1
Statistics: another short introduction

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; Stigler (2016) likewise. 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] And contains the funniest sentence ever written in Statistics, contributed by L.J. Savage.

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. I will refer to X as the *predictands*, Y as the *observables*, and y^{obs} as the *observations*; the observations are actual values.

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 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 $X = (X_1, \ldots, X_m)$, 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).

[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.

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'.

[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).

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 $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots\}$. The complete set of probability distributions for X is

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

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

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 particular way in which statisticians specify a subset of all distributions originates with Ronald Fisher in the 1920s; Stephen

Stigler calls it "One of Ronald A. Fisher's subtlest innovations" (Stigler, 2016, p. 180). The family is denoted by a *probability mass function (PMF)* f_X, a *parameter* θ, and a *parameter space* Ω, such that

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

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

$$f_X(\cdot\,; \theta') \neq f_X(\cdot\,; \theta''); \tag{1.3}$$

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

$$\mathcal{E} = \{\mathcal{X}, \Omega, f_X\}, \tag{1.4}$$

termed a *parametric model*. For example, the Poisson family is

$$\text{Poisson} = \{\mathbb{N}, \mathbb{R}_+, f_X\} \text{ where } f_X(x; \theta) = e^{-\theta} \frac{\theta^x}{x!},$$

although it is common in this case to use 'λ' rather than 'θ' as the label for the parameter.[6] Where $\mathcal{X}$ is embedded in a larger set, it is understood that $f_X(x; \cdot) = 0$ for $x \notin \mathcal{X}$. This would allow us to define the Poisson distribution over the realm $\mathbb{R}$, if that turned out to be convenient.

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

$$\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega, f_{X,Y}\}. \tag{1.5}$$

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. Some procedures start with a more general specification for f_X, termed *non-parametric* statistical models. The most common is that $f_X(x_1, \ldots, x_m)$ is a symmetric function of $(x_1, \ldots, x_m)$, termed *exchangeable*.

[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 semi-colon.

[6] $\mathbb{N}$ denotes the set of natural numbers, and $\mathbb{R}_+$ the set of non-negative real numbers. Mathematicians are flexible about whether $0 \in \mathbb{N}$: in our case it is.

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

[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).

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, 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

[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.

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.

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

1.3 *Prediction and inference*

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

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

[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.

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*

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

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

[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.

[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.

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.

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]

[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

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 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.

[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.

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

$$Y_i \mid X_i, \sigma^2 \sim f_{\epsilon_i}(X_i, \sigma^2) \qquad i = 1, \ldots, n \tag{1.14a}$$
$$X_i \mid \theta_i \sim f_{X_i}(\theta_i) \qquad i = 1, \ldots, m \tag{1.14b}$$
$$\theta_i \mid \psi \sim f_\theta(\psi) \qquad i = 1, \ldots, m \tag{1.14c}$$
$$(\sigma^2, \psi) \sim f_0\,. \tag{1.14d}$$

At the top (first) level is the measurement model for the sample $(Y_1, \ldots, Y_n)$, 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 $(X_1, \ldots, X_m)$, 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 $(\sigma^2, \theta_1, \ldots, \theta_m)$, 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]

[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).

[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).

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 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.

[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.

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]

[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.

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.[24]

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

[24] See Harford (2014) for an interesting essay about why big is not always better, and why in many situations we can expect statisticians to outperform 'data analysts'.