4
Modern Statistical Practice

The central part of this chapter describes modern statistical practice as a sequence of developments. This is *not* a history of statistics. Rather, it is a 'model' of statistics, where I understand a 'model' to be *an artefact used to organise our knowledge and beliefs*.[1] I stand by my definition of statistical inference in Chapter 1, and its recognition of our limitations when quantifying uncertainty. And yet we find very little trace of these limitations in modern statistical practice. Sec. 4.2 to Sec. 4.5 is my model of this anomaly; its sequential structure is an organisational device.

Before these middle sections, the next section does some preliminary spade-work, dispelling a naive interpretation of statistical practice ('learning') and replacing it with a naturalistic one, which I hope will be recognisable to any applied statistician, despite my need to describe it in rather abstract terms. And then after the middle sections there are two additional starred sections that cover two abiding issues in statistical practice. Sec. 4.6 considers when a Bayesian statistician can downplay his choice of prior distribution, based on a general result by L.J. Savage that has a wide range of applications. Sec. 4.7 considers invariance to the family of distributions specified by a family of statistical models.

4.1 *Some preliminary spadework*

This is a brief discussion about what statistics is not, and what it is (as actually practiced, not as theorised about).

4.1.1 *Statistical inference is not 'learning'*

Recall my definition of 'model' at the start of this chapter. There is a model of idealised learning, which runs as follows. There is a collection of random quantities, say $\boldsymbol{X} := (X_1, \dots, X_m)$, about which an agent has beliefs. These beliefs are represented as a conjunction of first-order sentences about $\boldsymbol{X}$ which he believes to be true, denoted by the proposition Ψ. The agent's complete set of beliefs is written as Bel_Ψ, where $\text{Bel}_\Psi(P)$ is his strength of belief in the proposition P. 'Learning' consists of adding new sentences to Ψ,

[1] My favourite example of a model in the natural sciences is the 'ocean conveyor belt'; see Lozier (2010).

and is represented formally as the arrow in

$$\mathrm{Bel}_\Psi(P) \longrightarrow \mathrm{Bel}_{Q\wedge\Psi}(P),$$

where Q is a sentence now believed by the agent to be true, and for which $Q \wedge \Psi$ is not a contradiction. $\mathrm{Bel}_{Q\wedge\Psi}$ might be termed the agent's *updated* beliefs.

Now we could choose to represent the belief function $\mathrm{Bel}_\Psi(\cdot)$ by the conditional probability $\Pr(\cdot \mid \Psi)$, and hence we could represent learning Q as

$$\Pr(P \mid \Psi) \longrightarrow \Pr(P \mid Q, \Psi).$$

This is known as *Bayesian conditionalisation*. Paris (1994) provides a very clear description of this model for learning from the point of view of computer science, and it has been popular with physicists such as R.T. Cox and Edwin Jaynes (Jaynes, 2003) and Harold Jeffreys (Jeffreys, 1961), and also philosophers (Jeffrey, 2004; Howson and Urbach, 2006).[2] Note that all parties see conditionalisation as a model for learning, and not a description for how we learn. As Jaynes expresses it, this model describes how we might program an agent such as a robot to operate on our behalf.

The learning rule in Bayesian conditionalisation has an interesting and attractive form, summarised in the following result.[3]

Theorem 4.1 (Muddy table theorem, again)**.** *Let* $\boldsymbol{X} := (X_1, \dots, X_m)$, Ψ *be a random proposition, and* $q(\boldsymbol{x})$ *be a first-order sentence with* $Q := q(\boldsymbol{X})$ *and* $\Pr(Q, \Psi) > 0$*. Then*

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q, \Psi) \propto \mathbb{1}_{q(\boldsymbol{x})} \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid \Psi)$$

where the constant of proportionality is $\Pr(Q \mid \Psi)^{-1}$.

Proof. Follows from the original Muddy Table table theorem (Thm 3.3) and the recursive property of hypothetical expectation (Sec. 3.2), which together imply that

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q, \Psi) \Pr(Q \mid \Psi) = \mathbb{1}_{q(\boldsymbol{x})} \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid \Psi)$$

provided that $\Pr(\Psi) > 0$. Then if $\Pr(Q, \Psi) > 0$ this can be rearranged, as required. □

Bas van Fraassen (1989, ch. 7) describes this learning rule in the following terms. We start with $\mathcal{X}$ represented as tiles on a tabletop (he suggested a Venn diagram). Each tile contains a heap of mud whose proportion in the total represents the agent's $\Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)} \mid \Psi)$, for tile j. When the agent learns that Q is true, he sweeps the mud off all the tiles for which $q(\boldsymbol{x}^{(j)})$ is false; i.e. all the tiles that are ruled out by the truth of Q. So Thm 3.3 is the *Muddy table theorem*.

Bayesian conditionalisation has three very attractive properties. First, it is consistent, so that if $\Psi \implies P$ then $\mathrm{Bel}_\Psi(P) = 1$. Second, it is property-preserving. If Bel_Ψ satisfies the three axioms of probability, then so will $\mathrm{Bel}_{Q\wedge\Psi}$; this follows from the recursive property (**??**). Third, it is order-invariant. If $q(\boldsymbol{x}) = q_1(\boldsymbol{x}) \wedge q_2(\boldsymbol{x})$, then $\mathrm{Bel}_{Q\wedge\Psi}$

[2] Sec. 4.6.1 presents a probabilistic result used as a model in the philosophy of science, but only as a bit of fun.

[3] Recall the comment on functional equalities in Sec. 1.5.1. The equality in Thm 4.1 holds for all $\boldsymbol{x}$ in the realm of $\boldsymbol{X}$.

will be the same whether the update is Q_1 then Q_2, or Q_2 then Q_1, or both together. These properties are so attractive that we should not be surprised to find that conditionalisation is also the basis of statistical practice.

Attractive as the learning model is, inference as practised by statisticians is *not* learning as described here. As described in Sec. 4.1.2, statistical practice tends to involve working *backwards* from the dataset (represented as the truth of Q above): this reverse direction is the antithesis of learning. But anything else would be completely impractical, because there is no end to the list of relevant things that might be learnt between two time-points, even in a highly controlled experiment. To have to anticipate all of these would unworkable, and even accounting for just a fraction of them would require a really massive $\mathfrak{X}$: far larger than we could compute with, and most of which would then be ruled out by the dataset.

One might think that this was too obvious to mention, but for the fact that statistical inference is constantly being confused with learning, a confusion that is even enshrined in our vocabulary—'prior' and 'posterior' distributions, for example: see Sec. 4.5. This confusion is also represented in the inane advice 'not to use the dataset more than once'. And in the practice of setting the characteristics of an inference ahead of analysing the dataset (as in hypothesis tests with prescribed error rates, or fixed thresholds for significance levels). Interestingly, it has affected some of the deepest thinkers in our profession:

> We are sometimes asked "If all rests ultimately on personal opinion, why should a person confronted with data not simply decide intuitively what opinions he finds himself with, having seen the data, rather than trouble to calculate these opinions from initial opinions and opinions about the mechanism of the experiment by means of Bayes' theorem?" The question has the merit of pointing out a false asymmetry between initial and final opinions. For it is only chronologically, not logically, that one has priority over the other.[4] (from the English summary of de Finetti and Savage, 1962)

The use of 'initially', 'final', and 'chronologically' all point to a learning interpretation, although the continuation of the quote (in the footnote) suggests a more iterative process.

4.1.2 *A more naturalistic description*

Sec. 4.1.1 made the point that statistics is not 'learning', in the sense of progressing from initial beliefs to updated beliefs through the application of an algorithm. What happens in practice? Statisticians do not work in a vacuum, so I will presume throughout these notes that the inference originates with a client.

Here are the steps in an actual statistical inference, from the point-of-view of the statistician:

1. Meet the client, take delivery of a *datapool*, a research objective, and some beliefs.

[4] The quote continues: "The reason to make Bayesian and other probability calculations is to allow a person to confront his various opinions with one another to see if they are coherent. If they are not, he will generally be able so to modify his opinions as to be satisfied that he has improved them. Often, but not always, it is the conclusions intuitively suggested by the experiment that will be so modified."

2. Sanity-check the datapool, and push errors back to the client. Keep doing this throughout the inference.

3. Meet the client again, refine the research objective and collect additional beliefs.

4. Specify a set of random quantities $\boldsymbol{X}$ which encompass some of the datapool (see Sec. 2.4), some of the client's beliefs, and her research objective, h.

5. Have a go at constructing E^*. Be prepared to return to any of the previous steps.

I write 'datapool' to emphasise that the dataset used in the inference is typically abstracted from a larger set of possible data. E^* is the resulting inference, suitable for all functions of $\boldsymbol{X}$, including h. This process is represented in the following Figure; as in Sec. 2.4, write the dataset as the truth of the random proposition Q.

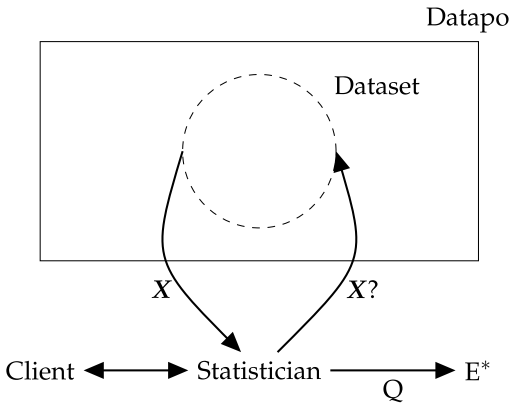

Figure 4.1: The steps in an actual inference. The statistician is responsible for synthesising the client's beliefs and as much of the datapool as is practical.

What makes this inferential process different from 'learning' are the extra arrows, from the statistician back to the client, and from the statistician back to the datapool. This backwards-and-forwards process subverts the simplistic view that there can be a separation between beliefs and data, and denies the notion that statistics is about proceeding from 'prior' beliefs E to 'posterior' beliefs E^* through conditioning on data. Instead, E^* should be thought of as an inference that satisfies two properties:

1. $\mathrm{Pr}^*(Q) = 1$, and

2. E^* reflects the beliefs of the client.

The desirability of these two properties is self-evident, although the ways in which they are achieved are different. The first property is built-in to the way in which E^* is constructed. The second involves diagnostic assessment by the statistician and the client (see Sec. 8.1).

* * *

The next four sections all present ways of constructing an E^* with property 1. I will conflate the roles of client and statistician, for simplicity, and focus on the latter. For convenience I will tend to refer to 'Bayesians' and 'Frequentists' as though they were two different tribes of statisticians, like the Houyhnhnms and Yahoos of *Gulliver's Travels*. But although many statisticians will self-identify as one or the other, generally a more pragmatic attitude prevails, and where I write, e.g., 'Bayesians' one ought to read this as 'statisticians operating in a Bayesian mode'.

4.2 *Brief review of Chapter 2*

The portrait of statistical inference given in Chapter 1 and Chapter 2 acknowledged from the outset our limitations as assessors of uncertainty. Hence the need for a calculus that imposes simple and intuitive rules. In the calculus of expectations, I specify my beliefs about $\boldsymbol{X}$ as expectations of a set of random quantities, including probabilities as a special case. Expectation is characterised, informally, in Def. 1.2, and defined by the properties given in Def. 1.4.

The approach of hypothetical expectations (and hypothetical probabilities) introduced in Chapter 3 provides a powerful approach to extending the set of expectations I can specify, for example by allowing me to reason causally within scenarios that I can construct, which may or may not happen.

Among my beliefs, represented as expectations and hypothetical expectations, I also include the dataset, represented as the truth of the random proposition Q. Formally I specify the belief $\Pr(Q) \leftarrow 1$, but this is equivalent to thinning the joint realm $\mathfrak{X}$, removing all of the $\boldsymbol{x}^{(j)}$ elements for which $q(\boldsymbol{x}^{(j)})$ is false. The mechanics of statistical inference were described in Chapter 2. My expectations can be checked for coherence, and they can be extended to expectations of those random quantities which are the objective of the inference. There is no need to distinguish between E and E^*.

Typically I will find that my expectations of many random quantities are not single values, but intervals. This 'undecided' aspect of my beliefs is a consequence of the limitations of my beliefs. The width of the interval can be reduced in a number of ways, all of which cost resources, but which might be justified by the need I have for tightly-constrained beliefs. For example, I might spend more time thinking about those beliefs which I have quantified. Or I might go out and extend my beliefs by polling experts. Or I might augment the dataset.

The computational tool is Linear Programming, see Sec. 2.2. But, as that section discussed, this tool does not scale well with large numbers of random quantities, because the size of the realm $\mathfrak{X}$ is exponential in the number of random quantities: that is why this chapter does not stop right here. The calculation we ought to do is often not practical, and an approximation must be found. As will be seen in the next three sections, the approximation is to adopt a level of personal omniscience which belies our limitations.

4.3 *Bayesian statistical inference*

Bayesians are bold. They laugh in the face of 'undecided', asserting that every expectation is specified. According to the FTP (Thm 1.6), this is equivalent to specifying a $\boldsymbol{p} \in \mathbb{S}^{s-1}$, where $p_j = \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)})$.[5] Were I being a Bayesian, I would have such a $\boldsymbol{p}$, and my expectation for any random quantity $h(\boldsymbol{X})$ would be computable using the FTP,

$$\mathrm{E}\{h(\boldsymbol{X})\} = \textstyle\sum_j h(\boldsymbol{x}^{(j)}) \cdot p_j.$$

As explained in the previous chapters, this expression also covers probabilities, hypothetical expectations, and hypothetical probabilities.

Now $\mathcal{X}$ might be huge, comprising thousands if not millions of elements. Clearly our Bayesian cannot think about each $\boldsymbol{x}^{(j)} \in \mathcal{X}$ and specify each p_j. Instead, he specifies a formula 'p' for which

$$p_j \leftarrow \mathrm{p}(\boldsymbol{x}^{(j)}) \qquad j = 1, \dots, s.$$

Here p is a PMF, as described in Sec. 1.5.1. In that section $\mathrm{p}(\boldsymbol{x})$ represented the single value $\Pr(\boldsymbol{X} \doteq \boldsymbol{x})$. But this is different. Now 'p' represents a function which returns a value for every $\boldsymbol{x} \in \mathcal{X}$. So were we to ask our Bayesian what his probability was for the random proposition $\boldsymbol{X} \doteq \boldsymbol{x}$ he would say "Hang on while I plug $\boldsymbol{x}$ into my function p ... Ah ha! It's 0.002347843." Now is this really his probability? Well, it is now! This is the basic challenge of the Bayesian framework, to propose a defensible function for specifying a probability for every element of $\mathcal{X}$. I come back to this in Sec. 4.5.[6]

Having surmounted this challenge, the Bayesian is then in good shape. The dataset, represented as the truth of data proposition Q, is incorporated into his beliefs by conditioning,

$$\mathrm{p}^*(\boldsymbol{x}) := \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q) \propto \mathbb{1}_{q(\boldsymbol{x})}\, \mathrm{p}(\boldsymbol{x}) \tag{4.1a}$$

by the Muddy table theorem (Thm 3.3). Summing over the whole of $\mathcal{X}$ supplies the missing constant of proportionality, which is $\Pr(Q)^{-1}$, presuming that this is positive. Finally, expectations of interest are computed as

$$\mathrm{E}^*\{h(\boldsymbol{X})\} := \mathrm{E}\{h(\boldsymbol{X}) \mid Q\} = \textstyle\sum_j h(\boldsymbol{x}^{(j)}) \cdot \mathrm{p}^*(\boldsymbol{x}^{(j)}), \tag{4.1b}$$

according to the CFTP (Thm 3.2).

Bayesians tend to take it for granted that the truth of Q is incorporated into beliefs by conditioning. The attractive features of this method were described in Sec. 4.1.1.

The last two decades have seen a statistical computing revolution in which this inferential calculation can be extended to joint realms which are non-finite and non-countable, using *Markov chain Monte Carlo (MCMC)* sampling techniques; see Besag *et al.* (1995) and Besag (2004) for summaries, and Robert and Casella (2004) for details.

[5] As usual, $\mathbb{S}^{s-1}$ is the $(s-1)$-dimensional simplex, defined in (1.2), and s is the size of $\mathcal{X}$, the realm of $\boldsymbol{X}$.

[6] An alternative notation to the function p is the function $f_{\boldsymbol{X}}$. I would prefer this alternative, but for the fact that it is often useful to keep the subscript slot that is filled by $\boldsymbol{X}$ free, to allow for a more extensible notation.

With these techniques it is not necessary to enumerate $\mathcal{X}$, and so the size of $\mathcal{X}$ is not, in itself, an impediment to inference, as it would be for the Linear Programming calculations described in Sec. 2.2. Also it is not necessary to know $\Pr(Q)$. It is hard to overstate the way in which the simultaneous development of MCMC sampling algorithms and computing power has revolutionised Bayesian statistical inference. There are now tools in which one supplies a p and a q, and everything else is automated; see, e.g., Lunn *et al.* (2013) for a description of BUGS.[7]

4.4 *Frequentist statistical inference*

Frequentists are ostensibly more cautious than Bayesians. Unlike Bayesians, they are unwilling to commit up-front to a single vector $\boldsymbol{p} \in \mathbb{S}^{m-1}$, preserving some 'undecided' in all of their expectations. They do this by proposing a *statistical model*, which is a family of PMFs indexed by a *parameter* $\theta \in \Omega$, where Ω is the *parameter space*. This family is represented by the function 'p' where, for any particular choice of θ,

$$p_j \leftarrow \mathrm{p}(\boldsymbol{x}^{(j)}; \theta) \quad j = 1, \dots, s.$$

In other words, the Frequentist's 'p' is a function with domain $\mathcal{X} \times \Omega$, where $\mathrm{p}(\cdot; \theta)$ is a PMF for $\boldsymbol{X}$, for any $\theta \in \Omega$. There is no risk of confusing the Bayesian's 'p' with the Frequentist's 'p' because the latter has a second argument.

With this statistical model, any expectation or probability is a function of θ:

$$\mathrm{E}\{h(\boldsymbol{X}); \theta\} = \textstyle\sum_j h(\boldsymbol{x}^{(j)}) \cdot \mathrm{p}(\boldsymbol{x}^{(j)}; \theta),$$

by the FTP. There is more discussion about families of distributions in Sec. 4.7, but the material in this section should be read first.

In fact, this notion of a statistical model is completely general. The vector $\boldsymbol{p}$ lives in $\mathbb{S}^{s-1}$, which has the cardinality of the continuum. Thus if Ω also has the cardinality of the continuum, e.g. the convex interval $[0, 1]$, then we can arrange a bijective relationship between $\mathbb{S}^{s-1}$ and Ω, and every possible $\boldsymbol{p}$ can be represented by a $\theta \in \Omega$. But what actually happens, of course, is that the Frequentist chooses the statistical model and Ω to severely restrict the set of possible $\boldsymbol{p}$.

A simple definition of the effective dimension of the statistical model is the minimum number of expectations it takes to completely specify $\boldsymbol{p}$. In standard statistical models this equates to the dimension of Ω, which is usually treated as a product space (see Sec. 4.7). For example, $\theta = (\lambda)$ for a Poisson model for X, or $\theta = (\mu, \sigma^2)$ for a Normal model for X. In the Poisson model, $\mathrm{E}(X; \lambda) \leftarrow v$ is sufficient to specify (λ) and thus $\boldsymbol{p}$, while in the Normal model $\mathrm{E}(X; \mu, \sigma^2) \leftarrow v_1$ and $\mathrm{E}(X^2; \mu, \sigma^2) \leftarrow v_2$ are sufficient to specify (μ, σ^2) and thus $\boldsymbol{p}$. So we should not get carried away by

[7] And I should mention JAGS, `http://mcmc-jags.sourceforge.net/` and the new kid on the block, STAN, `http://mc-stan.org/`.

the potential generality of the Frequentist approach: if the Bayesian approach has effective dimension zero ($\boldsymbol{p}$ specified directly), then typical Frequentist models might have effective dimension of 'few'. These are both a long way from the cardinality of the continuum!

Now to incorporate the belief that Q is true. The natural approach is for the Frequentist to consider each element of Ω. First,

$$\mathrm{p}^*(\boldsymbol{x};\theta) := \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q;\theta) \propto \mathbb{1}_{q(\boldsymbol{x})}\ \mathrm{p}(\boldsymbol{x};\theta) \tag{4.2a}$$

by the Muddy table theorem (Thm 3.3), where the missing constant of proportionality is $\Pr(Q;\theta)^{-1}$, presuming that this is positive. Then

$$\mathrm{E}^*\{h(\boldsymbol{X});\theta\} := \mathrm{E}\{h(\boldsymbol{X}) \mid Q;\theta\} = \textstyle\sum_j h(\boldsymbol{x}^{(j)}) \cdot \mathrm{p}^*(\boldsymbol{x}^{(j)};\theta) \tag{4.2b}$$

by the conditional FTP. So far, this is a direct analogue of the Bayesian approach, except with an added '$;\theta$'. What to do, though, about the 'width' of Ω, represented by the unspecified value θ? Our Frequentist could report the union of the set of values for $\mathrm{E}^*\{h(\boldsymbol{X});\theta\}$ that is generated over those $\theta \in \Omega$ for which $\Pr(Q;\theta) > 0$. Thus he remains 'undecided' because these expectations will not all coincide.[8]

Unfortunately, the resulting union is typically far too wide to be useful (or believable). The Frequentist solution is to constrain the domain of θ according to the truth of Q. By far the most popular approach is to use the *Maximum Likelihood (ML) estimate*

$$\hat{\theta}(q) := \operatorname*{argmax}_{\theta \in \Omega} \Pr(Q;\theta).$$

If this estimate is plugged-in for θ, then we have the resulting inference

$$\hat{\mathrm{E}}^*\{h(\boldsymbol{X})\} := \mathrm{E}^*\left\{h(\boldsymbol{X});\hat{\theta}(q)\right\}, \tag{4.3}$$

which is a point value with no 'undecided'.

Sec. 4.7 provides a justification for using the ML estimate as the 'plug-in' value for θ. Another reason for favouring the ML estimate is that it gives sensible answers in some simple situations (IID observations from the Exponential family, for example). Typically the ML estimate has to be computed numerically. In this case, the cost of evaluating $\Pr(Q;\cdot)$ is a major factor, where

$$\Pr(Q;\theta) = \textstyle\sum_j \mathbb{1}_{q(\boldsymbol{x}^{(j)})} \cdot \mathrm{p}(\boldsymbol{x}^{(j)};\theta).$$

This is a sum over $\mathcal{X}$, which might be massive; finding the maximum over Ω will require many (thousands, if Ω has more than one or two dimensions) of repetitions of this calculation. The one tractable situation is the simple observation model (SOM, see Sec. 2.4) combined with a statistical model with the product form

$$\mathrm{p}(\boldsymbol{x};\theta) = \prod_{i=1}^{m} \mathrm{p}(x_i;\theta),$$

[8] There is no guarantee that the set of expectations will be convex, though, so it would not be correct to call it an interval.

because then

$$\Pr(Q;\theta) = \prod_{i=1}^{n} \mathrm{p}(y_i^{\mathrm{obs}};\theta).$$

It is no coincidence that this is the typical example in textbooks. The more general situations are gruelling. The *Expectation Maximisation (EM)* algorithm has stood the test of time; see Hastie *et al.* (2009, sec. 8.5).[9]

4.5 *Bayesian/Frequentist synthesis*

In Sec. 4.3, where did the Bayesian's p come from? In the synthesis of the two approaches to statistical inference it comes from accepting the notion of a statistical model p and a parameter $\theta \in \Omega$, but choosing to treat θ itself as uncertain.[10] I hesitate to call θ a 'random quantity', because typically it does not have an operational definition. Instead I will call it a *random variable*. The Bayesian treats the statistical model as a conditional distribution, which is formally valid because $\mathrm{p}(\boldsymbol{x};t)$ behaves exactly like the hypothetical probability $\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid \theta \doteq t)$.

One difficulty immediately presents itself. I have argued that operationally-defined random quantities must have finite realms. But there is no such restriction on random variables,[11] and in fact in most cases random variables are constructed to have uncountably infinite realms. In this chapter I will treat θ as an *absolutely continuous* random variable, with a probability distribution represented as the *probability density function (PDF)* π_θ, for which

$$\pi_\theta(t)\,\mathrm{d}t = \Pr(t \mathrel{\dot{<}} \theta \mathrel{\dot{\leq}} t + \mathrm{d}t).$$

This implies

$$\pi_\theta(t) \geq 0 \quad \text{and} \quad \int_\Omega \pi_\theta(t)\,\mathrm{d}t = 1.$$

Treating θ as continuous provides some notational variety. All of the results presented so far generalise to such random variables, under a stronger countable additivity axiom for expectation (see Sec. 1.3).

Armed with a PDF for θ, the Bayesian constructs the joint distribution

$$\Pr(\boldsymbol{X} \doteq x, \theta \doteq t) \leftarrow \mathrm{p}(\boldsymbol{x};t)\,\pi_\theta(t)\,\mathrm{d}t, \tag{4.4}$$

following the template in (3.2).[12] The PDF π_θ is termed the *prior distribution*—a label I do not like for reasons given in Sec. 4.1.1, but we are stuck with it. The choice of p and π_θ induce a prior expectation E over functions of $\boldsymbol{X}$. However, this may not be a meaningful representation of anyone's beliefs. First, as noted in Sec. 4.1.2, the statistician's objective is to construct a meaningful E^*, and p and π_θ (and the E they imply) are just ingredients in this process. Second, Sec. 4.6 will explain why the Bayesian is sometimes able to replace a considered π_θ with something flatter and more tractable. This is not to downplay the difficulty of specifying π_θ, but to only to note

[9] And also the excellent Wikipedia page at `https://en.wikipedia.org/wiki/Expectation-maximization_algorithm`

[10] This synthesis position is somewhat conciliatory. A more general treatment of statistical modelling is given in Chapter 7.

[11] But we must be mindful of the PEP (Def. 1.1): whenever we use such random variables, they must not create pathologies in the distribution of the random quantities.

[12] For simplicity, I write $\theta \doteq t$ rather than $t \mathrel{\dot{<}} \theta \mathrel{\dot{\leq}} t + \mathrm{d}t$.

that we should be careful not to misjudge 'ostentatiously wrong' choices of π_θ.

With π_θ specified, the Bayesian is back on-track, except now the uncertain quantities are $(\boldsymbol{X}, \theta)$ rather than just $\boldsymbol{X}$. Hence, provided that $\Pr(Q) > 0$,

$$\begin{aligned}
&\Pr(\boldsymbol{X} \doteq \boldsymbol{x}, \theta \doteq t \mid Q) \\
&\quad = \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid \theta \doteq t, Q) \Pr(\theta \doteq t \mid Q) \quad \text{by seq. cond. (Thm 3.9)} \\
&\quad = \mathrm{p}^*(\boldsymbol{X} \doteq \boldsymbol{x}; t)\, \pi^*_\theta(t)\, \mathrm{d}t \qquad (4.5a)
\end{aligned}$$

where the first term was defined in (4.2a), and

$$\pi^*_\theta(t)\, \mathrm{d}t := \Pr(\theta \doteq t \mid Q) = \frac{\Pr(Q; t)\, \pi_\theta(t)\, \mathrm{d}t}{\Pr(Q)} \qquad (4.5b)$$

by Bayes's theorem (Thm 3.11); termed the *posterior distribution*. Treated as a function of $t \in \Omega$, $\Pr(Q; t)$ is termed the *likelihood function*. Hence the Bayesian mantra:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Finally,

$$\begin{aligned}
\mathrm{E}^*\{h(\boldsymbol{X})\} &:= \mathrm{E}\{h(\boldsymbol{X}) \mid Q\} \\
&= \int \textstyle\sum_j h(\boldsymbol{x}^{(j)}) \cdot \mathrm{p}^*(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}; t)\, \pi^*_\theta(t)\, \mathrm{d}t \quad \text{CFTP (Thm 3.2) and (4.5a)} \\
&= \int \mathrm{E}^*\left\{h(\boldsymbol{X}); t\right\} \pi^*_\theta(t)\, \mathrm{d}t, \qquad (4.5c)
\end{aligned}$$

where the first term in the integrand previously occurred in (4.2b). In other words, the Bayesian knows precisely how to handle the 'width' of Ω: he averages $\mathrm{E}^*\{h(\boldsymbol{X}); \theta\}$ over the posterior distribution π^*_θ. This result is important enough to have its own number.

Theorem 4.2. *In order to incorporate probabilistic parametric uncertainty into an inference, do the inference for each $t \in \Omega$ and then average the result over the posterior distribution π^*_θ.*

* * *

This, then, is the key difference between the Frequentist approach and the Bayesian approach. Let us suppose that all agree on the statistical model $\mathrm{p}(\cdot; \theta)$ with its parameter $\theta \in \Omega$ (but see the discussion at the end of Sec. 4.6). The Frequentist approach eschews the specification of a prior distribution π_θ but must then adduce an 'extra-probabilistic' principle for handling the width of Ω in the inference.[13] The difficulty for the Frequentist is that no compelling principle has been found. Although maximum likelihood is the most popular, collapsing Ω to a single point $\hat{\theta}(q)$ is a drastic step.

In contrast, the Bayesian approach specifies a prior distribution π_θ and is then able to handle the width of Ω within the standard rules of probability. The difficulty for the Bayesian approach is that θ is not operationally-defined, and so π_θ is not a very natural

[13] 'Extra' as in 'outside' or 'beyond'.

PMF to specify. And so, in this case also, 'extra-probabilistic' principles are often adduced to handle the choice of π_θ; see Kass and Wasserman (1996) or Robert (2007) for reviews. The next section tackles the tricky question of when it is relatively harmless for a Bayesian to replace a carefully considered prior distribution with a rule-based one.

4.6 *Stable estimation*

As already discussed, the course of an inference involves model development, in which the statistician and the client iterate through a sequence of models, possibly varying the subset of the datapool which is modelled directly. Each model might have a different parameter space, which is bad news for the Bayesian, who has to specify a prior distribution for the parameters of each model. Early on, he may well prefer to use a rather simple rule-based prior distribution, and focus his efforts, as the Frequentist would, on the development of the model. But he may find, as his choice for the model settles down, that the effect on his inference of changing the prior distribution is rather small. This insentitivity to the choice of prior distribution can be formally analysed, and the result is a set of qualitative guidelines under which the Bayesian can replace a carefully considered prior distribution with a rule-based one, at no serious detriment to his inference.

The analysis was presented in a classic paper, Edwards *et al.* (1963), but was almost certainly the work of the third author, L.J. Savage (see Lindley, 1980). I will adapt Savage's analysis to my own notation. Let $\Omega := \{t^{(1)}, \dots, t^{(k)}\}$ be the parameter space, which I take to be finite but otherwise unstructured.[14] Let $\boldsymbol{u} := (u_1, \dots, u_k)$ be proportional to the prior probabilities, and $\boldsymbol{v} := (v_1, \dots, v_k)$ be proportional to the likelihoods, i.e.

$$u_i \propto \pi_\theta(t^{(i)}) \quad \text{and} \quad v_i \propto \Pr(Q; t^{(i)}) \qquad i = 1, \dots, k.$$

The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are the primitive quantities; everything below is derived from these. Define

$$p_i := \frac{v_i}{\sum_j v_j} \quad \text{and} \quad q_i := \pi^*_\theta(t^{(i)}) = \frac{v_i \, u_i}{\sum_j v_j \, u_j} \qquad i = 1, \dots, k.$$

Thus $\boldsymbol{q} := (q_1, \dots, q_k)$ are the posterior probabilities, and $\boldsymbol{p} := (p_1, \dots, p_k)$ are the normalised likelihoods. In these terms our interest is in when we can approximate an inference based on $\boldsymbol{q}$ with one based on $\boldsymbol{p}$, for which we do not have to specify a prior distribution.

Savage provided three *stability conditions* that are sufficient to provide a small upper bound on the error of such an approximation, based around a subset $B \subset \Omega$. In each case I give the formal condition and its interpretation.

1. There is an $\alpha \ll 1$ for which

$$\sum_{i \notin B} p_i \le \alpha \sum_{i \in B} p_i.$$

[14] Finiteness is not important: I impose it for simplicity and because two applications of this result below (Sec. 6.6 and Sec. 8.2.1) are naturally expressed in terms of a finite set.

▶ The subset B contains almost all of the relative likelihood.

2. There is a $\beta \ll 1$ for which

$$\psi \le u_i \le \psi(1+\beta) \quad \text{for all } i \in B$$

(the value of ψ is unimportant).

▶ The prior probabilities change very little in B.

3. There is some θ for which

$$u_i \le \theta\psi \quad \text{for all } i,$$

where $\gamma := \alpha\theta \ll 1$.

▶ The prior probabilities are nowhere very large compared to their (nearly constant) values in B.

Savage proved the following result (Implication 5, p. 203).[15]

Theorem 4.3 (Stable estimation theorem).
Let $g : \Omega \to \mathbb{R}$ have upper bound G. Then

$$\left| \sum_t g(t^{(i)})\, p_i - \sum_t g(t^{(i)})\, q_i \right| \le G \cdot \big(\alpha + \beta + \gamma + \max\{\alpha, \gamma\}\big).$$

As an immediate corollary, set $g(t) \leftarrow \mathbb{1}_{t \in C}$ for any $C \subset \Omega$, and then Thm 4.3 implies that the *total variation distance* between the normalised likelihood and the posterior distribution is bounded above by $\alpha + \beta + \gamma + \max\{\alpha, \gamma\}$.

Referring back to Sec. 4.5, we are interested in the inference

$$\mathrm{E}^*\{h(\boldsymbol{X})\} = \int \mathrm{E}^*\{h(\boldsymbol{X}); t\}\, \pi^*_\theta(t)\, \mathrm{d}t = \int g(t)\, q(t)\, \mathrm{d}t$$

taking $g(t) \leftarrow \mathrm{E}^*\{h(\boldsymbol{X}); t\}$, and switching to an absolutely continuous θ, with $q \leftarrow \pi^*_\theta$. We might consider instead the approximation

$$\tilde{\mathrm{E}}^*\{h(\boldsymbol{X})\} := \int g(t)\, p(t)\, \mathrm{d}t$$

replacing the posterior distribution q with the normalised likelihood p. Thm 4.3 asserts that the relative absolute error in replacing E^* with $\tilde{\mathrm{E}}^*$ is bounded above by $\alpha + \beta + \gamma + \max\{\alpha, \gamma\}$. If the three stability conditions hold, then this value is close to zero, and

(i) the normalised likelihood is close to the posterior distribution in total variation distance, and

(ii) the approximate inference $\tilde{\mathrm{E}}^*\{h(\boldsymbol{X})\}$ is close to the actual inference $\mathrm{E}^*\{h(\boldsymbol{X})\}$ in relative absolute error.

In principle the crucial set B can be any subset of Ω. Usually, however, Ω has a topology. In this case there is a compelling reason to restrict B to a contiguous subset of Ω, because smoothness in the prior distribution will then imply a smaller β in condition 2 than would otherwise be the case. Suppose that $\theta := (\theta_1, \dots, \theta_p)$ with

[15] For those referring to the original paper, I have simplified the expression by approximating δ and ϵ in terms of α, β, and γ, using $\delta \approx \beta + \gamma$ and $\epsilon \approx \alpha + \beta$. So technically the result as stated is not quite true.

$\Omega \subset \mathbb{R}^p$. In this case a simple and effective strategy for identifying a B which respects the topology of Ω is to define it as a level set for the likelihood, i.e.

$$B := \{i : v_i \geq c\},$$

and adjust c from the maximum likelihood value downwards until a sufficiently small α is reached. This tends to generate connected B's because small perturbations in the parameter value tend to cause only small perturbations in the likelihood. Helpfully, the level sets of the likelihood are transformation-invariant, so it would not matter whether the parameter was θ or some bijection (see Sec. 4.7).

Once a B with a small α has been found, the statistician must decide whether β and γ are sufficiently small, to satisfy stability conditions 2 and 3. Of course he could do this explicitly if he has specified a prior distribution. But the attraction of Savage's stability conditions is that he may be able to do this qualitatively, without specifying a prior distribution. Edwards *et al.* (1963) provide a detailed illustration. This is not a trivial exercise, even in the simple models that were ubiquitous thirty years ago. Now, however, we can compute with a very diverse set of models, and there has also been a widening of the kinds of applications that statisticians consider. So it is really quite hard to know whether, for one particular model, stability conditions 2 and 3 apply.[16]

There are two guidelines that can help. First, to construct models in which the parameters are meaningful. For example, to map individual parameters to distinct beliefs about $\boldsymbol{X}$. It is tempting to treat the parameters in a purist manner, devoid of meaning except as an index into a family of distributions (Sec. 4.7). But in this case it is impossible to have well-formed beliefs about the prior probabilities, and in particular beliefs about whether these prior probabilities might be much larger in the complement of B than in B, violating stability condition 3. Hierarchical models (Sec. 7.4) are a powerful framework for constructing models with meaningful parameters.

The second guideline is to favour models with fewer parameters. This limits the opportunity for one parameter to offset another in the likelihood, and should result in a more compact B with a smaller β in stability condition 2. Also, of course, the more parameters there are, the harder it is to ascribe a distinct meaning to each parameter.

Both of these guidelines are basic tenets of statistical modelling, followed by experienced statisticians of all tribes; see, e.g., Lehmann (1990) and Cox (1990), in the same volume of *Statistical Science* (vol. 5, num. 2). For the Bayesian, though, they have the beneficial side-effect of enabling an assessment of whether the stability conditions hold. If so, the Bayesian can simplify his inference by replacing a carefully-specified prior distribution with a flat and tractable one, confident in the knowledge that the relative absolute error in his inference will be small.

When we inspect a published Bayesian inference, we often find

[16] Lindley (1980) alluded to this issue in his review of Savage's work.

such flat and tractable prior distributions, such as $\pi_{\theta_1}(t_1) \propto 1$ or $\pi_{\theta_2}(t_2) \propto 1/t_2$. It is vey important to appreciate that in this case the $\boldsymbol{X}$-margin of the joint distribution $(\boldsymbol{X}, \theta)$ is *not* a representation of the Bayesian's beliefs about $\boldsymbol{X}$. Instead, the Bayesian has replaced his actual joint distribution for $(\boldsymbol{X}, \theta)$ with another distribution which does not have the right $\boldsymbol{X}$-margin, but which still gives approximately the right posterior distribution and inference about $h(\boldsymbol{X})$ after conditioning on the dataset. This issue will resurface in Chapter 8.

4.6.1 A model for scientific induction

I deliberately refrained from framing the previous material in terms of the question "When might two statisticians agree in their inference?" This has been a traditional concern, and is a major preoccupation of 'learning' theories. The challenge for such theories is to explain why, if beliefs are subjective, we often come to hold many beliefs in common. A superficially attractive answer is to point to the Stable estimation theorem, or something like it, which indicates that agreement on the model and a sufficiently large dataset might do the trick, for Frequentists and Bayesians alike.

However, this answer fails, from a statistical point of view, because there is no reason for two statisticians to have the same model, or to make the same choice about which portion of the dataset to condition on. Hence the likelihood is just as subjective as the prior distribution.[17] This is not to say that the choices are not similar enough to lead to the same inferences. But we cannot prove similarity of beliefs in a formal sense unless there are strict conditions on the model.

But, just for fun, here is a probabilistic model for why scientists might come to agree on the predictions of a scientific theory. Suppose there is a sequence of experimental outcomes, $E_1, E_2, \dots$, all of which are implied by a scientific theory M. Represent this as

$$\Pr(E_A \mid M) = 1 \quad \text{for all } A, \tag{$\dagger$}$$

where E_A denotes the conjunction of any subset A of the experimental outcomes. Then we have the following remarkable result, termed the *First Induction Theorem* by Good (1975), and originally proved by Wrinch and Jeffreys (1921).

Theorem 4.4 (First Induction Theorem)**.**
Let $\Pr(E_A \mid M) = 1$ *for all* A*. If* $\Pr(M) > 0$ *then*

$$\lim_{n \to \infty} \Pr(E_{n+1} \mid E_1, \dots, E_n) = 1.$$

Proof. Under the conditions of the theorem,

$$\begin{aligned}
\Pr(E_A) &= \Pr(E_A \mid M)\Pr(M) + \Pr(E_A \mid \neg M)\Pr(\neg M) \quad \text{by the LTP, Thm 3.10} \\
&\geq \Pr(E_A \mid M)\Pr(M) \\
&= 1 \cdot \Pr(M).
\end{aligned}$$

[17] In fact I would go further, based on my own experience, which I doubt is unusual. I would be unlikely to choose the same model and dataset from one year to the next. New modelling frameworks are appearing all the time; often these were previously ruled out for computational reasons that are being eroded by faster CPUs, more memory, and parallel computation. Also, I am getting more experienced.

Now let $A \leftarrow \{1, \dots, n\}$ and write the lefthand side as

$$p_n := \Pr(E_1, \dots, E_n) = \Pr(E_1) \prod_{i=2}^{n} \Pr(E_i \mid E_1, \dots, E_{i-1})$$

using the Factorisation theorem (Thm 3.8). $\{p_n\}$ is a monotone decreasing sequence bounded below by $\Pr(M)$. Since $\Pr(M) > 0$ it converges to a positive limit, in which case $\Pr(E_n \mid E_1, \dots, E_{n-1})$ converges to 1. □

The remarkable thing about this result is that the displayed equation in Thm 4.4 makes no reference to the theory M at all. It indicates that anyone who believes that M implies the E's and that M is not utterly false is bound, sooner or later, on the accumulation of enough evidence, to act as though M is true, in terms of their expectations for other implications of M. I think we could criticise the condition $\Pr(M) > 0$, on the basis that scientific theories are always abstractions, and that no theory works alone in implying an experimental result (see, e.g., Cartwright, 1983). But it is a very cute proof.

A feature of the First Induction Theorem is that its conditions are not sufficient to imply that $\lim_n \Pr(M \mid E_1, \dots, E_n) = 1$. But the reason is straightforward: there may be another model which is equivalent to M. If this is ruled out then the result follows. Good (1975) calls this the *Second Induction Theorem*; it was first proved by Keynes (1921, ch. XX). I'm not sure it means anything—I am just including it for completeness.

Theorem 4.5 (Second Induction Theorem)**.** *Under the same conditions as Thm 4.4, and the additional condition* $\lim_n \Pr(E_1, \dots, E_n \mid \neg M) = 0$ *where* $\neg M$ *is the complement of* M*,*

$$\lim_{n \to \infty} \Pr(M \mid E_1, \dots, E_n) = 1.$$

Proof. Using Bayes's Theorem in odds form (see after Thm 3.11),

$$\begin{aligned} \frac{\Pr(M \mid E_1, \dots, E_n)}{\Pr(\neg M \mid E_1, \dots, E_n)} &= \frac{\Pr(E_1, \dots, E_n \mid M)}{\Pr(E_1, \dots, E_n \mid \neg M)} \cdot \frac{\Pr(M)}{\Pr(\neg M)} \\ &= \frac{1}{\Pr(E_1, \dots, E_n \mid \neg M)} \cdot \frac{\Pr(M)}{\Pr(\neg M)} \end{aligned}$$

which tends to ∞ under the conditions of the Theorem, so that $\Pr(M \mid E_1, \dots, E_n)$ tends to 1. □

4.7 *Models and invariance*

A family of models is simply a set $\mathcal{F}$, where each $p \in \mathcal{F}$ is a PMF for $\boldsymbol{X}$. The Frequentist in Sec. 4.4 must specify $\mathcal{F}$, likewise the Bayesian in Sec. 4.5. The crucial point, which can easily be obscured in textbook descriptions, is that the set $\mathcal{F}$ is the basis for any inference. In particular, inferences should be invariant to the way in which we label the elements of $\mathcal{F}$, because labels are ephemeral. In this sense, $\{\mathrm{p}, \Omega\}$ is just a way of labelling $\mathcal{F}$, and so there should be a property of invariance with respect to parameteric models. So first, a statement about when two apparently different models index the same $\mathcal{F}$.[18]

Definition 4.1 (Equivalent parametric models)**.** *The parametric models*

$$\{\mathrm{p}_\theta(\boldsymbol{x}; t),\, t \in \Omega\} \quad \textit{and} \quad \{\mathrm{p}_\phi(\boldsymbol{x}; s),\, s \in \Phi\}$$

are equivalent *exactly when there is a bijection $g : \Omega \to \Phi$ such that*

$$\mathrm{p}_\theta(\boldsymbol{x}; t) = \mathrm{p}_\phi(\boldsymbol{x}; s) \quad \textit{for } s = g(t).$$

For example, the Poisson model

$$\mathrm{p}_\theta(x; t) = e^{-t}\frac{t^x}{x!} \quad t \in \mathbb{R}_{++}$$

is equivalent to the model

$$\mathrm{p}_\phi(x; s) = e^{-e^s}\frac{e^{sx}}{x!} \quad s \in \mathbb{R}$$

because the two are linked by the bijection $g = \log$. This definition can now be used to state a principle of invariance.

Definition 4.2 (Model invariance)**.** *An inference about random quantities is* model-invariant *exactly when it is the same under equivalent parametric models.*

Now we can prove that the plug-in inference $\hat{\mathrm{E}}^*\{h(\boldsymbol{X})\}$ is model-invariant. But first we have to prove the extremely important result that the maximum likelihood (ML) estimate is invariant, although this has to be interpreted slightly differently from 'model-invariant', being a statement about parameters rather than about random quantities.

Theorem 4.6 (Invariance of maximum likelihood)**.**
The ML estimate is invariant*, in the sense that if $\hat{\theta}(q)$ and $\hat{\phi}(q)$ are the ML estimates of two equivalent parametric models for the dataset $Q := q(\boldsymbol{X})$, then $\hat{\phi}(q) = g(\hat{\theta}(q))$.*

Proof. The ML estimate of $\hat{\theta}(q)$ satisfies

$$\Pr_\theta\{Q; \hat{\theta}(q)\} \geq \Pr_\theta\{Q; t\} \qquad \text{for all } t \in \Omega.$$

As p_θ and p_ϕ are equivalent parametric models, substitute from (4.1) to give

$$\Pr_\phi\{Q; g(\hat{\theta}(q))\} \geq \Pr_\phi\{Q; g(t)\} \qquad \text{for all } t \in \Omega.$$

But the two sets $\{g(t) : t \in \Omega\}$ and Φ are equal because g is a bijection, and hence

$$\Pr_\phi\{Q; g(\hat{\theta}(q))\} \geq \Pr_\phi\{Q; s\} \qquad \text{for all } s \in \Phi,$$

showing that $g(\hat{\theta}(q))$ is the MLE of ϕ, as required. □

[18] In this definition and this section I suppress the $\boldsymbol{X}$ subscript to make space for the model label.

The next result follows straightforwardly.

Theorem 4.7. *The plug-in inference*

$$\hat{\mathrm{E}}^*_\theta\{h(\boldsymbol{X})\} := \mathrm{E}_\theta\{h(\boldsymbol{X}) \mid Q; \hat{\theta}(q)\}$$

is model-invariant.

Proof. Apply the Muddly table theorem (Thm 3.3) to verify that for equivalent p_θ and p_ϕ,

$$\mathrm{p}^*_\theta(\boldsymbol{x}; t) = \frac{\mathbb{1}_{q(x)}\, \mathrm{p}_\theta(\boldsymbol{x}; t)}{\Pr_\theta(Q; t)} = \frac{\mathbb{1}_{q(x)}\, \mathrm{p}_\phi(\boldsymbol{x}; s)}{\Pr_\phi(Q; s)} = \mathrm{p}^*_\phi(\boldsymbol{x}; s)$$

for $s = g(t)$. Hence $\mathrm{E}^*_\theta\{h(\boldsymbol{X}); t\} = \mathrm{E}^*_\phi\{h(\boldsymbol{X}); s\}$ for $s = g(t)$. Then

$$\begin{aligned}
\hat{\mathrm{E}}^*_\theta\{h(\boldsymbol{X})\} &= \mathrm{E}^*_\theta\{h(\boldsymbol{X}); \hat{\theta}(q)\} && \text{by definition} \\
&= \mathrm{E}^*_\phi\{h(\boldsymbol{X}); g(\hat{\theta}(q))\} && \text{see immediately above} \\
&= \mathrm{E}^*_\phi\{h(\boldsymbol{X}); \hat{\phi}(q)\} && \text{by Thm 4.6} \\
&= \hat{\mathrm{E}}^*_\phi\{h(\boldsymbol{X})\} && \text{by definition again,}
\end{aligned}$$

as required. □

It is important to appreciate that other types of parameter estimate may not have the invariance property (e.g. Method of Moments estimators), and thus using them as plug-ins would make the Frequentist inference sensitive to the labelling of $\mathcal{F}$, which must be undesirable.

* * *

As long as his inferential method is parameterisation-invariant, the statistician can choose whichever parameterisation of $\mathcal{F}$ is most convenient for him. This is an important practical point, because some choices of parameterisation are far more convenient than others. For example, when maximising over the parameter space to find the value of the MLE, it is very convenient if the parameter space can be written in a product form, i.e. if $\theta = (\theta_1, \dots, \theta_p)$ then

$$\Omega = \Omega_1 \times \cdots \times \Omega_p,$$

because it is far easier to explore rectangular regions than non-rectangular ones. In this case the parameters are said to be *variation independent*. For example, in the Normal statistical model we have $\theta = (\mu, \sigma^2) \in \Omega = \mathbb{R} \times \mathbb{R}_{++}$ for a parameterisation in terms of the expectation and the variance. Almost all practical models have parameters that are variation independent, although there is no theoretical reason for this property to be favoured.

Also, the statistician can choose a parameterisation in which

$$h^*(\theta) := \mathrm{E}^*\{h(\boldsymbol{X}); \theta\}$$

has a simple form (if this does not conflict with the previous property of variation independence). Forms such as $h^*(\theta) = \theta_1$, or some linear combination of the elements of θ, are popular. So popular, in fact, that a label is available for those elements of θ which are not in $h^*(\theta)$: they are called *nuisance parameters*. 'Old fashioned' textbooks devote a lot of material to particular families and parameterisations in which the nuisance parameters can be circumvented.[19] This material has largely been superseded by empirical methods such as the *bootstrap* (see, e.g., Davison *et al.*, 2003), or by a Bayesian approach, as discussed in Sec. 4.5.

Set estimators. Thm 4.6 showed that the ML estimate is invariant, in the sense that the results from equivalent models map onto each other. The same property can be demanded of other types of inference about the parameters. Here is a very useful invariance result for estimates of θ that are sets in Ω rather than points (it generalises Thm 4.6).

Theorem 4.8 (Invariance of likelihood function level sets). *The level sets of the likelihood function are invariant, in the sense that if*

$$\mathcal{C}_\theta := \{t : \Pr_\theta(Q; t) \geq c\} \quad \textit{and} \quad \mathcal{C}_\phi := \{s : \Pr_\phi(Q; s) \geq c\}$$

are level sets from two equivalent models, then $g(\mathcal{C}_\theta) = \mathcal{C}_\phi$.

Proof. Fix c. Then

$$\begin{aligned} \mathcal{C}_\theta &= \{t : \Pr_\theta(Q; t) \geq c\} \\ &= \{t : \Pr_\phi(Q; g(t)) \geq c\} && \text{if } \mathrm{p}_\theta \text{ and } \mathrm{p}_\phi \text{ are equivalent} \\ &= \{g^{-1}(s) : \Pr_\phi(Q; s) \geq c\} && \text{as } g \text{ is bijective} \end{aligned}$$

and hence $g(\mathcal{C}_\theta) = \{s : \Pr_\phi(Q; s) \geq c\} = \mathcal{C}_\phi$, as required. □

This result will be used in the definition of confidence sets (Sec. 8.5). Concerning the latter, note that 'Wald-style' confidence intervals of the form

$$\hat{\theta}(q) \mp k \times \widehat{\mathrm{SE}}_{\hat{\theta}(q)}$$

for some constant k (like 1.96), where $\widehat{\mathrm{SE}}_{\hat{\theta}(q)}$ is the estimated standard error, are *not* invariant in the sense given in Thm 4.8. Again, this is undesirable. Twenty years ago this would have been forgivable owing to computational cost, because approximating the estimated standard error is much cheaper than identifying the limits of an appropriate level set. These days, though, computational cost should not be an issue.

[19] See, e.g., Cox and Hinkley (1974). I say 'old fashioned', but this remains one of my favourite textbooks.