with a random N. But as long as we are satisfied that the parameter controlling the sampling process was auxiliary, the SCP asserts that we can treat n as known, and condition on it.

Here is another example, which will be familiar to all statisticians. A regression of Y on X appears to make a distinction between Y, which is random, and X, which is not. This distinction is insupportable, given that the roles of Y and X are often interchangeable, and determined by the *hypothèse du jour*. What is really happening is that (X, Y) is random, but X is being treated as ancillary for the parameters in $f_{Y|X}$, so that its parameters are auxiliary in the analysis. Then the SCP is invoked (implicitly), which justifies modelling Y conditionally on X, treating X as known.

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 $\mathcal{E}$ and $\mathcal{E}'$ be two experiments with the same parameter $\theta = (\theta_1, \theta_2)$. Let $\mathrm{Ev}_2(\bullet; \theta_1)$ denote the evidence for Θ_2, with $\Theta_1 = \theta_1$. If

$$\mathrm{Ev}_2(\mathcal{E}, x; \theta_1) = \mathrm{Ev}_2(\mathcal{E}', x'; \theta_1) \quad \text{for every } \theta_1 \in \Omega_1,$$

then $\mathrm{Ev}_2(\mathcal{E}, x) = \mathrm{Ev}_2(\mathcal{E}', x')$, where Ev_2 is the evidence for Θ_2.

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 ∧ STP) → SCP.*

Proof. Consider the experiment from Definition 2.6. Treat θ_1 as known, in which case the parameter is θ_2, X is ancillary, and the CP asserts that

$$\mathrm{Ev}_2\left(\mathcal{E}, (x, y); \theta_1\right) = \mathrm{Ev}_2(\mathcal{E}^{Y|x}, y; \theta_1).$$

As this equality holds for all $\theta_1 \in \Omega_1$, the STP implies that

$$\mathrm{Ev}_2\left(\mathcal{E}, (x, y)\right) = \mathrm{Ev}_2(\mathcal{E}^{Y|x}, y),$$

as required. □

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_2.

2.5 *Stopping rules*

Here is a surprising but gratifying consequence of the LP, which can be strengthened under the SCP.

Consider a sequence of random quantities $X_1, X_2, \dots$ with marginal PMFs[8]

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

In a sequential experiment, the number of X's that are observed is not fixed in advanced but depends deterministically on the values seen so far. That is, at time j, the decision to observe X_{j+1} can be modelled by a set $\mathcal{A}_j \subset \mathcal{X}^j$, where sampling stops if $(x_1, \dots, x_j) \in \mathcal{A}_j$, and continues otherwise.[9] We can assume, resources being finite, that the experiment must stop at specified time m, if it has not stopped already. Denote the stopping rule as $\tau := (\mathcal{A}_1, \dots, \mathcal{A}_m)$, where $\mathcal{A}_m = \mathcal{X}^m$.

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

[8] These must satisfy Kolmogorov's consistency theorem.

[9] Implicit in this definition is that $(x_1, \dots, x_{j-1}) \notin \mathcal{A}_{j-1}$.

Definition 2.11 (Stopping Rule Principle, SRP)**.** In a sequential experiment $\mathcal{E}^\tau$, $\text{Ev}\left(\mathcal{E}^\tau, (x_1, \dots, x_n)\right)$ 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 $(x_1, \dots, x_n)$ are available. Thus the statistician could proceed as though the simplest possible stopping rule were in effect, which is $\mathcal{A}_1 = \cdots = \mathcal{A}_{n-1} = \emptyset$, and $\mathcal{A}_n = \mathcal{X}^n$, 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. Let $\mathcal{Y} = \mathcal{X} \cup \{s\}$, where $Y_i = X_i$ while the experiment is running, and $Y_i = s$ once it has stopped. Then the experiment is

$$\mathcal{E}^\tau := (\mathcal{Y}^m, \Omega, f)$$

where

$$f(y_1, \dots, y_m; \theta) = \begin{cases} f_n(x_1, \dots x_n; \theta) & (y_1, \dots, y_n) \in \mathcal{A}_n \text{ and } y_{n+1} = \cdots = y_m = s \\ 0 & \text{otherwise.} \end{cases}$$

The condition in the top branch is a deterministic function of $y_1, \dots, y_m$, which we can write as $q(y_1, \dots, y_m) \in \{\mathsf{FALSE}, \mathsf{TRUE}\}$. Thus we have

$$f(y_1, \dots, y_m; \theta) = \mathbb{1}_{q(y_1, \dots, y_m)} \cdot f_n(x_1, \dots, x_n; \theta) \qquad \text{for all } \theta \in \Omega$$

where $\mathbb{1}_q$ is the indicator function of the first-order sentence q. Hence

$$f(y_1, \dots, y_m; \bullet) = c(y_1, \dots, y_m) \cdot f_n(x_1, \dots, x_n; \bullet)$$

and so, by the LP,

$$\text{Ev}\left(\mathcal{E}^\tau, (x_1, \dots, x_n, s, \dots, s)\right) = \text{Ev}\left(\mathcal{E}^n, (x_1, \dots, x_n)\right), \qquad (\dagger)$$

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 have the same outcome $x = (\text{T,H,T,T,H,H,T,H,H,H})$:

$\mathcal{E}_1$ 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, twice. First, to stopping rules which are stochastic functions of $(x_1, \dots, x_j)$, i.e. where the probability of stopping at j is some known function $p_j(x_1, \dots, x_j)$, for $j = 1, 2, \dots, m$, with $\text{p}_m(x_1, \dots, x_m) = 1$. This stronger version is still implied by the LP. Second, 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 $\text{p}_j(x_1, \dots, x_j; \psi)$ is unrelated to the true value Θ. This much stronger version is implied by the SCP (Definition 2.9). Both proofs are straightforward, although tedious to type. In the absence of any information about the experimenter's intentions, the strongest version of the SRP is the one 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 Twentieth 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

Finally in this chapter 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 and ignoring stopping rules.

2.6.1 Likelihood-based inference (LBI)

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

$$L : \theta \mapsto f(x; \theta). \tag{2.4}$$

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,

$$\text{Ev}(\mathcal{E}, x) = \phi(L) = \phi\big(c(x) \cdot L\big)$$

for some operator ϕ depending on Ev, and any $c > 0$.

The invariance of ϕ 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)*

$$\hat{\theta} = \operatorname*{argsup}_{\theta \in \Omega} L(\theta) \tag{2.5}$$

for point estimation, and *Wilks level sets*

$$\widehat{C}_k = \big\{\theta \in \Omega : \log L(\theta) \geq \log L(\hat{\theta}) - k\big\} \tag{2.6}$$

for set estimation and hypothesis testing, where k may depend on y. Wilks level sets have the interesting and reassuring property that they are invariant to bijective transformations of the parameter.[10]

Both of these operators satisfy $\phi(L) = \phi(c \cdot L)$. 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. 2.6).

LBI satisfies the LP by construction, so it also satisfies the CP. To see whether it satisfies the SCP requires a definition of Ev_2, the evidence for Θ_2 in the case where $\Theta = (\Theta_1, \Theta_2)$. The standard definition is based on the *profile likelihood*,

$$L_2 : \theta_2 \mapsto \sup_{\theta_1 \in \Omega_1} L(\theta_1, \theta_2), \tag{2.7}$$

from which

$$\text{Ev}_2(\mathcal{E}, x) := \phi(L_2). \tag{2.8}$$

Then we have the following result.

Theorem 2.7. *If profile likelihood is used for* Ev_2*, then LBI satisfies the SCP.*

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

Proof. Under the conditions of Definition 2.9 we have, putting '•' where the θ_2 argument goes,

$$\begin{aligned}
\mathrm{Ev}_2\{\mathcal{E}, (x, y)\} &= \phi\{\sup_{\theta_1} L(\theta_1, \bullet)\} \\
&= \phi\{\sup_{\theta_1} f_X(x; \theta_1) \cdot f_{Y|X}(y \mid x; \bullet)\} \\
&= \phi\{c(x) \cdot f_{Y|X}(y \mid x; \bullet)\} \\
&= \mathrm{Ev}(\mathcal{E}^{Y|x}, y),
\end{aligned}$$

where $\mathcal{E}^{Y|x}$ was defined in (2.3). □

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 $\mathcal{E} := \{\mathcal{X}, \Omega, f\}$ with a prior probability distribution π on Ω, representing initial beliegfs abourt Θ. The *posterior distribution* for Θ is found by conditioning on the outcome x, to give

$$\pi^*(\theta) \propto f(x; \theta) \cdot \pi(\theta) = L(\theta) \cdot \pi(\theta) \tag{2.9}$$

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 $1/\Pr(X = x)$.

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

$$\mathrm{Ev}(\mathcal{E}, x) = \phi(\pi^*) = \phi\big(c(x) \cdot \pi^*\big)$$

for some operator ϕ depending on Ev, and any $c > 0$.

The presence of c in ϕ indicates that the BCP will, if necessary, normalize the argument to ϕ if it does not integrate to 1 over Ω. So it is fine to write $\mathrm{Ev}(\mathcal{E}, x) = \phi(L \cdot \pi)$. Compared to LBI, Bayesian inference needs an extra object in order to compute Ev, namely the prior distribution π.

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

$$\hat{\theta}^* = \operatorname*{argsup}_{\theta \in \Omega} \pi^*(\theta). \tag{2.10}$$

The MAP estimator does not require the calculation of the multiplicative constant $1/\Pr(X = x)$. 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*

$$C_k^* := \left\{ \theta \in \Omega \,\middle|\, \log \pi^*(\theta) \geq k \right\}. \tag{2.11}$$

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

$$\int_{C_k^*} \pi^*(\theta) \, \mathrm{d}\theta = 0.95. \tag{2.12}$$

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.[11] 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

$$\begin{aligned} \mathrm{Ev}(\mathcal{E}_1, x_1) &= \phi(L_1 \cdot \pi) \\ &= \phi(c \cdot L_2 \cdot \pi) \\ &= \phi(L_2 \cdot \pi) \\ &= \mathrm{Ev}(\mathcal{E}_2, x_2). \end{aligned} \qquad \square$$

Hence BCP also satisfies the CP. What about the SCP? As for LBI in Section 2.6.1, this requires a definition of Ev_2. In the Bayesian approach there is only one choice, based on the marginal posterior distribution

$$\pi_2^* := \theta_2 \mapsto \int_{\theta_1 \in \Omega_1} \pi^*(\theta_1, \theta_2) \, \mathrm{d}\theta_1, \tag{2.13}$$

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

from which

$$\mathrm{Ev}_2(\mathcal{E}, x) = \phi(\pi_2^*) = \phi\big(c(x) \cdot \pi_2^*\big). \tag{2.14}$$

Then we have the following result.

Theorem 2.8. *If $\pi(\theta_1, \theta_2) = \pi_1(\theta_1) \cdot \pi_2(\theta_2)$, then Bayesian inference satisfies the SCP.*

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

$$\begin{aligned}\pi^*(\theta_1, \theta_2) &\propto L(\theta_1, \theta_2) \cdot \pi(\theta_1, \theta_2) \\ &= f_X(x; \theta_1) \cdot f_{Y|X}(y \mid x; \theta_2) \cdot \pi_1(\theta_1) \cdot \pi_2(\theta_2) \\ &\propto f_{Y|X}(y \mid x; \theta_2) \cdot \pi_2(\theta_2) \cdot \pi_1^*(\theta_1 \mid x).\end{aligned}$$

Integrating out θ_1 shows that

$$\pi_2^*(\bullet) \propto f_{Y|x}(y; \bullet) \cdot \pi_2(\bullet),$$

using the definition of $f_{Y|x}$ from (2.3). Thus

$$\begin{aligned}\mathrm{Ev}_2\left(\mathcal{E}, (x, y)\right) &= \phi(\pi_2^*) \\ &= \phi\big(f_{Y|x}(y; \bullet) \cdot \pi_2(\bullet)\big) \\ &= \mathrm{Ev}(\mathcal{E}^{Y|x}, y)\end{aligned}$$

□

Therefore, under the mild condition that $\pi = \pi_1 \cdot \pi_2$, 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 ϕ. 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 methods do not respect the LP, because the sampling distribution of the algorithm depends on values for f other than $f(x; \bullet)$.

Frequentist statisticians are caught between between Scylla and Charybdis.[12] To reject the LP is to reject one of the WIP and WCP, and these seem self-evident. On the other hand, in their everyday practice Frequentist statisticians use the (S)CP or SRP, both of which are most easily justified as consequences of the LP. The (S)CP and SRP are not self-evident. This means that, if we are to accept them without the support of the LP, we must do so on the personal authority of individual statisticians; see, e.g., Cox and Mayo (2010). No matter how much we respect these statisticians, this is not a scientific attitude.

As a practising statistician I want to be able to satisfy an auditor who asks about the logic of my approach.[13] I do not want to agree with him that the WIP and the WCP are self-evident, and then illogically choose to violate the LP. And I do not want to violate the LP, but use the (S)CP or SRP. In terms of volume, most Frequentist Applied Statistics is being done by non-statisticians (despite what it might say on their business cards). Non-statisticians do not know about statistical principles, and are ignorant about whether their approach violates the LP, and what this entails.

I offer this suggestion to auditors: ask which of the WIP or the WCP the Frequentist statistician rejects—this should elicit an informative response. I would not rule out, from a statistician, "I know that my practice is illogical, but if the alternative is to specify a prior distribution, then so be it." This is encouraging, and the basis for further discussion. But anything along the lines of "I'm not sure what you mean" suggests that the so-called statistician has misrepresented herself: she is in fact a 'data analyst'—which is fine, as long as that was what the client paid for.[14]

[12] Or, colloquially, between a rock and a hard place. This was not known before Birnbaum's Theorem, which is why we might think of this result as 'Birnbaum's bombshell'.

[13] As discussed in Smith (2010, ch. 1), there are three players in an inference problem, although two roles may be taken by the same person. There is the client, who has the problem, the statistician whom the client hires to help solve the problem, and the auditor whom the client hires to check the statistician's work.

[14] Another response 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. If you are this person, you can probably charge a lot of money.

3
Statistical Decision Theory

The basic premise of Statistical Decision Theory is that we want to make inferences about the parameter of a family of distributions. So the starting point of this chapter is a family of distributions for the observables $Y \in \mathcal{Y}$ of the general form

$$\mathcal{E} = \{\mathcal{Y}, \Omega, f\},$$

where f is the 'model', θ is the 'parameter', and Ω the 'parameter space', just as in Chapter 1 and Chapter 2. The parameter space Ω may be finite or non-finite, possibly uncountable. In this chapter I will treat it as finite, because this turns out to be much simpler, and the results generalize; hence

$$\Omega = \{\theta_1, \dots, \theta_k\}.$$

The value $f(y;\theta)$ denotes the probability that $Y = y$ under family member θ. I will assume throughout this chapter that $f(y;\theta)$ is easy to evaluate (see Section 1.2).

We accept as our working hypothesis that $\mathcal{E}$ is true, and then inference is learning about Θ, the true value of the parameter. More precisely, we would like to understand how to construct the 'Ev' function from Chapter 2, in such a way that it reflects our needs, which will vary from application to application.

3.1 General Decision Theory

There is a general theory of decision-making, of which Statistical Decision Theory is a special case. Here I outline the general theory, subject to one restriction which always holds for Statistical Decision Theory (to be introduced below). In general we should imagine the statistician applying decision theory on behalf of a client, but for simplicity of exposition I will assume the statistician is her own client.

There is a set of random quantities $X \in \mathcal{X}$. The statistician contemplates a set of *actions*, $a \in \mathcal{A}$. Associated with each action is a consequence which depends on X. This is quantified in terms of a *loss function*, $L : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$, with larger values indicating worse consequences. Thus $L(a, x)$ is the loss incurred by the statistician if action a is taken and X turns out to be x. Before making her choice of action, the statistician will observe $Y \in \mathcal{Y}$. Her choice should be some function of the value of Y, and this is represented as a *decision rule*, $\delta : \mathcal{Y} \to \mathcal{A}$.

The statistician's beliefs about (X, Y) are represented by a probability distribution $f_{X,Y}$, from which she can derive marginal distributions f_X and f_Y, and conditional distributions $f_{X|Y}$ and $f_{Y|X}$, should she need them. Of the many ways in which she might choose δ, one possibility is to minimize her expected loss, and this is termed the *Bayes rule*,

$$\delta^* := \operatorname*{argmin}_{\delta \in \mathcal{D}} \mathrm{E}\{L(\delta(Y), X)\},$$

where $\mathcal{D}$ is the set of all possible rules. The value $\mathrm{E}\{L(\delta(Y), X)\}$ is termed the *Bayes risk* of decision rule δ, and therefore the Bayes rule is the decision rule which minimizes the Bayes risk, for some specified action set, loss function, and joint distribution.

There is a justly famous result which gives the explicit form for a Bayes rule. I will give this result under the restriction anticipated above, which is that $f_{X|Y}$ does not depend on the choice of action. Decision theory can handle the more general case, but it is seldom appropriate for Statistical Decision Theory.

Theorem 3.1 (Bayes Rule Theorem, BRT). *A Bayes rule satisfies*

$$\delta^*(y) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, X) \mid Y = y\} \tag{3.1}$$

whenever $y \in \operatorname{supp} Y$.[1]

This astounding result indicates that the minimization of expected loss over the space of all functions from $\mathcal{Y}$ to $\mathcal{A}$ can be achieved by the pointwise minimization over $\mathcal{A}$ of the expected loss conditional on $Y = y$. It converts an apparently intractable problem into a simple one.

Proof. We have to show that $\mathrm{E}\{L(\delta(Y), X)\} \geq \mathrm{E}\{L(\delta^*(Y), X)\}$ for all $\delta : \mathcal{Y} \to \mathcal{A}$. So let δ be arbitrary. Then

$$\begin{aligned}
\mathrm{E}\{L(\delta(Y), X)\} &= \textstyle\sum_{x,y} L(\delta(y), x) \cdot f_{X,Y}(x, y) \\
&= \textstyle\sum_y \sum_x L(\delta(y), x) \cdot f_{X|Y}(x \mid y)\, f_Y(y) \\
&\geq \textstyle\sum_y \Big\{ \min_a \sum_x L(a, x) f_{X|Y}(x \mid y) \Big\} f_Y(y) \quad \text{as } f_Y \geq 0 \\
&= \textstyle\sum_y \Big\{ \sum_x L(\delta^*(y), x) f_{X|Y}(x \mid y) \Big\} f_Y(y) \\
&= \textstyle\sum_y \sum_x L(\delta^*(y), x) \cdot f_{X|Y}(x \mid y)\, f_Y(y) \\
&= \mathrm{E}\{L(\delta^*(Y), X)\},
\end{aligned}$$

as needed to be shown. □

3.2 *Inference about parameters*

Now consider the special case of Statistical Decision Theory, in which inference is not about some random quantities X, but about

[1] Here, $\operatorname{supp} Y = \{y : f_Y(y) > 0\}$.

the true value of the parameter, denoted Θ. The three main types of inference about Θ are (i) point estimation, (ii) set estimation, and (iii) hypothesis testing. It is a great conceptual and practical simplification that Statistical Decision Theory distinguishes between these three types simply according to their action sets, which are:

Type of inference	Action set $\mathcal{A}$
Point estimation	The parameter space, Ω. See Section 3.4.
Set estimation	The set of all subsets of Ω, denoted 2^{Ω}. See Section 3.5.
Hypothesis testing	A specified partition of Ω, denoted $\mathcal{P}$ below. See Section 3.6.

One challenge for Statistical Decision Theory is that finding the Bayes rule requires specifying a prior distribution over Ω, which I will denote

$$\boldsymbol{\pi} := (\pi_1, \dots, \pi_k) \in \mathbb{S}^{k-1}$$

where $\mathbb{S}^{k-1}$ is the $(k-1)$-dimensional unit simplex.[2] Applying the BRT (Theorem 3.1),

$$\begin{aligned} \delta^*(y) &= \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, \Theta) \mid Y = y\} \\ &= \operatorname*{argmin}_{a \in \mathcal{A}} \textstyle\sum_j L(a, \theta_j) \cdot \pi_j^*(y) \end{aligned}$$

where $\boldsymbol{\pi}^*(y)$ is the posterior distribution, which must of course depend on the prior distribution $\boldsymbol{\pi}$. So the Bayes rule will not be an attractive way to choose a decision rule for Frequentist statisticians, who are reluctant to specify a prior distribution for Θ. These statisticians need a different approach to choosing a decision rule.

The accepted approach for Frequentist statisticians is to narrow the set of possible decision rules by ruling out those that are obviously bad. Define the *risk function* for rule δ as

$$\begin{aligned} R(\delta, \theta) &:= \mathrm{E}\{L(\delta(Y), \theta); \theta\} \\ &= \textstyle\sum_y L(\delta(y), \theta) \cdot f(y; \theta). \end{aligned} \tag{3.2}$$

That is, $R(\delta, \theta)$ is the expected loss from rule δ in family member θ. A decision rule δ *dominates* another rule δ' exactly when

$$R(\delta, \theta) \le R(\delta', \theta) \quad \text{for all } \theta \in \Omega,$$

with a strict inequality for at least one $\theta \in \Omega$. If you had both δ and δ', you would never want to use δ'.[3] A decison rule is *admissible* exactly when it is not dominated by any other rule; otherwise it is *inadmissible*. So the accepted approach is to reduce the set of possible decision rules under consideration by only using admissible rules.

It is hard to disagree with this approach, although one wonders how big the set of admissible rules will be, and how easy it is to enumerate the set of admissible rules in order to choose between them. This is the subject of Section 3.3. To summarise,

[2] That is, the set
$$\left\{p \in \mathbb{R}^k : p_i \ge 0, \textstyle\sum_i p_i = 1\right\}.$$

[3] Here I am assuming that all other considerations are the same in the two cases: e.g. $\delta(y)$ and $\delta'(y)$ take about the same amount of resource to compute.

Theorem 3.2 (Wald's Complete Class Theorem, CCT). *In the case where both the action set $\mathcal{A}$ and the parameter space Ω are finite, a decision rule δ is admissible* if and only if *it is a Bayes rule for some prior distribution $\boldsymbol{\pi}$ with strictly positive values.*

There are generalisations of this theorem to non-finite realms for Y, non-finite action sets, and non-finite parameter spaces; however, the results are highly technical. See Schervish (1995, ch. 3), Berger (1985, chs 4, 8), and Ghosh and Meeden (1997, ch. 2) for more details and references to the original literature.

So what does the CCT say? First of all, if you select a Bayes rule according to some prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$ then you cannot ever choose an inadmissible decision rule.[4] So the CCT states that there is a very simple way to protect yourself from choosing an inadmissible decision rule. Second, if you cannot produce a $\boldsymbol{\pi} \gg \mathbf{0}$ for which your proposed rule δ is a Bayes Rule, then you cannot show that δ is admissible.

But here is where you must pay close attention to logic. Suppose that δ' is inadmissible and δ is admissible. It does not follow that δ dominates δ'. So just knowing of an admissible rule does not mean that you should abandon your inadmissible rule δ'. You can argue that although you know that δ' is inadmissible, you do not know of a rule which dominates it. All you know, from the CCT, is the family of rules within which the dominating rule must live: it will be a Bayes rule for some $\boldsymbol{\pi} \gg \mathbf{0}$. This may seem a bit esoteric, but it is crucial in understanding modern parametric inference. Statisticians sometimes use inadmissible rules according to standard loss functions. They can argue that yes, their rule δ is or may be inadmissible, which is unfortunate, but since the identity of the dominating rule is not known, it is not wrong to go on using δ. Do not attempt this line of reasoning with your client!

3.3 The Complete Class Theorem

This section can be skipped once the previous section has been read. But it describes a very beautiful result, Theorem 3.2 above, originally due to an iconic figure in Statistics, Abraham Wald.[5] I assume throughout this section that all sets are finite: the realm $\mathcal{Y}$, the action set $\mathcal{A}$, and the parameter space Ω.

The CCT is if-and-only-if. Let $\boldsymbol{\pi}$ be any prior distribution on Ω. Both branches use a simple result that relates the Bayes Risk of a decision rule δ to its Risk Function:

$$\begin{aligned}\mathrm{E}\{L(\delta(Y),\Theta)\} &= \textstyle\sum_j \mathrm{E}\{L(\delta(Y),\theta_j);\theta_j\} \cdot \pi_j \qquad \text{by the LIE} \\ &= \textstyle\sum_j R(\delta,\theta_j) \cdot \pi_j, \qquad (\dagger)\end{aligned}$$

where 'LIE' is the *Law of Iterated Expectation*.[6] The first branch is easy to prove.

Theorem 3.3. *If δ is a Bayes rule for prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$, then it is admissible.*

[4] Here I am using a fairly common notion for vector inequalities. If all components of $\boldsymbol{x}$ are non-negative, I write $\boldsymbol{x} \geq \mathbf{0}$. It in addition at least one component is positive, I write $\boldsymbol{x} > \mathbf{0}$. If all components are positive I write $\boldsymbol{x} \gg \mathbf{0}$. For comparing two vectors, $\boldsymbol{x} \geq \boldsymbol{y}$ exactly when $\boldsymbol{x} - \boldsymbol{y} \geq \mathbf{0}$, and so on.

[5] For his tragic story, see https://en.wikipedia.org/wiki/Abraham_Wald.

[6] Sometimes called the 'Tower Property' of Expectation.

Proof. By contradiction. Suppose that the Bayes rule δ is not admissible; i.e. there exists a rule δ' which dominates it. In this case

$$\begin{aligned} \mathrm{E}\{L(\delta(Y), \Theta)\} &= \textstyle\sum_j R(\delta, \theta_j) \cdot \pi_j && \text{from } (\dagger) \\ &> \textstyle\sum_j R(\delta', \theta_j) \cdot \pi_j && \text{if } \boldsymbol{\pi} \gg \mathbf{0} \\ &= \mathrm{E}\{L(\delta'(Y), \theta)\} \end{aligned}$$

and hence δ cannot have been a Bayes rule, because δ' has a smaller expected loss. The strict inequality holds if δ' dominates δ *and* $\boldsymbol{\pi} \gg \mathbf{0}$. Without it, we cannot deduce a contradiction. □

The second branch of the CCT is harder to prove. The proof uses one of the great theorems in Mathematics, the Supporting Hyperplane Theorem (SHT, given below in Theorem 3.5).

Theorem 3.4. *If δ is admissible, then it is a Bayes rule for some prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$.*

I will give an algebraic proof here, but blackboard proof in the simple case where $\Omega = \{\theta_1, \theta_2\}$ is more compelling. The blackboard proof is given in Cox and Hinkley (1974, sec. 11.6).

For a given loss function L and model f, construct the *risk matrix*,

$$R_{ij} := R(\delta_i, \theta_j)$$

over the set of all decision rules. If there are m decision rules altogether (m is finite because $\mathcal{Y}$ and $\mathcal{A}$ are both finite), then R represents m points in k-dimensional space, where k is the cardinality of Ω.

Now consider *randomised rules*, indexed by $\boldsymbol{w} \in \mathbb{S}^{m-1}$. For randomised rule $\boldsymbol{w}$, actual rule δ_i is selected with probability w_i. The risk for rule $\boldsymbol{w}$ is

$$\begin{aligned} R(\boldsymbol{w}, \theta_j) &:= \textstyle\sum_i \mathrm{E}\{L(\delta_i(Y), \theta_j); \theta_j\} \cdot w_i && \text{by the LIE} \\ &= \textstyle\sum_i R(\delta_i, \theta_j) \cdot w_i. \end{aligned}$$

If we also allow randomised rules—and there is no reason to disallow them, as the original rules are all still available as special cases—then the set of risks for all possible randomised rules is the *convex hull* of the rows of the risk matrix R, denoted $[R] \subset \mathbb{R}^k$, and termed the *risk set*.[7] We can focus on the risk set because every point in $[R]$ corresponds to at least one choice of $\boldsymbol{w} \in \mathbb{S}^{m-1}$.

Only a very small subset of the risk set will be admissible. A point $\boldsymbol{r} \in [R]$ is admissible exactly when it is on the lower boundary of $[R]$. More formally, define the 'quantant' of $\boldsymbol{r}$ to be the set

$$Q(\boldsymbol{r}) := \left\{\boldsymbol{x} \in \mathbb{R}^k : \boldsymbol{x} \leq \boldsymbol{r}\right\}$$

(see footnote 4). By definition, $\boldsymbol{r}$ is dominated by every $\boldsymbol{r}'$ for which $\boldsymbol{r}' \in Q(\boldsymbol{r}) \setminus \{\boldsymbol{r}\}$. So $\boldsymbol{r} \in [R]$ is admissible exactly when $[R] \cap Q(\boldsymbol{r}) = \{\boldsymbol{r}\}$. The set of $\boldsymbol{r}$ for satisfying this condition is the lower boundary of $[R]$, denoted $\lambda(R)$.

[7] If $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}$ are m points in $\mathbb{R}^k$, then the convex hull of these points is the set of $\boldsymbol{x} \in \mathbb{R}^k$ for which $\boldsymbol{x} = w_1 \boldsymbol{x}^{(1)} + \cdots + w_m \boldsymbol{x}^{(m)}$ for some $\boldsymbol{w} \in \mathbb{S}^{m-1}$.

Now we have to show that every point in $\lambda(R)$ is a Bayes rule for some $\boldsymbol{\pi} \gg \mathbf{0}$. For this we use the SHT, the proof of which can be found in any book on convex analysis (e.g., Çınlar and Vanderbei, 2013).

Theorem 3.5 (Supporting Hyperplane Theorem, SHT). *Let $[R]$ be a convex set in $\mathbb{R}^k$, and let $\boldsymbol{r}$ be a point on the boundary of $[R]$. Then there exists an $\boldsymbol{a} \in \mathbb{R}^k$ not equal to $\mathbf{0}$ such that*

$$\boldsymbol{a}^T \boldsymbol{r} = \min_{\boldsymbol{r}' \in [R]} \boldsymbol{a}^T \boldsymbol{r}'.$$

So let $\boldsymbol{r} \in \lambda(R)$ be any admissible risk. Let $\boldsymbol{a} \in \mathbb{R}^k$ be the coefficients of its supporting hyperplane. Because $\boldsymbol{r}$ is on the lower boundary of $[R]$, $\boldsymbol{a} \gg \mathbf{0}$.[8] Set

$$\pi_j := \frac{a_j}{\sum_{j'} a_{j'}} \qquad j = 1, \dots, k,$$

so that $\boldsymbol{\pi} \in \mathbb{S}^{k-1}$ and $\boldsymbol{\pi} \gg \mathbf{0}$. Then the SHT asserts that

$$\textstyle\sum_j r_j \cdot \pi_j \leq \sum_j r'_j \cdot \pi_j \quad \text{for all } \boldsymbol{r}' \in [R]. \tag{‡}$$

Let $\boldsymbol{w}$ be any randomised strategy with risk $\boldsymbol{r}$. Since $\sum_j r_j \cdot \pi_j$ is the expected loss of $\boldsymbol{w}$ (see †), (‡) asserts that $\boldsymbol{w}$ is a Bayes rule for prior distribution $\boldsymbol{\pi}$. Because $\boldsymbol{r}$ was an arbitrary point on $\lambda(R)$, and hence an arbitrary admissible rule, this completes the proof of Theorem 3.4.

3.4 *Point estimation*

For point estimation the action space is $\mathcal{A} = \Omega$, and the loss function $L(\theta, \theta')$ represents the (negative) consequence of choosing θ as a point estimate of Θ, when in fact $\Theta = \theta'$.

There will be situations where an obvious loss function $L : \Omega \times \Omega \to \mathbb{R}$ presents itself. But not very often. Hence the need for a generic loss function which is acceptable over a wide range of situations. A natural choice in the very common case where Ω is a convex subset of $\mathbb{R}^d$ is a *convex loss function*,[9]

$$L(\theta, \theta') = h(\theta - \theta') \tag{3.3}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is a smooth non-negative convex function with $h(\mathbf{0}) = 0$. This type of loss function asserts that small errors are much more tolerable than large ones. One possible further restriction would be that h is an even function.[10] This would assert that under-prediction incurs the same loss as over-prediction. There are many situations where this is *not* appropriate, but in these cases a generic loss function should be replaced by a more specific one.

Proceeding further along the same lines, an even, differentiable and strictly convex loss function can be approximated by a *quadratic loss function*,

$$h(\boldsymbol{x}) \propto \boldsymbol{x}^T Q \, \boldsymbol{x} \tag{3.4}$$

[8] Proof: because if $\boldsymbol{r}$ is on the lower boundary, the slightest decrease in any component of $\boldsymbol{r}$ must move $\boldsymbol{r}$ outside $[R]$.

[9] If Ω is convex then it is uncountable, and hence definitely not finite. But this does not have any disturbing implications for the following analysis.

[10] I.e. $h(-\boldsymbol{x}) = h(\boldsymbol{x})$.

where Q is a symmetric positive-definite $d \times d$ matrix. This follows directly from a Taylor series expansion of h around $\mathbf{0}$:

$$h(\boldsymbol{x}) = 0 + 0 + \tfrac{1}{2}\boldsymbol{x}^T \nabla^2 h(\mathbf{0})\, \boldsymbol{x} + 0 + O(\|\boldsymbol{x}\|^4)$$

where the first 0 is because $h(\mathbf{0}) = 0$, the second 0 is because $\nabla h(\mathbf{0}) = 0$ since h is minimized at $\boldsymbol{x} = \mathbf{0}$, and the third 0 is because h is an even function. $\nabla^2 h$ is the *hessian matrix* of second derivatives, and it is symmetric by construction, and positive definite at $\boldsymbol{x} = \mathbf{0}$, if h is strictly convex and minimized at $\mathbf{0}$.

In the absence of anything more specific the quadratic loss function is the generic loss function for point estimation. Hence the following result is widely applicable.

Theorem 3.6. *Under a quadratic loss function, the Bayes rule for point prediction is the conditional expectation*

$$\delta^*(y) = \mathrm{E}(\Theta \mid Y = y).$$

A Bayes rule for a point estimation is known as a *Bayes estimator*. Note that although the matrix Q is involved in defining the quadratic loss function in (3.4), it does not influence the Bayes estimator. Thus the Bayes estimator is the same for an uncountably large class of loss functions. Depending on your point of view, this is either its most attractive or its most disturbing feature.

Proof. Here is a proof that does not involve differentiation. The BRT (Theorem 3.1) asserts that

$$\delta^*(y) = \operatorname*{argmin}_{t \in \Omega} \mathrm{E}\{L(t, \Theta) \mid Y = y\}. \tag{3.5}$$

So let $\psi(y) := \mathrm{E}(\Theta \mid Y = y)$. For simplicity, treat θ as a scalar. Then

$$\begin{aligned} L(t, \theta) &\propto (t - \theta)^2 \\ &= (t - \psi(y) + \psi(y) - \theta)^2 \\ &= (t - \psi(y))^2 + 2(t - \psi(y))(\psi(y) - \theta) + (\psi(y) - \theta)^2. \end{aligned}$$

Take expectations conditional on $Y = y$ to get

$$\mathrm{E}\{L(t, \Theta) \mid Y = y\} \propto (t - \psi(y))^2 + \mathrm{E}\{(\psi(y) - \theta)^2 \mid Y = y\}. \tag{†}$$

Only the first term contains t, and this term is minimized over t by setting $t \leftarrow \psi(y)$, as was to be shown.

The extension to vector θ with loss function (3.4) is straightforward, but involves more ink. It is crucial that Q in (3.4) is positive definite, because otherwise the first term in (†), which becomes $(t - \psi(y))^T Q\, (t - \psi(y))$, is not minimized if and only if $t = \psi(y)$. □

Note that the same result holds in the more general case of a point prediction of random quantities X based on observables Y: under quadratic loss, the Bayes estimator is $\mathrm{E}(X \mid Y = y)$.

* * *

Now apply the CCT (Theorem 3.2) to this result. For quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some prior distribution $\pi \gg \mathbf{0}$.[11] Among the casualties of this conclusion is the Maximum Likelihood Estimator (MLE),

$$\hat{\theta}(y) := \operatorname*{argsup}_{\theta \in \Omega} f(y; \theta).$$

Stein's paradox showed that under quadratic loss, the MLE is not always admissible in the case of a Multinormal distribution with known variance, by producing an estimator which dominated it. This result caused such consternation when first published that it might be termed 'Stein's bombshell'. See Efron and Morris (1977) for more details, and Samworth (2012) for an accessible proof. Persi Diaconis thought this was such a powerful result that he focused on it for his brief article on Mathematical Statistics in the *The Princeton Companion to Mathematics* (Ed. T. Gowers, 2008, 1056 pages). Interestingly, the MLE is still the dominant point estimator in applied statistics, even though its admissibility under quadratic loss is questionable.

3.5 *Set estimation*

For set estimation the action space is $\mathcal{A} = 2^{\Omega}$, and the loss function $L(C, \theta)$ represents the (negative) consequences of choosing $C \subset \Omega$ as a set estimate of Θ, when the true value of Θ is θ.

There are two contradictory requirements for set estimators of Θ. We want the sets to be small, but we also want them to contain Θ. There is a simple way to represent these two requirements as a loss function, which is to use

$$L(C, \theta) = |C| + \kappa \cdot (1 - \mathbb{1}_{\theta \in C}) \quad \text{for some } \kappa > 0 \tag{3.6a}$$

where $|C|$ is the cardinality of C.[12] The value of κ controls the trade-off between the two requirements. If $\kappa \downarrow 0$ then minimizing the expected loss will always produce the empty set. If $\kappa \uparrow \infty$ then minimizing the expected loss will always produce Ω. For κ in-between, the outcome will depend on beliefs about Y and the value y.

It is important to note that the crucial result, Theorem 3.7 below, continues to hold for the much more general set of loss functions

$$L(C, \theta) = g(|C|) + h(1 - \mathbb{1}_{\theta \in C}) \tag{3.6b}$$

where g is non-decreasing and h is strictly increasing. This is a large set of loss functions, which should satisfy most statisticians who do not have a specific loss function already in mind.

For point estimators there was a simple characterisation of the Bayes rule for quadratic loss functions (Theorem 3.6). For set estimators the situation is not so simple. However, for loss functions of the form (3.6) there is a simple necessary condition for a rule to be a Bayes rule.

[11] This is under the conditions of Theorem 3.2, or with appropriate extensions of them in the non-finite cases.

[12] Here and below I am treating Ω as countable, for simplicity; otherwise $|\bullet|$ would denote volume.

Theorem 3.7 (Level set property, LSP). *Say that $C : \mathcal{Y} \to 2^{\Omega}$ has the 'level set property' exactly when $C(y)$ is a subset of a level set of $\pi^*(y)$ for every y.*[13] *If C is a Bayes rule for the loss function in* (3.6a), *then it has the level set property.*

Proof. Let 'BR' denote 'C is a Bayes rule for (3.6a)' and let 'LSP' denote 'C has the level set property'. The theorem asserts that BR → LSP, showing that LSP is a necessary condition for BR. We prove the theorem by proving the contra-positive, that ¬LSP → ¬BR.

¬LSP asserts that there is a y for which:

$$\exists \theta_j \in C, \exists \theta_{j'} \notin C \quad \text{such that} \quad \pi^*_{j'} > \pi^*_j,$$

where I have suppressed the y argument on C and $\boldsymbol{\pi}^*$. For this y, j, and j', let $C' \subset \Omega$ be the same as C, except with θ_j swapped for $\theta_{j'}$. In this case $|C'| = |C|$, but

$$\Pr(\Theta \notin C \mid Y = y) > \Pr(\Theta \notin C' \mid Y = y).$$

Hence

$$\begin{aligned} \mathrm{E}\{L(C, \Theta) \mid Y = y\} &= |C| + \kappa \cdot \Pr(\Theta \notin C \mid Y = y) \\ &> |C'| + \kappa \cdot \Pr(\Theta \notin C' \mid Y = y) \\ &= \mathrm{E}\{L(C', \Theta) \mid Y = y\}, \end{aligned}$$

i.e.

$$C \neq \operatorname*{argmin}_{C'} \mathrm{E}\{L(C', \Theta) \mid Y = y\}$$

which shows that C is not a Bayes rule, by the BRT (Theorem 3.1). □

Now relate this result to the CCT (Theorem 3.2). First, Theorem 3.7 asserts that C having the LSP is necessary (but not sufficient) for C to be a Bayes rule for loss functions of the form (3.6a). Second, the CCT asserts that being a Bayes rule is a necessary (but not sufficient) condition for C to be admissible.[14] So unless C has the LSP then it is impossible for C to be admissible for loss functions of the form (3.6a). Bayesian HPD regions (see eq. 2.11) satisfy this necessary condition for admissibility.

Things are trickier for Frequentist set estimators, which must proceed without a prior distribution $\boldsymbol{\pi}$, and thus cannot compute $\boldsymbol{\pi}^*(y)$. But, at least in the case where Ω is finite (and more generally when it is bounded) a prior of $\pi_j \propto 1$ would imply that $\pi^*_j(y) \propto f(y; \theta_j)$, by Bayes's Theorem. So in this case levels sets of $f(y; \bullet)$ would also be level sets of $\boldsymbol{\pi}^*(y)$, and hence would satisfy the necessary condition for admissibility. So my strong recommendation for Frequentist set estimators is

- In the absence of a prior distribution, base set estimators on level sets of $f(y; \bullet)$, i.e.

  $$C(y) = \{\theta : f(y; \theta) \geq k(y)\}$$

  for some $k > 0$ which may depend on y.

These are effectively Wilks set estimators from Section 2.6.1. I will be adopting this recommendation in Chapter 4.

[13] Dropping the y argument, C is a level set of $\boldsymbol{\pi}^*$ exactly when $C = \{\theta_j : \pi^*_j \geq k\}$ for some k.

[14] As before, terms and conditions apply in the non-finite cases.

3.6 Hypothesis tests

For hypothesis tests, the action space is a partition of Ω, denoted

$$\mathcal{H} := \{H_0, H_1, \dots, H_d\}.$$

Each element of $\mathcal{H}$ is termed a *hypothesis*; it is traditional to number the hypotheses from zero. The loss function $L(H_i, \theta)$ represents the (negative) consequences of choosing element H_i, when the true value of Θ is θ. It would be usual for the loss function to satisfy

$$\theta \in H_i \implies L(H_i, \theta) = \min_{i'} L(H_{i'}, \theta)$$

on the grounds that an incorrect choice of element should never incur a smaller loss than the correct choice.

I will be quite cavalier about hypothesis tests. If the statistician has a complete loss function, then the CCT (Theorem 3.2) applies, a $\boldsymbol{\pi} \gg \mathbf{0}$ must be found, and there is nothing more to be said. The famous *Neyman-Pearson (NP) Lemma* is of this type. It has $\Omega = \{\theta_0, \theta_1\}$, with $H_i = \{\theta_i\}$, and loss function

L	θ_0	θ_1
H_0	0	ℓ_1
H_1	ℓ_0	0

with $\ell_0, \ell_1 > 0$. The NP Lemma asserts that a decision rule for choosing between H_0 and H_1 is admissible if and only if it has the form

$$\frac{f(y;\theta_0)}{f(y;\theta_1)} \begin{cases} < c & \text{choose } H_1 \\ = c & \text{toss a coin} \\ > c & \text{choose } H_0 \end{cases}$$

for some $c > 0$. This is just the CCT (Theorem 3.2).[15]

The NP Lemma is particularly simple, corresponding to a choice in a family with only two elements. In situations more complicated than this, it is extremely challenging and time-consuming to specify a loss function. And yet statisticians would still like to choose between hypotheses, in decision problems whose outcome does not seem to justify the effort required to specify the loss function.[16]

There is a generic loss function for hypothesis tests, but it is hardly defensible. The *0-1 ('zero-one') loss function* is

$$L(H_i, \theta) = 1 - \mathbb{1}_{\theta \in H_i},$$

i.e., zero if θ is in H_i, and one if it is not. Its Bayes rule is to select the hypothesis with the largest conditional probability. It is hard to think of a reason why the 0-1 loss function would approximate a wide range of actual loss functions, unlike in the cases of generic loss functions for point estimation and set estimation. This is not to say that it is wrong to select the hypothesis with the largest conditional probability; only that the 0-1 loss function does not provide a very compelling reason.

[15] In fact, $c = (\pi_1/\pi_0) \cdot (\ell_1/\ell_0)$, where (π_0, π_1) is the prior probability for which $\pi_1 = 1 - \pi_0$.

[16] Just to be clear, *important* decisions should not be based on cut-price procedures: an important decision warrants the effort required to specify a loss function.

* * *

There is another approach which has proved much more popular. In fact, it is the dominant approach to hypothesis testing. This is to co-opt the theory of set estimators, for which there *is* a defensible generic loss function, which has strong implications for the selection of decision rules (see Section 3.5). The statistician can use her set estimator $C : \mathcal{Y} \to 2^{\Omega}$ to make at least some distinctions between the members of $\mathcal{H}$, on the basis of the value of the observable, y^{obs}:

- 'Accept' H_i exactly when $C(y^{\text{obs}}) \subset H_i$,
- 'Reject' H_i exactly when $C(y^{\text{obs}}) \cap H_i = \emptyset$,
- 'Undecided' about H_i otherwise.

Note that these three terms are given in scare quotes, to indicate that they acquire a technical meaning in this context. We do not use the scare quotes in practice, but we always bear in mind that we are not "accepting H_i" in the vernacular sense, but simply asserting that $C(y^{\text{obs}}) \subset H_i$ for our particular choice of δ.

Looking at the three options above, there are two classes of outcome. If we accept H_i then we must reject all of the other hypotheses. But if we are undecided about H_i then we cannot accept any hypothesis. One very common case is where $\mathcal{H} = \{H_0, H_1\}$, where H_0 is the *null hypothesis* and H_1 is the *alternative hypothesis*. There are two versions. In the first, known as a *two-sided test* (or 'two-tailed test'), H_0 is a tiny subset of Ω, too small for $C(y^{\text{obs}})$ to get inside. Therefore it is impossible to accept H_0, and all that we can do is reject H_0 and accept H_1, or be undecided. In the second case, known as a *one-sided test* (or 'one-tailed test'), H_0 is a sizeable subset of Ω, and then it is possible to accept H_0 and reject H_1.

For example, suppose that the model is $Y \sim \text{Norm}(\mu, \sigma^2)$, for which $\theta = (\mu, \sigma^2) \in \mathbb{R}_{++} \times \mathbb{R}_{++}$. Consider two different tests:

Test A	Test B
$H_0 : \kappa = c$	$H_0 : \kappa \geq c$
$H_1 : \kappa \neq c$	$H_1 : \kappa < c$

where $\kappa := \sigma/\mu \in \mathbb{R}_{++}$, known as the 'coefficient of variation', and c is some specified constant. Test A is a two-sided test, in which it is impossible to accept H_0, and so there are only two outcomes: to reject H_0, or to be undecided, which is usually termed 'fail to reject H_0'. Test B is a one-sided test in which we can accept H_0 and reject H_1, or accept H_1 and reject H_0, or be undecided.

In applications we usually want to do a one-sided test. For example, if μ is the performance of a new treatment relative to a control, then we can be fairly sure *a priori* that $\mu = 0$ is false: different treatments seldom have identical effects. What we want to know is whether the new treatment is worse or better than the control: i.e. we want $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. In this case we can find in favour of H_0, or in favour of H_1, or be undecided. In a one-sided test, it would be sensible to push the upper bound of H_0 above $\mu = 0$ to some value $\mu_0 > 0$, which is the *minimial clinically significant difference (MCSD)*.

Hypothesis testing is practiced mainly by Frequentist statisticians, and so I will continue in a Frequentist vein. In the Frequentist approach, it is conventional to use a 95% confidence set as the set estimator for hypothesis testing. Other levels, notably 90% and 99%, are occasionally used. If H_0 is rejected using a 95% confidence set, then this is reported as "H_0 is rejected at a significance level of 5%" (occasionally 10% or 1%). Confidence sets are covered in detail in Chapter 4.

This confidence set approach to hypothesis testing seems quite clear-cut, but we must end on a note of caution. First, the statistician has not solved the decision problem of choosing an element of $\mathcal{H}$. She has solved a different problem. Based on a set estimator, she may reject H_0 on the basis of y^{obs}, but that does not mean she should proceed as though H_0 is false. This would require her to solve the correct decision problem, for which she would have to supply a loss function. So, first caution:

- Rejecting H_0 is not the same as deciding that H_0 is false. Hypothesis tests do not solve decision problems.

Second, loss functions of the form (3.6) may be generic, but that does not mean that there is only one 95% confidence procedure. As Chapter 4 will show, there are an uncountable number of ways of constructing a 95% confidence procedure. In fact, there are an uncountable number of ways of constructing a 95% confidence procedure based on level sets of $f(y; \bullet)$. So the statistician still needs to make and to justify two subjective choices, leading to the second caution:

- Accepting or rejecting a hypothesis is contingent on the choice of confidence procedure, as well as on the level.

4
Confidence sets

This chapter is a continuation of Chapter 3, and the same conditions hold; re-read the introduction to Chapter 3 if necessary.

In this chapter we have the tricky situation in which a specified function $g : \mathcal{Y} \times \Omega \to \mathbb{R}$ becomes a random quantity when Y is a random quantity. Then the distribution of $g(Y, \theta)$ depends on the value in Ω controlling the distribution of Y, which need not be the same value as θ in the argument. However, in this chapter the value in Ω controlling the distribution of Y will always be the same value as θ. Hence $g(Y, \theta)$ has the distribution induced by $Y \sim f(\bullet; \theta)$.

4.1 *Confidence procedures and confidence sets*

A confidence procedure is a special type of decision rule for the problem of set estimation. Hence it is a function of the form $C : \mathcal{Y} \to 2^{\Omega}$, where 2^{Ω} is the set of all sets of Ω.[1] Decision rules for set estimators were discussed in Section 3.5. A confidence set is *not* a Bayes Rule for the loss function in (3.6a).

Definition 4.1 (Confidence procedure). $C : \mathcal{Y} \to 2^{\Omega}$ is a level-$(1-\alpha)$ confidence procedure exactly when

$$\Pr\{\theta \in C(Y); \theta\} \geq 1 - \alpha \quad \text{for all } \theta \in \Omega.$$

If the probability equals $(1-\alpha)$ for all θ, then C is an *exact* level-$(1-\alpha)$ confidence procedure.[2]

The value $\Pr\{\theta \in C(Y); \theta\}$ is termed the *coverage* of C at θ. Thus a 95% confidence procedure has coverage of at least 95% for all θ, and an exact 95% confidence procedure has coverage of exactly 95% for all θ. The diameter of $C(y)$ can grow rapidly with its coverage.[3] In fact, the relation must be extrememly convex when coverage is nearly one, because, in the case where $\Omega = \mathbb{R}$, the diameter at coverage $= 1$ is unbounded. So an increase in the coverage from, say 95% to 99%, could correspond to a doubling or more of the diameter of the confidence procedure. For this reason, exact confidence procedures are highly valued, because a conservative 95% confidence procedure can deliver sets that are much larger than an exact one.

[1] In this chapter I am using 'C' for a confidence procedure, rather than 'δ' for a decision rule.

[2] Exact is a special case. But when it necessary to emphasize that C is not exact, the term 'conservative' is used.

[3] The diameter of a set in a metric space such as Euclidean space is the maximum of the distance between two points in the set.

But, immediately a note of caution. It seems obvious that exact confidence procedures should be preferred to conservative ones, but this is easily exposed as a mistake. Suppose that $\Omega = \mathbb{R}$. Then the following procedure is an exact level-$(1 - \alpha)$ confidence procedure for θ. First, draw a random variable U with a standard uniform distribution.[4] Then set

$$C(y) := \begin{cases} \mathbb{R} & U \leq 1 - \alpha \\ \{0\} & \text{otherwise.} \end{cases} \tag{\dagger}$$

This is an exact level-$(1 - \alpha)$ confidence procedure for θ, but also a meaningless one because it does not depend on y. If it is objected that this procedure is invalid because it includes an auxiliary random variable, then this rules out the method of generating approximately exact confidence procedures using bootstrap calibration (Section 4.3.3). And if it is objected that confidence procedures must depend on y, then (†) could easily be adapted so that y is the seed of a numerical random number generator for U. So something else is wrong with (†). In fact, it fails a necessary condition for admissibility that was derived in Section 3.5. This will be discussed in Section 4.2.

It is helpful to distinguish between the confidence procedure C, which is a function of y, and the result when C is evaluated at the observations y^{obs}, which is a set in Ω. I like the terms used in Morey et al. (2016), which I will also adapt to p-values in Section 4.5.

Definition 4.2 (Confidence set)**.** $C(y^{\text{obs}})$ is a level-$(1 - \alpha)$ confidence set exactly when C is a level-$(1 - \alpha)$ confidence procedure.

So a confidence procedure is a function, and a confidence set is a set. If $\Omega \subset \mathbb{R}$ and $C(y^{\text{obs}})$ is convex, i.e. an interval, then a confidence set (interval) is represented by a lower and upper value. We should write, for example, "using procedure C, the 95% confidence interval for θ is $[0.55, 0.74]$", inserting "exact" if the confidence procedure C is exact.

4.2 *Families of confidence procedures*

The challenge with confidence procedures is to construct one with a specified level (look back to Section 1.4). One could propose an arbitrary $C : \mathcal{Y} \to 2^{\Omega}$, and then laboriously compute the coverage for every $\theta \in \Omega$. At that point one would know the level of C as a confidence procedure, but it is unlikely to be 95%; adjusting C and iterating this procedure many times until the minimum coverage was equal to 95% would be exceedingly tedious. So we need to go backwards: start with the level, e.g. 95%, then construct a C guaranteed to have this level.

Define a *family of confidence procedures* as $C : \mathcal{Y} \times [0, 1] \to 2^{\Omega}$, where $C(\cdot; \alpha)$ is a level-$(1 - \alpha)$ confidence procedure for each α. If we start

[4] See footnote 6.

with a family of confidence procedures for a specified model, then we can compute a confidence set for any level we choose.

One class of families of confidence procedures has a natural and convenient form. The key concept is *stochastic dominance*. Let X and Y be two scalar random quantities. Then X stochastically dominates Y exactly when

$$\Pr(X \leq v) \leq \Pr(Y \leq v) \quad \text{for all } v \in \mathbb{R}.$$

Visually, the distribution function for X is never to the left of the distribution function for Y.[5] Although it is not in general use, I define the following term.

Definition 4.3 (Super-uniform)**.** The random quantity X is *super-uniform* exactly when it stochastically dominates a standard uniform random quantity.[6]

In other words, X is super-uniform exactly when $\Pr(X \leq u) \leq u$ for all $0 \leq u \leq 1$. Note that if X is super-uniform then its support is bounded below by 0, but not necessarily bounded above by 1. Now here is a representation theorem for families of confidence procedures.[7]

Theorem 4.1 (Families of Confidence Procedures, FCP)**.** *Let $g : \mathcal{Y} \times \Omega \to \mathbb{R}$. Then*

$$C(y;\alpha) := \big\{\theta \in \Omega : g(y,\theta) > \alpha\big\} \tag{4.1}$$

is a family of level-$(1-\alpha)$ confidence procedures if and only if $g(Y,\theta)$ is super-uniform for all $\theta \in \Omega$. C is exact if and only if $g(Y,\theta)$ is uniform for all θ.

Proof.
($\Leftarrow$). Let $g(Y,\theta)$ be super-uniform for all θ. Then, for arbitrary θ,

$$\begin{aligned}\Pr\{\theta \in C(Y;\alpha);\theta\} &= \Pr\{g(Y,\theta) > \alpha;\theta\}\\ &= 1 - \Pr\{g(Y,\theta) \leq \alpha;\theta\}\\ &= 1 - (\leq \alpha) \geq 1 - \alpha\end{aligned}$$

as required. For the case where $g(Y,\theta)$ is uniform, the inequality is replaced by an equality.

($\Rightarrow$). This is basically the same argument in reverse. Let $C(\cdot;\alpha)$ defined in (4.1) be a level-$(1-\alpha)$ confidence procedure. Then, for arbtrary θ,

$$\Pr\{g(Y,\theta) > \alpha;\theta\} \geq 1 - \alpha.$$

Hence $\Pr\{g(Y,\theta) \leq \alpha;\theta\} \leq \alpha$, showing that $g(Y,\theta)$ is super-uniform as required. Again, if $C(\cdot;\alpha)$ is exact, then the inequality is replaced by a equality, and $g(Y,\theta)$ is uniform. □

Families of confidence procedures have the very intuitive *nesting property*, that

$$\alpha < \alpha' \implies C(y;\alpha) \supset C(y;\alpha'). \tag{4.2}$$

[5] Recollect that the distribution function of X has the form $F(x) := \Pr(X \leq x)$ for $x \in \mathbb{R}$.

[6] A standard uniform random quantity being one with distribution function $F(u) = \max\{0, \min\{u, 1\}\}$.

[7] Look back to 'New notation' at the start of the Chapter for the definition of $g(Y;\theta)$.

In other words, higher-level confidence sets are always supersets of lower-level confidence sets from the same family. This has sometimes been used as part of the definition of a family of confidence procedures (see, e.g., Cox and Hinkley, 1974, ch. 7), but I prefer to see it as a consequence of a construction such as (4.1).

* * *

Section 3.5 made a recommendation about set estimators for θ, which was that they should be based on level sets of $f(y; \bullet)$. This was to satisfy a necessary condition to be admissible under the loss function (3.6). I call this the *Level Set Property (LSP)*. A family of confidence procedures does not necessarily have the LSP. So it is not obvious, but highly gratifying, that it is possible to construct families of confidence procedures with the LSP. Three different approaches are given in the next section.

4.3 *Methods for constructing confidence procedures*

All three of these methods produce families of confidence procedures with the LSP. This is a long section, and there is a summary in Section 4.3.4.

4.3.1 *Markov's inequality*

Here is a result that has pedagogic value, because it can be used to generate an uncountable number of families of confidence procedures, each with the LSP.

Theorem 4.2. *Let h be any PMF for Y. Then*

$$C(y; \alpha) := \big\{\theta \in \Omega : f(y, \theta) > \alpha \cdot h(y)\big\} \tag{4.3}$$

is a family of confidence procedures, with the LSP.

Proof. Define $g(y, \theta) := f(y; \theta)/h(y)$, which may be ∞. Then the result follows immediately from Theorem 4.1 because $g(Y, \theta)$ is super-uniform for each θ:

$$\begin{aligned}\Pr\{f(Y;\theta)/h(Y) \le u; \theta\} &= \Pr\{h(Y)/f(Y;\theta) \ge 1/u; \theta\} \\ &\le \frac{\mathrm{E}\{h(Y)/f(Y;\theta); \theta\}}{1/u} \qquad \text{Markov's inequality} \\ &\le \frac{1}{1/u} = u.\end{aligned}$$

For the final inequality,

$$\begin{aligned}\mathrm{E}\{h(Y)/f(Y;\theta); \theta\} &= \sum_{y \in \operatorname{supp} f(\bullet;\theta)} \frac{h(y)}{f(y;\theta)} \cdot f(y;\theta) \\ &= \sum_{y \in \operatorname{supp} f(\bullet;\theta)} h(y) \\ &\le 1.\end{aligned}$$

If $\operatorname{supp} h \subset \operatorname{supp} f(\bullet; \theta)$, then this inequality is an equality. □

Among the interesting choices for g, one possibility is $g = f(\bullet;\theta)$, for some $\theta \in \Omega$. Note that with this choice, the confidence set of (4.3) always contains θ. So we know that we can construct a level-$(1-\alpha)$ confidence procedure whose confidence sets will always contain θ, for any $\theta \in \Omega$.

This is another illustration of the fact that the definition of a confidence procedure given in Definition 4.1 is too broad to be useful. But now we see that insisting on the LSP is not enough to resolve the issue. Two statisticians can both construct 95% confidence sets for θ which satisfy the LSP, using different families of confidence procedures. Yet the first statistician may reject the null hypothesis that $H_0 : \Theta = \theta_0$ (see Section 3.6), and the second statistician may fail to reject it, for any $\theta_0 \in \Omega$.

Actually, the situation is not as grim as it seems. Markov's inequality is very slack, and so the coverage of the family of confidence procedures defined in Theorem 4.2 is likely to be much larger than $(1-\alpha)$, e.g. much larger than 95%. Remembering the comment about the rapid increase in the diameter of the confidence set as the coverage increases, from Section 4.1, a more likely outcome is that $C(y; 0.05)$ is large for many different choices of h, in which case no one rejects the null hypothesis.

All in all, it would be much better to use an exact family of confidence procedures, if one existed. And, for perhaps the most popular model in the whole of Statistics, this is the case.

4.3.2 The Linear Model

The Linear Model (LM) can be expressed as

$$Y \overset{\text{D}}{=} X\beta + \epsilon \quad \text{where } \epsilon \sim \text{N}_n(\mathbf{0}, \sigma^2 I_n) \tag{4.4}$$

where Y is an n-vector of observables, X is a specified $n \times p$ matrix of *regressors*, β is a p-vector of *regression coefficients*, and ϵ is an n-vector of *residuals*.[8] The parameter is $\theta = (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_{++}$, and where it is necessary to refer to the true parameter value I would use (Θ_1, Θ_2).

'$\text{N}_n(\cdot)$' denotes the n-dimensional *Multinormal distribution* with specified expectation vector and variance matrix (see, e.g., Mardia et al., 1979, ch. 3). The symbol '$\overset{\text{D}}{=}$' denotes 'equal in distribution'; this notation is useful here because the Multinormal distribution is closed under affine transformations. Hence Y has a Multinormal distribution, because it is an affine transformation of ϵ. So the LM must be restricted to applications for which Y can be thought of, at least approximately, as a collection of n random quantities each with realm $\mathbb{R}$, and for each of which our uncertainty is approximately symmetric. Many observables fail to meet these necessary conditions (e.g. applications in which Y is a collection of counts); for these applications, we have *Generalized Linear Models (GLMs)*. GLMs retain many of the attractive properties of LMs.

[8] Usually I would make Y and ϵ bold, being vectors, and I would prefer not to use X for a specified matrix, but this is the standard notation.

Wood (2015, ch. 7) provides an insightful summary of the LM, while Draper and Smith (1998) give many practical details.

Now I show that the Maximum Likelihood Estimator (MLE) of (4.4) is

$$\begin{aligned} \hat{\beta}(y) &= (X^T X)^{-1} X^T y \\ \widehat{\sigma^2}(y) &= n^{-1}(y - \hat{y})^T (y - \hat{y}) \end{aligned}$$

where $\hat{y} := X\hat{\beta}(y)$.

Proof. For a LM, it is more convenient to minimise $-2 \log f(y; \beta, \sigma^2)$ over (β, σ^2) than to maximise $f(y; \beta, \sigma^2)$. Then

$$-2 \log f(y; \beta, \sigma^2) = n \log(2\pi\sigma^2) + \frac{1}{\sigma^2}(y - X\beta)^T (y - X\beta)$$

from the PDF of the Multinormal distribution. Now use a simple device to show that this is minimised at $\beta = \hat{\beta}(y)$ for all values of σ^2. I will write $\hat{\beta}$ rather than $\hat{\beta}(y)$:

$$\begin{aligned} &(y - X\beta)^T (y - X\beta) \\ &\quad = (y - X\hat{\beta} + X\hat{\beta} - X\beta)^T (y - X\hat{\beta} + X\hat{\beta} - X\beta) \\ &\quad = (y - \hat{y})^T (y - \hat{y}) + 0 + (X\hat{\beta} - X\beta)^T (X\hat{\beta} - X\beta) \qquad (\dagger) \end{aligned}$$

where multiplying out shows that the cross-product term in the middle is zero. Only the final term contains β. Writing this term as

$$(\hat{\beta} - \beta)^T (X^T X)(\hat{\beta} - \beta)$$

shows that if X has full column rank, so that $X^T X$ is positive definite, then (†) is minimised if and only if $\beta = \hat{\beta}$. Then

$$-2 \log f(y; \hat{\beta}, \sigma^2) = n \log(2\pi\sigma^2) + \frac{1}{\sigma^2}(y - \hat{y})^T (y - \hat{y}).$$

Solving the first-order condition gives the MLE for $\widehat{\sigma^2}(y)$, and it is easily checked that this is a global minimum. □

Now suppose we want a confidence procedure for β. For simplicity, I will assume that σ^2 is specified, and for practical purposes I would replace it by $\widehat{\sigma^2}(y^{\text{obs}})$ in calculations. This is known as *plugging in* for σ^2. The LM extends to the case where σ^2 is not specified, but, as long as $n/(n-p) \approx 1$, it makes little difference in practice to plug in.[9]

With β representing an element of the β-parameter space $\mathbb{R}^p$, and σ^2 specified, we have, from the results above,

$$-2 \log \left(\frac{f(y; \beta, \sigma^2)}{f(y; \hat{\beta}(y), \sigma^2)} \right) = \frac{1}{\sigma^2} \{\hat{\beta}(y) - \beta\}^T (X^T X) \{\hat{\beta}(y) - \beta\}. \quad (4.5)$$

Now suppose we could prove the following.

Theorem 4.3. *With σ^2 specified,*

$$\frac{1}{\sigma^2} \{\hat{\beta}(Y) - \beta\}^T (X^T X) \{\hat{\beta}(Y) - \beta\}$$

has a χ^2_p distribution.

[9] As an eminent applied statistician remarked to me: it if matters to your conclusions whether you use a standard Normal distribution or a Student-*t* distribution, then you probability have bigger things to worry about. This is good advice.

We could define the decision rule:

$$C(y;\alpha) := \left\{ \beta \in \mathbb{R}^p : -2\log\left(\frac{f(y;\beta,\sigma^2)}{f(y;\hat{\beta}(y),\sigma^2)}\right) < \chi_p^{-2}(1-\alpha) \right\}. \tag{4.6}$$

where $\chi_p^{-2}(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the χ_p^2 distribution. Under Theorem 4.3, (4.5) shows that C in (4.6) would be an exact level-$(1-\alpha)$ confidence procedure for β; i.e. it provides a family of exact confidence procedures. Also note that it satisfies the LSP.

After that build-up, it will come as no surprise to find out that Theorem 4.3 is true. Substituting Y for y in the MLE of β gives

$$\hat{\beta}(Y) \overset{\text{D}}{=} (X^T X)^{-1} X^T (X\beta + \epsilon) \overset{\text{D}}{=} \beta + (X^T X)^{-1} X^T \epsilon,$$

writing σ for $\sqrt{\sigma^2}$. So the distribution of $\hat{\beta}(Y)$ is another Multinormal distribution

$$\hat{\beta}(Y) \sim \text{N}_p(\beta, \Sigma) \quad \text{where } \Sigma := \sigma^2 (X^T X)^{-1}.$$

Now apply a standard result for the Multinormal distribution to deduce

$$\{\hat{\beta}(Y) - \beta\}^T \Sigma^{-1} \{\hat{\beta}(Y) - \beta\}|_{\beta=\beta} \sim \chi_p^2 \tag{†}$$

(see Mardia et al., 1979, Thm 2.5.2). This proves Theorem 4.3 above. Let's celebrate this result!

Theorem 4.4. *For the LM with σ^2 specified, C defined in* (4.6) *is a family of exact confidence procedures for β, which has the LSP.*

Of course, when we plug-in for σ^2 we slightly degrade this result, but not by much if $n/(n-p) \approx 1$.

This happy outcome where we can find a family of exact confidence procedures with the LSP is more-or-less unique to the regression parameters in the LM. but it is found, approximately, in the large-n behaviour of a much wider class of models, including GLMs, as explained next.

4.3.3 Wilks confidence procedures

There is a beautiful theory which explains how the results from Section 4.3.2 generalise to a much wider class of models than the LM. The theory is quite strict, but it almost-holds over relaxations of some of its conditions. Stated informally, if $Y := (Y_1, \dots, Y_n)$ and

$$f(y;\theta) = \prod_{i=1}^{n} f_1(y_i;\theta) \tag{4.7}$$

and f_1 is a *regular model*, and the parameter space Ω is an open convex subset of $\mathbb{R}^p$ (and invariant to n), then

$$-2\log\left(\frac{f(Y;\theta)}{f(Y;\hat{\theta}(Y))}\right)\Bigg|_{\theta=\theta} \xrightarrow{\text{D}} \chi_p^2 \tag{4.8}$$

where $\hat{\theta}$ is the Maximum Likelihood Estimator (MLE) of θ, and '$\xrightarrow{\text{D}}$' denotes 'convergence in distribution' as n increases without bound. Eq. (4.8) is sometimes termed *Wilks's Theorem*, hence the name of this subsection.

The definition of 'regular model' is quite technical, but a working guideline is that f_1 must be smooth and differentiable in θ; in particular, $\operatorname{supp} Y_1$ must not depend on θ. Cox (2006, ch. 6) provides a summary of this result and others like it, and more details can be found in Casella and Berger (2002, ch. 10), or, for the full story, in van der Vaart (1998).

This result is true for the LM, because we showed that it is exactly true for any n provided that σ^2 is specified, and the ML plug-in for σ^2 converges on the true value as $n/(n-p) \to 1$.[10] In general, we can use it the same way as in the LM, to derive a decision rule:

$$C(y; \alpha) := \left\{ \theta \in \Omega : -2 \log \left(\frac{f(Y; \theta)}{f(Y; \hat{\theta}(Y))} \right) < \chi_p^{-2}(1-\alpha) \right\}. \tag{4.9}$$

As already noted, this C satisfies the LSP. Further, under the conditions for which (4.8) is true, C is also a family of approximately exact confidence procedures.

Eq. (4.9) can be written differently, perhaps more intuitively. Define

$$L(\bullet; y) := f(y; \bullet)$$

known as the *likelihood function* of θ; sometimes the y argument is suppressed, notably when $y = y^{\text{obs}}$. Let $\ell := \log L$, the *log-likelihood function*. Then (4.9) can be written

$$C(y; \alpha) = \left\{ \theta \in \Omega : \ell(\theta; y) > \ell(\hat{\theta}(y); y) - \kappa(\alpha) \right\} \tag{4.10}$$

where $\kappa(\alpha) := \chi_p^{-2}(1-\alpha)/2$. In this procedure we keep all $\theta \in \Omega$ whose log-likelihood values are within $\kappa(\alpha)$ of the maximum log-likelihood. In the common case where $\Omega \subset \mathbb{R}$, (4.10) gives *'Allan's Rule of Thumb'*:[11]

- For an approximate 95% confidence procedure for a scalar parameter, keep all values of $\theta \in \Omega$ for which the log-likelihood is within 2 of the maximum log-likelihood.

The value 2 is from $\chi_1^{-2}(0.95)/2 = 1.9207\ldots \approx 2$.

* * *

The pertinent question, as always with methods based on asymptotic properties for particular types of model, is whether the approximation is a good one. The crucial concept here is *level error*. The coverage that we want is at least $(1-\alpha)$ everywhere, which is termed the 'nominal level'. But were we to evaluate a confidence procedure such as (4.10) for a general model (not a LM) we would find that, over all $\theta \in \Omega$, that the minimum coverage was not $(1-\alpha)$ but something else; usually something less than $(1-\alpha)$. This is the 'actual level'. The difference is

$$\text{level error} := \text{nominal level} - \text{actual level}.$$

[10] This is a general property of the MLE, that it is *consistent* when f has the product form given in (4.7).

[11] After Allan Seheult, who first taught it to me.

Level error exists because the conditions under which (4.10) provides an exact confidence procedure are not met in practice, outside the LM. Although it is tempting to ignore level error, experience suggests that it can be large, and that we should attempt to correct for level error if we can.

One method for making this correction is *bootstrap calibration*, described in DiCiccio and Efron (1996). I used this method in Rougier et al. (2016); you will have to read the Appendix.

4.3.4 *Summary*

With the Linear Model (LM) described in Section 4.3.2, we can construct a family of exact confidence procedures, with the LSP, for the parameters β. Additionally—I did not show it but it follows directly—we can do the same for all affine functions of the parameters β, including individual components.

In general we are not so fortunate. It is not that we cannot construct families of confidence procedures with the LSP: Section 4.3.1 shows that we can, in an uncountable number of different ways. But their levels will be conservative, and hence they are not very informative. A better alternative, which ought to work well in large-n simple models like (4.7) is to use Wilks's Theorem to construct a family of approximately exact confidence procedures, which have the LSP, see Section 4.3.3.

The Wilks approximation can be checked and—one hopes—improved, using bootstrap calibration. Bootstrap calibration is a necessary precaution for small n or more complicated models (e.g. time series or spatial applications). But in these cases a Bayesian approach is likely to be a better choice, which is reflected in modern practice.

4.4 *Marginalisation*

Suppose that $g : \theta \mapsto \phi$ is some specified function, and we would like a confidence procedure for ϕ. If C is a level-$(1-\alpha)$ confidence procedure for ϕ then it must have ϕ-coverage of at least $(1-\alpha)$ for all $\theta \in \Omega$. The most common situation is where $\Omega \subset \mathbb{R}^p$, and g extracts a single component of θ: for example, $\theta = (\mu, \sigma^2)$ and $g(\theta) = \mu$. So I call the following result the Confidence Procedure Marginalisation Theorem.

Theorem 4.5 (Confidence Procedure Marginalisation, CPM)**.** *Suppose that $g : \theta \mapsto \phi$, and that C is a level-$(1-\alpha)$ procedure for θ. Then gC is a level-$(1-\alpha)$ confidence procedure for ϕ.*[12]

Proof. Follows immediately from the fact that $\theta \in C(y)$ implies that $\phi \in gC(y)$ for all y, and hence

$$\Pr\{\theta \in C(Y); \theta\} \le \Pr\{\phi \in gC(Y); \theta\}$$

for all $\theta \in \Omega$. So if C has θ-coverage of at least $(1-\alpha)$, then gC has ϕ-coverage of at least $(1-\alpha)$ as well. □

[12] gC $:= \Big\{\phi : \phi = g(\theta) \text{ for some } \theta \in C\Big\}$.

This result shows that we can derive level-$(1 - \alpha)$ confidence procedures for functions of θ directly from level-$(1 - \alpha)$ confidence procedures for θ. But it also shows that the coverage of such derived procedures will typically be more than $(1 - \alpha)$, even if the original confidence procedure is exact.

4.5 p-values

There is a general theory for p-values, also known as *significance levels*, which is outlined in Section 4.5.2, and critiqued in Section 4.5.3 and **??**. But first I want to focus on p-values as used in Hypothesis Tests, which is a very common situation.

As discussed in Section 4.3, we have methods for constructing families of good confidence procedures, and the knowledge that there are also families of confidence procedures which are poor (including completely uninformative). In this section I will take it for granted that a family of good confidence procedures has been used.

4.5.1 p-values and confidence sets

Hypothesis Tests (HTs) were discussed in Section 3.6. In a HT the parameter space is partitioned as

$$\Omega = \{H_0, H_1\},$$

where typically H_0 is a very small set, maybe even a singleton. We 'reject' H_0 at a significance level of α exactly when a level-$(1 - \alpha)$ confidence set $C(y^{\text{obs}}; \alpha)$ does not intersect H_0; otherwise we 'fail to reject' H_0 at a significance level of α.

In practice, then, a hypothesis test with a significance level of 5% (or any other specified value) returns one bit of information, 'reject', or 'fail to reject'. We do not know whether the decision was borderline or nearly conclusive; i.e. whether, for rejection, H_0 and $C(y^{\text{obs}}; 0.05)$ were close, or well-separated. We can increase the amount of information if C is a family of confidence procedures, in the following way.

Definition 4.4 (p-value, confidence set)**.** Let $C(\cdot\,; \alpha)$ be a family of confidence procedures. The p-value of H_0 is the smallest value α for which $C(y^{\text{obs}}; \alpha)$ does not intersect H_0.

The picture for determining the p-value is to dial up the value of α from 0 and shrink the set $C(y^{\text{obs}}; \alpha)$, until it is just clear of H_0. Of course we do not have to do this in practice. From the Representation Theorem (Theorem 4.1) we take $C(y^{\text{obs}}; \alpha)$ to be synonymous with a function $g : \mathcal{Y} \times \Omega \to \mathbb{R}$. Then $C(y^{\text{obs}}; \alpha)$ does not intersect with H_0 if and only if

$$\forall \theta \in H_0 : g(y^{\text{obs}}, \theta) \leq \alpha.$$

Thus the p-value is computed as

$$p_t(y^{\text{obs}}; H_0) := \max_{\theta \in H_0} g(y^{\text{obs}}, \theta), \tag{4.11}$$

for a specified family of confidence procedures (represented by the choice of g). Here is an interesting and suggestive result.[13] This will be the basis for the generalisation in Section 4.5.2.

Theorem 4.6. *Under Definition 4.4 and* (4.11), $p_t(Y; H_0)$ *is super-uniform for every* $\theta \in H_0$.

Proof. $p_t(y; H_0) \leq u$ implies that $g(y, \theta) \leq u$ for all $\theta \in H_0$. Hence

$$\Pr\{p_t(Y; H_0) \leq u; \theta\} \leq \Pr\{g(Y, \theta) \leq u; \theta\} \leq u \qquad : \theta \in H_0$$

where the final inequality follows because $g(Y, \theta)$ is super-uniform for all $\theta \in \Omega$, from Theorem 4.1. □

If interest concerns H_0, then $p_t(y^{\text{obs}}; H_0)$ definitely returns more information than a hypothesis test at any fixed significance level, because $p_t(y^{\text{obs}}; H_0) \leq \alpha$ implies 'reject H_0' at significance level α, and $p_t(y^{\text{obs}}; H_0) > \alpha$ implies 'fail to reject H_0' at signficance level α. But a p-value of, say, 0.045 would indicate a borderline 'reject H_0' at $\alpha = 0.05$, and a p-value of 0.001 would indicate nearly conclusive 'reject H_0' at $\alpha = 0.05$. So the following conclusion is rock-solid:

- When performing a HT, a p-value is more informative than a simple 'reject H_0' or 'fail to reject H_0' at a specified significance level (such as 0.05).

4.5.2 *The general theory of p-values*

Theorem 4.6 suggests a more general definition of a p-value, which does not just apply to hypothesis tests for parametric models, but which holds much more generally, for any PMF or model for Y. In the following f_0 is any *null model* for Y, including as a special case $f_0 = f(\bullet; \theta_0)$ for some specified $\theta_0 \in \Omega$.

Definition 4.5 (Significance procedure). $p : \mathcal{Y} \to \mathbb{R}$ is a *significance procedure* for f_0 exactly when $p_t(Y)$ is super-uniform under f_0; if $p_t(Y)$ is uniform under $Y \sim f_0$, then p is an *exact* significance procedure for f_0. The value $p_t(y^{\text{obs}})$ is a *significance level* or *p-value* for f_0 exactly when p is a significance procedure for f_0.

This definition can be extended to a set of PMFs for Y by requiring that p is a significance procedure for every element in the set; this is consistent with the definition of $p_t(y; H_0)$ in Section 4.5.1. The usual extension would be to take the maximum of the p-values over the set.[14]

For any specified f, there are a lot of significance procedures for $H_0 : Y \sim f$. An uncountable number, actually, because *every test statistic* $t : \mathcal{Y} \to \mathbb{R}$ *induces a significance procedure*. See Section 4.6 for the probability theory which underpins the following result.

[13] Recollect the definition of 'super-uniform' from Definition 4.3.

[14] Although Berger and Boos (1994) have an interesting suggestion for parametric models.

Theorem 4.7. *Let $t : \mathcal{Y} \to R$. Define*

$$p_t(y; f_0) := \Pr\{t(Y) \geq t(y); f_0\}.$$

Then $p_t(Y; f_0)$ is super-uniform under $Y \sim f_0$. That is, $p_t(\cdot\,; t)$ is a significance procedure for $H_0 : Y \sim f_0$. If the distribution function of $t(Y)$ is continuous, then $p_t(\cdot\,; f_0)$ is an exact *significance procedure for H_0.*

Proof.

$$p_t(y; f_0) = \Pr\{t(Y) \geq t(y); f_0\} = \Pr\{-t(Y) \leq -t(y); f_0\} =: G(-t(y))$$

where G is the distribution function of $-t(Y)$ under $Y \sim f_0$. Then

$$p_t(Y; f_0) = G(-t(Y))$$

which is super-uniform under $Y \sim f_0$ according to the Probability Integral Transform (see Section 4.6, notably Theorem 4.9). The PIT also covers the case where the distribution function of $t(Y)$ is continuous, in which case $p_t(\cdot\,; f_0)$ is uniform under $Y \sim f_0$. □

Like confidence procedures, significance procedures suffer from being too broadly defined. Every test statistic induces a significance procedure. This includes, for example, $t(y) = c$ for some specified constant c; but clearly a p-value based on this test statistic is useless.[15] So some additional criteria are required to separate out good from poor significance procedures. The most pertinent criterion is:

- select a test statistic for which $t(Y)$ which will tend to be larger for decision-relevant departures from H_0.

This will ensure that $p_t(Y; f_0)$ will tend to be smaller under decision-relevant departures from H_0. Thus p-values offer a 'halfway house' in which an alterntive to H_0 is contemplated, but not stated explicitly.

Here is an example. Suppose that there are two sets of observations, characterised as $\boldsymbol{Y} \overset{\text{iid}}{\sim} f_0$ and $\boldsymbol{Z} \overset{\text{iid}}{\sim} f_1$, for unspecified PMFs f_0 and f_1. A common question is whether $\boldsymbol{Y}$ and $\boldsymbol{Z}$ have the same PMF, so we make this the null hypothesis:

$$H_0 : f_0 = f_1.$$

Under H_0, $(\boldsymbol{Y}, \boldsymbol{Z}) \overset{\text{iid}}{\sim} f_0$. Every test statistic $t(\boldsymbol{y}, \boldsymbol{z})$ induces a significance procedure. A few different options for the test statistic are:

1. The sum of the ranks of $\boldsymbol{y}$ in the ordered set of $(\boldsymbol{y}, \boldsymbol{z})$. This will tend to be larger if f_0 stochastically dominates f_1.
2. As above, but with $\boldsymbol{z}$ instead of $\boldsymbol{y}$.
3. The maximum rank of $\boldsymbol{y}$ in the ordered set of $(\boldsymbol{y}, \boldsymbol{z})$. This will tend to be larger if the righthand tail of f_0 is longer than that of f_1.

[15] It is a good exercise to check that $t(y) = c$ does indeed induce a super-uniform $p_t(Y; f_0)$ for every f_0.

4. As above, but with $\boldsymbol{z}$ instead of $\boldsymbol{y}$.

5. The difference between the maximum and minimum ranks of $\boldsymbol{y}$ in the ordered set of $(\boldsymbol{y}, \boldsymbol{z})$. This will tend to be larger if f_0 and f_1 have the same location, but f_0 is more dispersed than f_1.

6. As above, but with $\boldsymbol{z}$ instead of $\boldsymbol{y}$.

7. And so on . . .

There is no 'portmanteau' test statistic to examine H_0, and in my view H_0 should always be replaced by a much more specific null hypothesis which suggests a specific test statistic. For example,

$$H_0 : f_1 \text{ stochastically dominates } f_0.$$

In this case (2.) above is a useful test statistic. It is implemented as the *Wilcoxon rank sum test* (in its one-sided variant).

4.5.3 Being realistic about significance procedures

Section 4.5.1 made the case for reporting a HT in terms of a p-value. But what can be said about the more general use of p-values to 'score' the hypothsis $H_0 : Y \sim f_0$? Let's look at the logic. As Fisher himself stated, in reference to a very small p-value,

> The force with which such a conclusion is supported is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution [i.e. the null hypothesis] is not true. (Fisher, 1956, p. 39).

Fisher encourages us to accept that rare events seldom happen, and we should therefore conclude with him that a very small p-value strongly suggests that H_0 is not true. This is uncontroversial, although how small 'very small' should be is more mysterious; Cowles and Davis (1982) discuss the origin of the $\alpha = 0.05$ convention.

But what would he have written if the p-value had turned out to be large? The p-value is only useful if we conclude something different in this case, namely that H_0 is not rejected. But this is where Fisher would run into difficulties, because H_0 is an artefact: f_0 is a distribution chosen from among a small set of candidates for our convenience. So we know *a priori* that H_0 is false: nature is more complex than we can envisage or represent. Fisher's logical disjunction is trivial because the second proposition is always true (i.e. H_0 is always false). So either we confirm what we already know (small p-value, H_0 is false) or we fail to confirm what we already know (large p-value, but H_0 is still false). In the latter case, all that we have found out is that our choice of test statistic is not powerful enough to tell us what we already know to be true.

This is not how people who use p-values want to interpret them. They want a large p-value to mean "No reason to reject H_0", so that when the p-value is small, they can "Reject H_0". They do not want it to mean "My test statistic is not powerful enough to tell me what I already know to be true, namely that H_0 is false." But unfortunately that is what it means.

Statisticians have been warning about misinterpreting p-values for nearly 60 years (dating from Lindley, 1957). They continue to do so in fields which use statistical methods to examine hypotheses, indicating that the message has yet to sink in. So there is now a huge literature on this topic. A good place to start is Greenland and Poole (2013), and then work backwards.

4.6 *The Probability Integral Transform*

Here is a very elegant and useful piece of probability theory. Let X be a scalar random quantity with realm $\mathcal{X}$ and distribution function $F(x) := \Pr(X \leq x)$. By convention, F is defined for all $x \in \mathbb{R}$. By construction, $\lim_{x \downarrow -\infty} F(x) = 0$, $\lim_{x \uparrow \infty} F(x) = 1$, F is non-decreasing, and F is continuous from the right, i.e.

$$\lim_{x' \downarrow x} F(x') = F(x).$$

Define the *quantile function*

$$F^{-}(u) := \inf \{ x \in \mathbb{R} : F(x) \geq u \}. \tag{4.12}$$

The following result is a cornerstone of generating random quantities with easy-to-evaluate quantile functions.

Theorem 4.8 (Probability Integral Transform, PIT)**.** *Let U have a standard uniform distribution. If F^{-} is the quantile function of X, then $F^{-}(U)$ and X have the same distribution.*

Proof. Let F be the distribution function of X. We must show that

$$F^{-}(u) \leq x \iff u \leq F(x) \tag{†}$$

because then

$$\Pr\{F^{-}(U) \leq x\} = \Pr\{U \leq F(x)\} = F(x)$$

as required. So stare at Figure 4.1 for a while.

It is easy to check that

$$u \leq F(x) \implies F^{-}(u) \leq x,$$

which is one half of (†). It is also easy to check that

$$u' > F(x) \implies F^{-}(u') > x.$$

Taking the contrapositive of this second implication gives

$$F^{-}(u') \leq x \implies u' \leq F(x),$$

which is the other half of (†). □

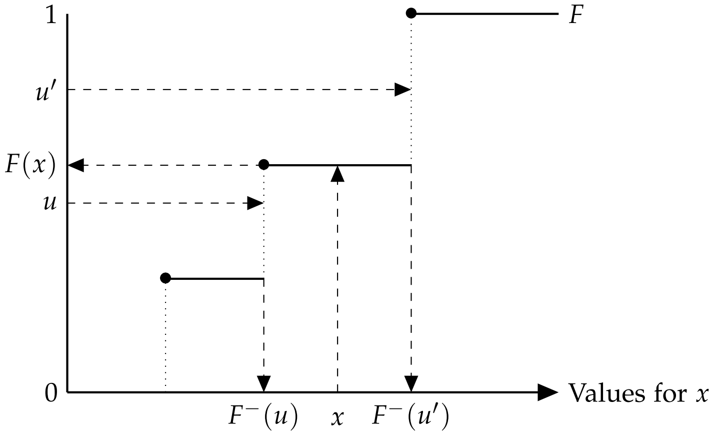

Figure 4.1: Figure for the proof of Theorem 4.8. The distribution function F is non-decreasing and continuous from the right. The quantile function F^- is defined in (4.12).

Theorem 4.8 is the basis for the following result; recollect the definition of a super-uniform random quantity from Definition 4.3. This result is used in Theorem 4.7.

Theorem 4.9. *If F is the distribution function of X, then $F(X)$ has a super-uniform distribution. If F is continuous then $F(X)$ has a uniform distribution.*

Proof. Check from Figure 4.1 that $F(F^-(u)) \geq u$. Then

$$\begin{aligned}\Pr\{F(X) \leq u\} &= \Pr\{F(F^-(U)) \leq u\} \qquad \text{from Theorem 4.8}\\ &\leq \Pr\{U \leq u\}\\ &= u.\end{aligned}$$

In the case where F is continuous, it is strictly increasing except on sets which have probability zero. Then

$$\Pr\{F(X) \leq u\} = \Pr\{F(F^-(U)) \leq u\} = \Pr\{U \leq u\} = u,$$

as required. □

5
Bibliography

Bartlett, M. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44:533–534. 57

Basu, D. (1975). Statistical information and likelihood. *Sankhyā*, 37(1):1–71. With discussion. 14, 15, 16, 20

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York, Inc., NY, USA, second edition. 30

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. 49

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). 25

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269–306. 13, 14, 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. 3, 5, 46

Çınlar, E. and Vanderbei, R. (2013). *Real and Convex Analysis*. Springer, New York NY, USA. 32

Cormen, T., Leiserson, C., and Rivest, R. (1990). *Introduction to Algorithms*. The MIT Press, Cambridge, MA. 12

Cowles, M. and Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37(5):553–558. 51

Cox, D. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge, UK. 3, 46

Cox, D. and Donnelly, C. (2011). *Principles of Applied Statistics*. Cambridge University Press, Cambridge, UK. 3

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London, UK. 15, 17, 31, 42

Cox, D. and Mayo, D. (2010). Objectivity and conditionality in Frequentist inference. In Mayo, D. and Spanos, A., editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge University Press, Cambridge, UK. 26

Davison, A. (2003). *Statistical Models*. Cambridge University Press, Cambridge, UK. 5

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. 47

Draper, N. and Smith, H. (1998). *Applied Regression Analysis*. New York: John Wiley & Sons, 3rd edition. 44

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`. 34

Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Edinburgh and London: Oliver and Boyd. 16, 51

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/`. 11

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London, UK. 30

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. 52

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. 3

Hacking, I. (2014). *Why is there a Philosophy of Mathematics at all?* Cambridge University Press, Cambridge, UK. 4

Lad, F. (1996). *Operational Subjective Statistical Methods*. New York: John Wiley & Sons. 4

Le Cam, L. (1990). Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171. 6, 22

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192. See also Bartlett (1957). 52

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. 11

MacKay, D. (2009). *Sustainable Energy – Without the Hot Air*. UIT Cambridge Ltd, Cambridge, UK. available online, at `http://www.withouthotair.com/`. 4

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. 10

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Harcourt Brace & Co., London, UK. 43, 45

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. 40

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. New York: Springer, 2nd edition. 6

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. 47

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. 34

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. 3, 21

Schervish, M. (1995). *Theory of Statistics*. Springer, New York NY, USA. Corrected 2nd printing, 1997. 3, 7, 10, 30

Smith, J. (2010). *Bayesian Decision Analysis: Principle and Practice*. Cambridge University Press, Cambridge, UK. 26

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. 11

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. 11

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK. 24, 46

Wood, S. (2015). *Core Statistics*. Cambridge University Press, Cambridge, UK. 43