

II Objectivity and Conditionality in Frequentist Inference

David Cox and Deborah G. Mayo

1 Preliminaries

Statistical methods are used to some extent in virtually all areas of science, technology, public affairs, and private enterprise. The variety of applications makes any single unifying discussion difficult if not impossible. We concentrate on the role of statistics in research in the natural and social sciences and the associated technologies. Our aim is to give a relatively non-technical discussion of some of the conceptual issues involved and to bring out some connections with general epistemological problems of statistical inference in science. In the first part of this [chapter \(7\(I\)\)](#), we considered how frequentist statistics may serve as an account of inductive inference, but because this depends on being able to apply its methods to appropriately circumscribed contexts, we need to address some of the problems in obtaining the methods with the properties we wish them to have. Given the variety of judgments and background information this requires, it may be questioned whether any account of inductive learning can succeed in being “objective.” However, statistical methods do, we think, promote the aim of achieving enhanced understanding of the real world, in some broad sense, and in this some notion of objectivity is crucial. We begin by briefly discussing this concept as it arises in statistical inference in science.

2 Objectivity

Objectivity in statistics, as in science more generally, is a matter of both aims and methods. Objective science, in our view, aims to find out what is the case as regards aspects of the world, independently of our beliefs, biases, and interests; thus objective methods aim for the critical control of inferences and hypotheses, constraining them by evidence and checks of error.

The statistician is sometimes regarded as the gatekeeper of objectivity when the aim is to learn about those aspects of the world that exhibit haphazard variability, especially where methods take into account the uncertainties and errors by using probabilistic ideas in one way or another. In one form, probability arises to quantify the relative frequencies of errors in a hypothetical long run; in a second context, probability purports to quantify the “rational” degree of belief, confirmation, or credibility in hypotheses. In the “frequentist” approach, the aim of objective learning about the world is framed within a statistical model of the process postulated to have generated data.

Frequentist methods achieve an objective connection to hypotheses about the data-generating process by being constrained and calibrated by the method’s error probabilities in relation to these models: the probabilities derived from the modeled phenomena are equal or close to the actual relative frequencies of results in applying the method. In the second, degree of belief construal by contrast, objectivity is bought by attempting to identify ideally rational degrees of belief controlled by inner coherency. What are often called “objective” Bayesian methods fall under this second banner, and many, although of course not all, current Bayesian approaches appear to favor the use of special prior probabilities, representing in some sense an indifferent or neutral attitude (Berger, 2004). This is both because of the difficulty of eliciting subjective priors and because of the reluctance among scientists to allow subjective beliefs to be conflated with the information provided by data. However, since it is acknowledged that strictly noninformative priors do not exist, the “objective” (or default) priors are regarded largely as conventionally stipulated reference points to serve as weights in a Bayesian computation. We return to this issue in Section 11.

Given our view of what is required to achieve an objective connection to underlying data-generating processes, questions immediately arise as to how statistical methods can successfully accomplish this aim. We begin by considering the nature and role of statistical analysis in its relations to a very general conception of learning from data.

3 Roles of Statistics

Statistical methods, broadly conceived, are directed to the numerous gaps and uncertainties scientists face in learning about the world with limited and fallible data. Any account of scientific method that begins its work only once well-defined evidence claims and unambiguous hypotheses and theories are available forfeits the ability to be relevant to understanding

the actual processes behind the success of science. Because the contexts in which statistical methods are most needed are ones that compel us to be most aware of errors and threats to reliability, considering the nature of statistical methods in the collection, modeling, and analysis of data is a good way to obtain a more realistic account of science. Statistical methods are called on at a variety of stages of inquiry even in explorations where only a vague research question is contemplated. A major chapter in statistical theory addresses the design of experiments and observational studies aiming to achieve unambiguous conclusions of as high a precision as is required. Preliminary checks of data quality and simple graphical and tabular displays of the data are made; sometimes, especially with very skillful design, little additional analysis may be needed. We focus, however, on cases where more formal analysis is required, both to extract as much information as possible from the data about the research questions of concern and to assess the security of any interpretation reached.

A central goal behind the cluster of ideas we may call frequentist methods is to extract what can be *learned from data* that can also be vouched for. Essential to this school is the recognition that it is typically necessary to *communicate* to others what has been learned and its associated uncertainty. A fundamental requirement that it sets for itself is to provide means to address legitimate critical questions and to give information about which conclusions are likely to stand up to further probing and where weak spots remain. The whole idea of Fisherian theory implicitly and of Neyman–Pearson more explicitly is that formal methods of statistical inference become relevant primarily when the probing one can otherwise accomplish is of relatively borderline effectiveness, so that the effects are neither totally swamped by noise nor so clear-cut that formal assessment of errors is relatively unimportant. The roles played by statistical methods in these equivocal cases are what make them especially relevant for the epistemological question of how reliable inferences are possible despite uncertainty and error. Where the recognition that data are always fallible presents a challenge to traditional empiricist foundations, the cornerstone of statistical induction is the ability to move from less accurate to more accurate data. Fisher put it thus:

It should never be true, though it is still often *said*, that the conclusions are no more accurate than the data on which they are based. Statistical data are always erroneous, in greater or less degree. The study of inductive reasoning is the study of the embryology of knowledge, of the processes by means of which truth is extracted from its native ore in which it is fused with much error. (Fisher, 1935, p. 39)

4 Formal Statistical Analysis

The basic inferential goal shared by formal statistical theories of inference is to pose and answer questions about aspects of statistical models in light of the data. To this end, empirical data, denoted by \mathbf{y} , are viewed as the observed value of a vector random variable \mathbf{Y} . The question of interest may be posed in terms of a probability distribution of \mathbf{Y} as defined by the relevant statistical model. A model (or family of models) gives the probability distribution (or density) of \mathbf{Y} , $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$, which may be regarded as an abstract and idealized representation of the underlying data-generating process. Statistical inferences are usually couched in terms of the unknown parameter $\boldsymbol{\theta}$.

Example 1. *Bernoulli trials.* Consider n independent trials (Y_1, Y_2, \dots, Y_n) , each with a binary outcome – success or failure – where the probability of success at each trial is an unknown constant θ with a value between 0 and 1. This model is a standard probability model which is often used to represent “coin-tossing” trials, where the hypothesis of a “fair” coin is $\theta = .5$. It serves as a standard in modeling aspects of many cases that are appropriately analogous.

Crucial conceptual issues concern the nature of the probability model and in particular the role of probability in it. An important concept that arises in all model-based statistical inference is that of *likelihood*. If \mathbf{y} is a realized data set from $f(\mathbf{y}; \boldsymbol{\theta})$, the likelihood is a function of $\boldsymbol{\theta}$ with \mathbf{y} fixed: $\text{Lik}(\boldsymbol{\theta}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})$.

In the case of binomial trials, the data $\mathbf{y} = (y_1, y_2, \dots, y_n)$ forms a sequence of r “successes” and $n - r$ “failures,” with the $\text{Lik}(\theta; \mathbf{y}) = \theta^r(1 - \theta)^{n - r}$.

Likelihoods do not obey the probability laws: for example, the sum of the likelihoods of a hypothesis and its denial is not 1.

4.1 Three Different Approaches

Three broad categories represent different approaches that are taken regarding model-based statistical inference. Each has given rise to philosophical and methodological controversies that have rumbled on for anywhere from fifty to two hundred years, which we do not plan to review here.¹ The following discussion is merely an outline of a few of the issues involved, which sets the stage for elucidating the unifying principle that enables frequentist

¹ See references in 7(IV).

methods, when properly formulated, to obtain objective information about underlying data-generating processes.

Likelihood Methods. The first approach rests on a comparative appraisal of rival statistical hypotheses H_0 and H_1 according to the ratio of their likelihoods. The basic premise is what Hacking (1965) called the *law of likelihood*: that the hypothesis with the higher likelihood has the higher evidential “support,” is the more “plausible,” or does a better job of “explaining” the data. The formal analysis is based on looking at ratios of likelihoods of two different parameter values $f_S(\mathbf{s}; \theta_1)/f_S(\mathbf{s}; \theta_0)$; the ratios depend only on statistic \mathbf{s} , and are regarded as comparative summaries of “what the data convey” about θ .

Bayesian Methods. In Bayesian approaches, the parameter θ is modeled as a realized value of a random variable Θ with a probability distribution $f_\Theta(\theta)$, called the prior distribution. Having observed \mathbf{y} , inference proceeds by computing the conditional distribution of Θ , given $\mathbf{Y} = \mathbf{y}$, the posterior distribution.

Under this broad category, two notable subgroups reflect contrasting uses of probability: (1) to represent personalistic degree of belief and (2) to represent impersonal or rational degree of belief or some broadly equivalent notion. The central point is that the focus of interest, ψ , is typically an unknown constant, usually a component of θ , and if we were to aim at talking about a probability distribution for ψ , an extended notion of probability, beyond immediately frequentist concepts, would usually then be unavoidable.

With a generalized notion of probability as degree of belief, it is possible, formally at least, to assign a probability distribution to ψ given the data. Many purport to use this notion to assess the strength of evidence or degree of credibility that some hypothesis about ψ is in some sense true. This is done by what used to be called inverse probability and is nowadays referred to as a Bayesian argument. Once the relevant probabilities are agreed on, the calculation uses Bayes’s theorem, an entirely uncontroversial result in probability theory that stems immediately from the definition of conditional probability. We obtain the posterior distribution of the full parameter by multiplying the likelihood by the prior density and then multiplying by a suitable constant to make the total posterior probability equal to 1.

Frequentist (Sampling) Methods. The third paradigm makes use of the frequentist view of probability to characterize methods of analysis by means

of their performance characteristics in a hypothetical sequence of repetitions. Two main formulations are used in the frequentist approach to the summarization of evidence about the parameter of interest, ψ . For simplicity, we suppose from now on that for each research question of interest ψ is one-dimensional.

The first is the provision of sets or intervals within which ψ is in some sense likely to lie (confidence intervals) and the other is the assessment of concordance and discordance with a specified value ψ_0 (significance tests). Although we concentrate on the latter, the estimation of ψ via sets of intervals at various confidence levels is the preferred method of analysis in many contexts. The two are closely connected, as our interpretation of tests makes plain: confidence intervals consist of parameter values that are not inconsistent with the data at specified levels. Parameter values outside a $(1-c)$ confidence interval are those that contradict the data at significance level c .

4.2 Our Focus

We concentrate here on the frequentist approach. The personalistic approach, whatever merits it may have as a representation of personal belief and personal decision making, is in our view inappropriate for the public communication of information that is the core of scientific research, and of other areas, too. To some extent, the objectivist Bayesian view claims to address the same issues as the frequentist approach; some of the reasons for preferring the frequentist approach are sketched in Section 12. The likelihood approach also shares the goal to learn “what can be said” about parameters of interest; however, except for very simple problems, the pure likelihood account is inadequate to address the complications common in applications and we do not specifically discuss it here.

The key difference between the frequentist approach and the other paradigms is its focus on the sampling distribution of the test (or other) statistic (i.e., its distribution in hypothetical repetition). In our view, the sampling distribution, when properly used and interpreted, is at the heart of the objectivity of frequentist methods. We will discuss the formulation and implementation of these methods in order to address central questions about the relevance and appropriate interpretation of its core notions of hypothetical repetitions and sampling distributions. In the first part of this chapter (Mayo and Cox), we considered the reasoning based on p -values; our considerations now pertain to constructing tests that (1) permit p -values to be calculated under a variety of null hypotheses and (2)

ensure the relevance of the hypothetical long run that is used in particular inferences.

5 Embarking on Formal Frequentist Analysis

In a fairly wide variety of contexts, the formal analysis may be seen to proceed broadly as follows. First, we divide the features to be analyzed into two parts and denote their full set of values collectively by \mathbf{y} and by \mathbf{x} , which are typically multidimensional. A probability model is formulated according to which \mathbf{y} is the observed value of a vector random variable \mathbf{Y} whose distribution depends on \mathbf{x} , regarded as fixed. Note that especially in observational studies it may be that \mathbf{x} could have been regarded as random but, given the question of interest, we chose not to do so. This choice leads to the relevant sampling distribution.

Example 2. *Conditioning by model formulation.* For a random sample of men, we measure systolic blood pressure, weight, height, and age. The research question may concern the relation between systolic blood pressure and weight, allowing for height and age, and if so, specifically for that question, one would condition on the last three variables and represent by a model the conditional distribution of Y , systolic blood pressure, given \mathbf{x} , the other three variables: $f(y | x_1, x_2, x_3; \theta)$. The linear regression

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i, \quad i = 1, 2, \dots, n,$$

is an example of a statistical model based on such a conditional distribution. Here the unknown parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ represent the effect of changing one explanatory variable while the others are held fixed and are typically the focus of interest. The u_i ($i = 1, 2, \dots, n$) are not directly observed random terms of zero expectation, representing the haphazard component of the variation.

One would condition on the explanatory variables even if, say, one knew the distribution of age in the population. We may call this *conditioning by model formulation*. This serves to constrain what is allowed to vary conceptually in determining the sampling distribution for inference.

Of course, for different purposes the additional information of how x varies *would* be relevant and the appropriate statistical model would reflect this. For example, if one were interested in the correlation between systolic blood pressure and age, the relevant distribution would be the joint distribution of Y and X_3 , say $f(y, x_3; \psi)$. In that case, the question is how

the two variables covary with each other and, thus, the sample space should include all their possible values and the associated probabilities.

Here we consider parametric models in which the probability density of Y is in the form $f_Y(y; \theta)$, where θ is typically a vector of parameters $\theta = (\psi, \lambda)$. Dependence on x is not shown explicitly in this notation. The *parameter of interest*, ψ , addresses the research question of concern; additional parameters are typically needed to complete the probability specification. Because these additions may get in the way of the primary focus of research, they are dubbed *nuisance parameters*. In contrasting different statistical approaches, we draw some comparisons based on how each handles such nuisance parameters.

Virtually all such models are to some extent provisional, which is precisely what is expected in the building up of knowledge. Probability models range from purely empirical representations of the pattern of observed haphazard variability to representations that include substantial elements of the underlying science base. The former get their importance from providing a framework for broad families of statistical methods – for example, some form of regression analysis – that find fruitful application across many fields of study. The latter provide a stronger link to interpretation. An intermediate class of models of increasing importance in observational studies, especially in the social sciences, represents a potential data-generating process and, hence, may point toward a causal interpretation. Parameters – especially the parameters of interest, ψ – are intended to encapsulate important aspects of the data-generating process separated off from the accidents of the specific data under analysis. Probability is to be regarded as directly or indirectly based on the empirical stability of frequencies under real or hypothetical repetition. This allows for a wide latitude of imaginative applications of probability, not limited to phenomena exhibiting actual repetitions.

Example 3. Cox and Brandwood (1959) put in order, possibly of time, the works of Plato, taking *Laws* and *Republic* as reference points. The data were the stresses on the last five syllables of each sentence and these were assumed to have in each book a probability distribution over the thirty-two possibilities. What does probability mean in such a case?

This is a good example of how historical questions, lacking literal repetitions, may be tackled statistically: an attribute such as the relative frequencies of the thirty-two ending types, together with deliberately chosen reference standards, may be used to discriminate patterns. By taking two large known works of Plato, between which the works being dated were written, as giving probability distributions for the thirty-two possible endings, one can assign

a (stylometric) score to the relative frequencies of ending types observed in the works whose relative dates are unknown. This allows one to determine what would be expected *statistically* were the two works identical with respect to the particular stylometric score, and thereby to probe null hypotheses of form “these two works are identical (with respect to the time written).” The other works may thus be ordered according to their differences from, or affinity to, the two reference standards (given assumptions about changes in literary style).

Two general challenges now arise. How do we use the data as effectively as possible to learn about θ ? How can we check on the appropriateness of the model?

6 Reducing the Data by Sufficiency

Suppose then that data $\mathbf{y} = (y_1, y_2, \dots, y_n)$ are modeled as a realization of random variable Y , and that a family F of possible distributions is specified; we seek a way to reduce the data so that what we wish to learn about the unknown θ may be “extracted from its native ore.” We seek a function of the data, a *statistic* $S(\mathbf{Y})$, such that knowing its value $s(\mathbf{y})$ would suffice to encompass the statistical information in the n -dimensional data as regards the parameter of interest. This is called a *sufficient* statistic. We aim to choose a sufficient statistic that minimizes the dimensionality of $\mathbf{s} = s(\mathbf{y})$ such that the distribution of the sample factorizes as

$$f(\mathbf{y}; \theta) = f_S(s; \theta) f_{Y|S}(\mathbf{y} | s),$$

where $f_{Y|S}(\mathbf{y} | s)$, the conditional distribution of \mathbf{Y} given the value of S , does not depend on the unknown parameter θ . In other words, knowing the distribution of the sufficient statistic S suffices to compute the probability of any given \mathbf{y} . The process of reducing the data in this way may be called *reduction by sufficiency*.

Example 4. Binomial model. Consider n independent trials (Y_1, Y_2, \dots, Y_n) , each with a binary outcome (success or failure), where the probability of success is an unknown constant θ . These are called Bernoulli trials. The sufficient statistic in this case is $S = \sum_{k=1}^n Y_k$, the number of successes, and has a binomial sampling distribution determined by the constants n and θ . It can be shown that the distribution of the sample reduces as earlier, where $f_S(s; \theta)$ is a binomial, and $f_{Y|S}(\mathbf{y} | s)$ is a discrete uniform, distribution; all permutations of the sequence of successes and failures are equally likely. We

now argue as follows. The experiment would be equivalent to having been given the data $\mathbf{y} = (y_1, y_2, \dots, y_n)$ in two stages:

First, we are told the value of $s(\mathbf{y})$ (e.g., $S = s$ successes out of n Bernoulli trials). Then some inference can be drawn about θ using the sampling distribution of S , $f_S(s; \theta)$, in some way.

Second, we learn the value of the remaining parts of the data (e.g., the first k trials were all successes, the rest failures). Now, if the model is appropriate, then the second phase is equivalent to a random draw from a totally known distribution and could just as well be the outcome of a random number generator. Therefore, all the information about θ is, so long as the model is appropriate, locked in s and in the dependence on θ of the distribution of the random variable S .

This second stage is essentially to observe a realization of the conditional distribution of Y given $S = s$, generated by observing \mathbf{y} in the distribution $f_{Y|S}(\mathbf{y} | s)$. Because this conditional distribution is totally known, it can be used to assess the validity of the assumed model. Insofar as the remaining parts of the data show discordancy, in some relevant respect, with being from the known distribution, doubt is thrown on the model. In Example 4, for instance, the fact that any permutation of the r successes in n trials has known probability, assuming the correctness of the model, gives us a standard to check if the model is violated. It is crucial that any account of statistical inference provides a conceptual framework for this process of model criticism, even if in practice the criticism is often done relatively informally. The ability of the frequentist paradigm to offer a battery of simple significance tests for model checking and possible improvement is an important part of its ability to supply objective tools for learning.

The appropriate reduction to a sufficient statistic s is usually best found by considering the likelihood function, which is the probability of the data considered as a function of the unknown parameter θ . The aim is to factor this as a function of s times a function of \mathbf{y} not involving θ . That is, we wish to write the following:

$$f_Y(\mathbf{y}; \theta) = m_1(\mathbf{y})m_2(s; \theta),$$

say, taking the minimal s for which this factorization holds. One important aspect is that, for any given \mathbf{y} from distribution $f_Y(\mathbf{y}; \theta)$, in the *same experiment*, the ratio of likelihoods at two different values of θ depends on the data only through s .

7 Some Confusion over the Role of Sufficiency

Sufficiency as such is not specifically a frequentist concept. Unfortunately some confusion has appeared in the literature over the role of sufficiency in frequentist statistics. We can address this by considering two contrasting experimental procedures in relation to Example 4.

Example 5. *Binomial versus negative binomial.* Consider independent Bernoulli trials, where the probability of success is an unknown constant θ , but imagine two different experimental procedures by which they may be produced. Suppose first that a preassigned number of trials, n , is observed. As noted earlier, the sufficient statistic is r , the number of successes, and this is the observed value of a random variable R having a *binomial* distribution with parameters n and θ . Suppose now instead that trials continue until a preassigned number of successes, r has occurred after n trials. Such an observational process is often called inverse sampling. In this second case the sufficient statistic is n , the observed value of a random variable N having a *negative binomial distribution* determined by the constants r and θ . We may denote the two experiments by E_N and E_R , respectively.

Now it has been argued that, because r and n determine the likelihood in the same form proportional to $\theta^n(1 - \theta)^{n-r}$, whether arising from E_N or E_R , that in both cases the same inference should be drawn. It is clear, however, from the present perspective that the roles of n and r are quite different in the two situations and there is no necessary reason to draw the same conclusions. Experiments E_N and E_R have different sample spaces, and because the sampling distributions of the respective sufficient statistics differ, the same string of successes and failures would result in a difference in p -values (or confidence-level) assessments, depending on whether it arose from E_N or E_R , although the difference is typically minor. Perhaps the confusion stems in part because the various inference schools accept the broad, but not the detailed, implications of sufficiency: the difference emanates from holding different notions of inference. We now explain this.

7.1 Sufficiency Principle (General)

If random variable \mathbf{Y} , in a given experiment E , has probability density $f_y(\mathbf{y}; \boldsymbol{\theta})$ and \mathbf{S} is minimal sufficient for $\boldsymbol{\theta}$, then as long as the model for E is adequate, identical inferences about $\boldsymbol{\theta}$ should be drawn from data \mathbf{y}' and \mathbf{y}'' whenever \mathbf{y}' and \mathbf{y}'' yield the same value of \mathbf{s} .

We may abbreviate this as follows:

If \mathbf{s} is minimal sufficient for θ in experiment E , and $\mathbf{s}(\mathbf{y}') = \mathbf{s}(\mathbf{y}'')$, then the inference from \mathbf{y}' and \mathbf{y}'' about θ should be identical; that is, $\text{Infr}_E(\mathbf{y}') = \text{Infr}_E(\mathbf{y}'')$.

However, when proposing to apply the sufficiency principle to a particular inference account, the relevant method for inference must be taken into account. That is, Infr_E is relative to the inference account.

7.2 Sufficiency in Sampling Theory

If a random variable \mathbf{Y} , in a given experiment E , arises from $f(\mathbf{y}; \theta)$, and the assumptions of the model are valid, then all the information about θ contained in the data is obtained from consideration of its minimal sufficient statistic \mathbf{S} and its *sampling distribution* $f_{\mathbf{S}}(\mathbf{s}; \theta)$.

An inference in sampling theory, therefore, needs to include the relevant sampling distribution, whether it was for testing or estimation. Thus, in using the abbreviation $\text{Infr}_E(\mathbf{y})$ to refer to an inference from \mathbf{y} in a sampling theory experiment E , we assume for simplicity that E includes a statement of the probability model, parameters, and sampling distribution corresponding to the inference in question. This abbreviation emphasizes that the inference that is licensed is relative to the particular experiment, the type of inference, and the overall statistical approach being discussed.²

In the case of frequentist sampling theory, features of the experiment that alter the sampling distribution must be taken account of in determining what inferences about θ are warranted, and when the same inferences from given experiments may be drawn. Even if \mathbf{y}' and \mathbf{y}'' have proportional likelihoods but are associated with different relevant sampling distributions, corresponding to E' and E'' , \mathbf{y}' and \mathbf{y}'' each provides different relevant information for inference. It is thus incorrect to suppose, within the sampling paradigm, that it is appropriate to equate $\text{Infr}_{E'}(\mathbf{y}')$ and $\text{Infr}_{E''}(\mathbf{y}'')$.

These points show that sampling theory violates what is called the *strong likelihood principle*.

² This abbreviation, like Birnbaum's $\text{Ev}(E, \mathbf{x})$, may be used to discuss general claims about principles of evidence. Birnbaum's $\text{Ev}(E, \mathbf{x})$, "the evidence about the parameter arising from experiment E and result \mathbf{x} ," is, for Birnbaum, the inference, conclusion or report, and thus is in sync with our notion (Birnbaum, 1962).

We prefer it because it helps avoid assuming a single measure of "the" evidence associated with an experimental outcome. By referring to the inference licensed by the result, it underscores the need to consider the associated methodology and context.

7.3 The Strong Likelihood Principle (SLP)

Suppose that we have *two* experiments, E' and E'' , with different probability models $f'_{Y'}(\mathbf{y}'; \theta)$ and $f''_{Y''}(\mathbf{y}''; \theta)$, respectively, with the same unknown parameter θ . If \mathbf{y}'^* and \mathbf{y}''^* are observed data from E' and E'' , respectively, where the likelihoods of \mathbf{y}'^* and \mathbf{y}''^* are proportional, then \mathbf{y}'^* and \mathbf{y}''^* have the identical evidential import for any inference about θ .

Here proportionality means that, for all θ , $f''_{Y''}(\mathbf{y}''; \theta)/f'_{Y'}(\mathbf{y}'; \theta)$ is equal to a constant that does not depend on θ . A sample of, say, six successes in twenty trials would, according to the SLP, have the identical evidential import whether it came from a binomial experiment, with sample size fixed at twenty, or from a negative binomial experiment where it took twenty trials to obtain six successes.

By contrast, suppose a frequentist is interested in making an inference about θ on the basis of data \mathbf{y}' consisting of r successes in n trials in a binomial experiment E' . Relevant information would be lost if the report were reduced to the following: there were r successes in n Bernoulli trials, generated from *either* a binomial experiment with n fixed, \mathbf{y}'^* , or a negative binomial experiment with r fixed, \mathbf{y}''^* – concealing which was actually the source of the data. Information is lost because $\text{Infr}_{E'}(\mathbf{y}'^*)$ is *not* equal to $\text{Infr}_{E''}(\mathbf{y}''^*)$ due to the difference in the associated sampling distributions. Equivalences that hold with respect to a single experiment, as is the case with sufficiency, cannot be assumed to hold in comparing data from different experiments.

8 Sufficient Statistics and Test Statistics

How then are we to extract answers to the research question out of $f_S(\mathbf{s}; \boldsymbol{\theta})$; all that the reduction to \mathbf{s} has done is to reduce the dimensionality of the data. To establish a significance test, we need to choose an appropriate test statistic $T(\mathbf{Y})$ and find a distribution for assessing its concordancy with H_0 . To warrant the interpretations of the various significance tests that we delineated in the first part of this chapter (Mayo and Cox), we need to consider how to identify test statistics to construct appropriate tests.

To interpret t , the observed value of T , we compare it with its predicted value under the null hypothesis by finding, for any observed value t , $p = P(T \geq t; H_0)$. That is, we examine how extreme t is in its probability distribution under H_0 . Thus, we need both to choose an appropriate test statistic $T(\mathbf{Y})$ and also to compute its distribution in order to compare t with what is expected under H_0 . To this end we find a suitable feature t of the data,

in light of the previous discussion, a function of sufficient statistic s , such that

- [1] The larger the value of t , the greater the discrepancy with the null hypothesis in the respect of concern.
- [2] The probability distribution of the random variable T is exactly known when the null hypothesis is true, so that in particular the distribution does not depend on nuisance parameters.

We must then collect data \mathbf{y} to compute $t(\mathbf{y})$, ensuring the data satisfy adequately the assumptions of the relevant probability model. Note that the p -value can itself be regarded as a random variable P ; and the probability that P takes different values under alternatives to the null hypothesis may be calculated.

To satisfy condition [1], the larger the value of t , the smaller the corresponding p -value must be. Satisfying condition [2] is the frequentists' way of ensuring as far as possible that observed discordancies are attributable to discrepancies between the null hypothesis and the actual phenomena giving rise to the data. This is key to avoiding ambiguities in pinpointing the source of observed anomalies (Duhem's problem).

The choice of test statistic depends on the type of null hypothesis involved (see the delineation in 7(I), p. 257). We deal first with an important situation where an essentially unique answer is possible. Suppose there is a full model covering both null and alternative possibilities. With such "embedded" nulls, there is formulated not only a probability model for the null hypothesis but also models that represent other possibilities in which the null hypothesis is false and usually, therefore, represent possibilities whose presence we would wish to detect.

Among the number of possible situations, a common one involves a parametric family of distributions indexed by an unknown parameter θ . Suppose that θ is one-dimensional. We reduce by sufficiency to $S(\mathbf{Y})$. If $S(\mathbf{Y})$ itself is one-dimensional, the test statistic must be a function of $S(\mathbf{Y})$ and we can almost always arrange that $S(\mathbf{Y})$ itself can be taken as the test statistic and its distribution thus found.

The null hypothesis is typically not logically contradicted however far t is from what is expected under the null hypothesis except in those rare cases where certain values of t are logically impossible under the null hypothesis; however, the p -value indicates the level at which the data contradict the null hypothesis. In selecting tests or, in the embedded case, corresponding confidence limits, two perspectives are possible. One focuses on being able to give objective guarantees of low long-run error rates and optimality

properties that hold regardless of unknown nuisance parameters. A second focuses on being able to objectively determine how consistent data are from various values of the parameter of interest. The former relates to the behavioristic perspective traditionally associated with Neyman–Pearson theory, the latter with the inductive inference perspective that we advance here.

Consider generating a statistic for the $1 - \alpha$ upper confidence bound, $CI^U(\mathbf{Y}; \alpha)$ for estimating a normal mean μ . This statistic is directly related to a test of $\mu = \mu_0$ against $\mu < \mu_0$. In particular, Y is statistically significantly smaller than those values of μ in excess of $CI^U(\mathbf{Y}; \alpha)$ at level α . Mathematically, the same intervals emerge from following the Neyman–Pearson or Fisherian perspective. Both aim to guarantee the sensitivity of the analysis by ensuring $P(\mu' < CI^U(\mathbf{Y}; \alpha))$ is minimal for $\mu' > \mu$, subject to the requirement that, with high probability $(1 - \alpha)$, $CI^U(\mathbf{Y}; \alpha)$ exceeds the true value of μ . That is,

$$P(\mu < CI^U(\mathbf{Y}; \alpha)) = 1 - \alpha.$$

To contrast the differences in interpretation and justification, consider forming $CI^U(\mathbf{Y}; \alpha)$ for a normal mean where the variance σ^2 is known. The observed upper limit is $\bar{y}_0 + k(\alpha)\sigma_y$, where $\bar{y}_0 = \sum_{k=1}^n y_k$, $k(\alpha)$ is the upper α -point of the standard normal distribution, and $\sigma_y = \sigma/\sqrt{n}$. Consider the inference $\mu < \bar{y}_0 + k(\alpha)\sigma_y$. One rationale that may be given to warrant this inference is that it instantiates an inference rule that yields true claims with high probability $(1 - \alpha)$ since

$$P(\mu < \bar{Y} + k(\alpha)\sigma_y) = 1 - \alpha.$$

The procedure, it is often said, has high long-run “coverage probabilities.” A somewhat different justification, based on the same probabilistic facts, is to view $\mu \leq \bar{y}_0 + k(\alpha)\sigma_y$ as an inference based on a type of *reductio ad absurdum* argument: suppose in fact that this inference is false and the true mean is μ' , where $\mu' > \bar{y}_0 + k(\alpha)\sigma_y$. Then it is very probable that we would have observed a larger sample mean since

$$P(\bar{Y} > \bar{y}_0; \mu') > 1 - \alpha.$$

Therefore, one can reason, \bar{y}_0 is inconsistent at level $(1 - \alpha)$, with having been generated from a population with μ in excess of the upper confidence limit. This reasoning is captured in the frequentist principle of evidence FEV that we set out in 7(I), p. 254.

The Neyman–Pearson formulation arrives at essentially the same test or confidence interval but proceeds in a sense in the opposite direction. Rather than beginning with sufficient statistic S , optimality criteria are set up for

arriving at the most sensitive analysis possible with the data. Solving the optimality problem, one arrives at a procedure in which the data enter via sufficient statistic S . In the Neyman–Pearson theory, sensitivity is assessed by means of the power – the probability of reaching a preset level of significance under the assumption that various alternative hypotheses are true. In the approach described here, sensitivity is assessed by means of the distribution of the random variable P , considered under the assumption of various alternatives. In confidence intervals, corresponding sensitivity assessments are not directly in terms of length of intervals but rather in terms of the probability of including false values of the parameter.

The two avenues to sufficient statistic s often lead to the same destination but with some differences of interpretation and justification. These differences can lead to more flexible specifications and uses of the same statistical tools. For example, it suffices for our purposes that the error probabilities are only approximate. Whereas Neyman–Pearson confidence intervals fix a single confidence level for a parameter of interest, in the current approach one would want to report several confidence limits at different levels. These benchmarks serve to more fully convey what the data are saying with respect to which values are, and are not, consistent with the data at different levels.

This interpretation of confidence intervals also scotches criticisms of examples where, due to given restrictions, it can happen that a $(1 - \alpha)$ estimate contains all possible parameter values. Although such an inference is “trivially true,” it is scarcely vacuous in our construal. That all parameter values are consistent with the data is an informative statement about the limitations of the data to detect discrepancies at the particular level.

9 Conditioning for Separation from Nuisance Parameters

In most realistic situations there is a nuisance parameter λ in addition to the parameter of interest. In this section we consider the formulation of tests to accommodate such nuisance parameters – first from the current perspective and then in their relation to tests developed from the traditional Neyman–Pearson perspective. In order to take a small value of p as evidence that it is due to a discrepancy between the null hypothesis and the actual data-generating procedure, we need a test statistic with a distribution that is split off from that of the unknown nuisance parameter λ . The parameter θ may be partitioned into components $\theta = (\psi, \lambda)$ such that the null hypothesis is that $\psi = \psi_0$, where λ is an unknown nuisance parameter. Interest may focus on alternatives $\psi > \psi_0$.

In this case one aim is to achieve a factorization $\mathbf{s} = (t, v)$, where t is one-dimensional such that

- the random variable V has a distribution depending only on λ and
- the conditional distribution of T given $V = v$ depends only on ψ .

In constructing significance tests, these conditions may sometimes be achieved by conditioning on a sufficient statistic V for the nuisance parameter, thereby reducing the null hypothesis to a simple hypothesis, where ψ is the only unknown. The test statistic is $T|V$ (i.e., T given $V = v$). Although the distribution of V depends on an unknown, the fact that it is disconnected from the parameter under test, ψ , allows values of p for a hypothesis about ψ (e.g., $H_0: \psi = \psi_0$) to be calculated from this conditional distribution, $T|V$. This may be called *technical conditioning for separation from nuisance parameters*. It has the additional advantage that it largely determines the appropriate test statistic by the requirement of producing the most sensitive test possible with the data at hand.

Example 6. *Conditioning for separation from nuisance parameters.* Suppose that Y_1 and Y_2 have independent Poisson distributions of means μ_1 and μ_2 , respectively, but that it is only the ratio of the means, μ_1/μ_2 , that is of interest; that is, the null hypothesis concerns $\psi = \mu_1/\mu_2$ and, therefore, $H_0: \psi = \psi_0$. Thus, the nuisance parameter λ is $\mu_1 + \mu_2$. In fact, for any given value of ψ , it can be shown that there is a sufficiency reduction to $V = y_1 + y_2$. That is, for any given value of ψ , the observed value v contains all the information about nuisance parameter λ . (V is a *complete sufficient statistic* for λ .) There is a factorization into information about λ and the complementary term of the distribution of, say, Y_1 given $V = v$, which depends only on ψ and, thus, contains all the information about ψ so long as there is no other information about the nuisance parameter. The variable Y , given $V = v$ has a binomial distribution with probability of success $\psi / (\psi + 1)$; accordingly, the test rejects the null hypothesis for large values of γ . This conditional distribution serves as our test statistic for the hypothesis of interest. In addition to achieving separation from nuisance parameters, the observed value of V also indicates the precision of the inference to be drawn. The same test would emerge based on the goal of achieving a uniformly most powerful size α *similar* test.

9.1 Conditioning to Achieve UMP Size α Rejection Regions

In the most familiar class of cases, the aforementioned strategy for constructing appropriately sensitive tests, separate from nuisance parameters,

produces the same tests entailed by Neyman–Pearson theory, albeit with a difference in rationale. In particular, when certain requirements are satisfied rendering the statistic V a “complete” sufficient statistic for nuisance parameter λ , there is no other way of achieving the Neyman–Pearson goal of an exactly α -level rejection region that is fixed regardless of nuisance parameters – exactly *similar* tests.³ These requirements are satisfied in many familiar classes of significance tests. In all such cases, exactly similar size α rejection regions are equivalent to regions where the conditional probability of Y being significant at level α is independent of v :

$$\Pr(T(Y) \text{ is significant at level } \alpha \mid v; H_0) = \alpha,$$

where v is the value of the statistic V that is used to eliminate dependence on the nuisance parameter. Rejection regions where this condition holds are called regions of *Neyman structure*. Having reduced the null hypothesis to a simple hypothesis, one may then ensure the test has maximum power against alternatives to the null within the class of α -level tests. In the most familiar cases, therefore, conditioning on a sufficient statistic for a nuisance parameter may be regarded as an outgrowth of the aim of calculating the relevant p -value independent of unknowns, or, alternatively, as a by-product of seeking to obtain the most powerful similar tests.

However, requiring exactly similar rejection regions precludes tests that merely satisfy the weaker requirement of being able to calculate p approximately, with only minimal dependence on nuisance parameters; and yet these tests may be superior from the perspective of ensuring adequate sensitivity to departures, given the particular data and inference of relevance. This fact is especially relevant when optimal tests are absent. Some examples are considered in Section 10.

9.2 Some Limitations

The constructions sketched in the preceding sections reveal the underpinnings of a substantial part of what may be called elementary statistical methods, including standard problems about binomial, Poisson, and normal distributions, and the method of least squares for so-called linear models. When we go to more complicated situations, the factorizations that underlie the arguments no longer hold. In some generality, however, we may show that they hold approximately and we may use that fact to obtain p -values whose interpretation is only very mildly dependent on the values of nuisance parameters. The distributional calculations needed to

³ For a discussion on the technical notion of (bounded) completeness, see Lehmann (1986).

find p often involve appeal to so-called asymptotic theory or, perhaps more commonly nowadays, involve computer simulation. The goal of ensuring minimal dependence of the validity of primary inferences on unknown nuisance parameters is thereby achieved. A contrasting way to assess hypothesis H_0 in the face of several parameters is to assign (prior) probability distributions to each and integrate out the uninteresting parameters to arrive at the posterior for the null hypothesis given the data. Then, however, the resulting inference depends on introducing probability distributions for the unknown nuisance parameters, and the primary inference may be vitiated by faulty priors. The corresponding Bayesian treatment does not involve mathematical approximations, except where forced by the numerical complexity of some applications, but it does depend, often relatively critically, on a precise formulation of the prior distribution. Even in so-called impersonal Bayesian accounts, it can depend on the particular ordering of importance of the nuisance parameters. (We return to this in Section 12.)

10 Conditioning to Induce Relevance to the Particular Inference

Being able to calculate the p -value under the null, split off from nuisance parameters, although a necessary accomplishment, does not by itself entail that the calculation will be appropriately relevant for purposes of inference from the data. Although conditioning on sufficient statistics for nuisance parameters is also to tailor the inference to the best estimates of the background or nuisance parameters, more may be required in certain cases to ensure relevance to the given question of interest. We now turn to this issue.

Suppose then that one can calculate the p -value associated with an observed difference t_{obs} , namely $P(T \geq t_{\text{obs}}; \psi = \psi_0)$. If $P(T \geq t_{\text{obs}}; \psi = \psi_0)$ is very low (e.g., .001), then t_{obs} is grounds to reject H_0 or to infer a discordance with H_0 in the direction of the specified alternative at the corresponding level .001. There are two main rationales for this interpretation:

1. It is to follow a rule with low error rates (i.e., erroneous rejections) in the long run when H_0 is true. In particular, we may give any particular value p the following hypothetical interpretation. Suppose that we were to treat the data as just decisive evidence against H_0 ; then, in hypothetical repetitions, H_0 would be rejected in a long-run proportion p of the cases in which it is actually true.

However, this theoretical calibration of a significance test may be used as a measuring instrument to make inferences about how consistent or

inconsistent these data show this hypothesis to be in the *particular case* at hand. In such contexts the justification is that

2. It is to follow a rule where the low p -value corresponds to the *specific data set* providing evidence of inconsistency with or discrepancy from H_0 .

This evidential construal follows the frequentist principle FEV in 7(I). This aim is accomplished only to the extent that it can be assured that the small observed p -value is due to the actual data-generating process being discrepant from that described in H_0 . Moreover, the p -values (or corresponding confidence levels) associated with the inference should validly reflect the stringency and sensitivity of the actual test and the specific data observed.

Once these requirements in rationale 2 are satisfied, the low-error-rate rationale 1 follows, but the converse is not true. Many criticisms of frequentist significance tests (and related methods) are based on arguments that overlook the avenues open in frequentist theory for ensuring the relevancy of the sampling distribution on which p -values are to be based. It is one of the concerns addressed by the conditionality principle.

10.1 Weak Conditionality Principle (WCP)

Example 7. *Two measuring instruments of different precisions.* Suppose a single observation Y is made on a normally distributed random variable with unknown mean μ . A randomizing device chooses which of two instruments to use in measuring y : E' or E'' , with probabilities v' or v'' . The first instrument has known small variance, say 10^{-4} , whereas the second has known large variance, say 10^4 . The full data indicate whether E' or E'' was performed, and the value of Y , y' or y'' , respectively. The randomizer may be seen as an indicator of which experiment is performed to produce the data; for this purpose we typically consider an indicator statistic A , which takes values 1 and 2 with probabilities v' and v'' , respectively; $S = (Y, A)$ is sufficient. Statistic A , being a subset of S whose distribution is independent of the parameter of interest, is an example of an *ancillary statistic* (Cox, 1958).

Using this setup, one may define a *mixture test*. First let the device (e.g., a coin toss) choose the instrument to use, then report the result of using it and calculate the p -value. In testing a null hypothesis, say, $\mu = 0$, the same y measurement would correspond to a much smaller p -value were it to have

come from E' (Y is normal $N(\mu, 10^{-4})$) than if it had come from E'' (Y is normal $N(\mu, 10^4)$): denote them as $p'(y)$ and $p''(y)$, respectively. However, if one were to consider the overall type I error of the mixture corresponding to the observed y , one would average: $[p'(y) + p''(y)]/2$. This is the convex combination of the p -values averaged over the probabilities from E' and E'' , chosen by the randomizer. The p -value associated with an inference – if calculated using (the unconditional distribution of) the mixture of tests, abbreviated as $\text{Infr}_{E\text{-mix}}(y)$ – would be based on this average.

The point essentially is that the marginal distribution of a p -value averaged over the two possible configurations is misleading for a particular set of data. It would mean that an individual fortunate in obtaining the use of a precise instrument in effect sacrifices some of that information in order to rescue an investigator who has been unfortunate enough to have the randomizer choose a far less precise tool. From the perspective of interpreting the specific data that are actually available, this makes no sense. Once it is known whether E' or E'' has been run, the p -value assessment should be made conditional on the experiment actually run. In some other cases, the basis for conditioning may not be so obvious; therefore, there is a need for a systematic formulation.

Weak Conditionality Principle (WCP): If a mixture experiment (of the aforementioned type) is performed, then, if it is known which experiment produced the data, inferences about θ are *appropriately drawn in terms of the sampling behavior* in the experiment known to have been performed.

To avoid equivocation, it is important to understand what is being asserted. The WCP does not state a mathematical identity but it asserts that the *appropriate* way to draw the inference is not by means of the unconditional but rather by means of the conditional, sampling distribution of the experiment known to have produced the data. Once we know the data have been generated by E_j , given that our inference is about some aspect of E_j , our inference should not be influenced by whether a coin was tossed to decide which of two experiments to perform, and the result was to perform E_j . WCP is a *normative* epistemological claim about the appropriate manner of reaching an inference in the given context. We are assuming, of course, that all the stipulations in WCP are satisfied. Another example very often referred to in this context is the following.

Example 8. Suppose two independent and identically distributed random variables Y_1, Y_2 can each take values $\varphi - 1$ or $\varphi + 1$ with probability .5, φ unknown. The data take one of two possible configurations: either both

values are the same, say $y_1 = y_2 = y'$, or there are two different values, say $y_1 = y'' - 1$ and $y_2 = y'' + 1$. Let A be an indicator of which of the two configurations obtains in a given sample. The minimal sufficient statistic S is (Y, A) . In the case of the second configuration, $A = 2$, the sample values differ by 2 and, thus, ψ is exactly known to be y'' . In the first configuration ($A = 1$), the observed y -values are the same; thus, the two possible values for ψ , namely, $y' \pm 1$, are equally concordant with the data. Although the distribution of A is fixed independently of the parameter of interest, ψ , learning whether $A = 1$ or $A = 2$ is very relevant to the precision achieved; hence, the relevant inference would be conditional on its value.

As with Example 7, the sufficient statistic S being of dimension 2, while there is only one parameter, indicates the *incompleteness* of S . This opens the door to different p -values or confidence levels when calculated conditionally. In particular, the marginal distribution of a p -value averaged over the two possible configurations $(.5(0) + .5(.5) = .25)$ would be misleading for any particular set of data. Here the problem is generally given as estimating ψ with a confidence set with $n = 2$. If the two observed values are the same, then infer $\psi = y' - 1$; if they are different, infer ψ is y'' . Overall, the probability of an erroneous inference is .25. If two distinct values have been observed, all but one of the parameter values are ruled out with highest severity (p -value is zero), whereas when both observed values are the same, the test fails to discriminate between the two logically possible parameter values (the p -value for either value is .25).

The general argument here is analogous to the ones seen earlier. We seek a factorization $\mathbf{s} = (t, \mathbf{a})$, where t is one-dimensional, and we can write

$$f_{\mathbf{s}}(\mathbf{s}; \psi) = f_A(\mathbf{a}) f_{T|A}(t; \mathbf{a}, \psi),$$

where the first factor A has a fixed distribution that does not depend on θ . Now we argue that it is equivalent to obtain the data in two steps.

First, we observe that $A = \mathbf{a}$ (either $A = 2$, we are lucky enough to have observed two different values, or $A = 1$, we are unlucky enough to have observed both the same).

Second, we observe, conditionally on the first step, that $T = t$, an observation from the conditional distribution $f_{T|A}$ (e.g., given E'' is performed, observe y''). In the case of the mixture in Example 7, observing A corresponds to applying the randomizer, indicating which experiment to perform. In other kinds of examples, the second step might correspond, in an analogous manner, to conditioning on a statistic A that is indicative of the level of precision achieved. The second step defines a unique p . We may

call this process *technical conditioning to induce relevance* of the frequentist probability to the inference at hand.

Because it is given that the distribution of A does not involve ψ , merely learning the value of A at the first step tells us nothing directly about the value of ψ . However, it may, and indeed in general will, say something about the amount of information actually obtained; and thus is relevant to determining what is learned from the observed data as regards the actual data-generating procedure. If, for example, y' results from E' , then it is the properties of E' that are relevant for evaluating warranted inferences about E' .

The concern about average p -values in mixtures and related examples that underwrites the need to condition often arises in relation to the pre-data emphasis typically associated with the behavioristic “accept/reject” accounts of testing from which we have already distinguished the present approach. The justification for the WCP is fully within the frequentist sampling philosophy for contexts of scientific inference. There is no suggestion, for example, that only the particular data set should be considered. That would entail abandoning altogether the sampling distribution as the basis for inference. It is rather a matter of identifying an appropriate sampling distribution for the inference goal at hand.

It is not uncommon to see statistics texts argue that in frequentist theory one is faced with the following dilemma: either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or else to embrace the strong likelihood principle, which entails that frequentist sampling distributions are irrelevant to inference once the data are obtained. This is a false dilemma. Conditioning is warranted to achieve objective frequentist goals, and the conditionality principle coupled with sufficiency does not entail the strong likelihood principle. The “dilemma” argument is therefore an illusion (see Mayo, 7(III)).

11 Bayesian Alternatives

There are two further possible approaches to these issues. One involves a notion of probability as a personalistic degree of belief. It allows the incorporation of evidence other than that which can be modeled via a frequency concept of probability but, by its very nature, is not focused on the extraction and presentation of evidence of a public and objective kind. Indeed the founders of this approach emphasized its connection with individual decision making. Its main appeal in statistical work is some mixture of internal formal coherency with the apparent ability to incorporate information that is of a broader kind than that represented by a probabilistic

model based on frequencies. The essential focus is too far from our concern with objectivity for this to be a generally satisfactory basis for statistical analysis in science.

The other approach, based on a notion of rational degree of belief, has in some respects similar objectives to the frequentist view sketched earlier and often leads to formally very similar or even numerically identical answers. There are, however, substantial difficulties over the interpretation to be given to the probabilities used to specify an initial state of knowledge, and hence also to the final, or posterior, probabilities. We now turn to this issue.

11.1 What Do Reference Posteriors Measure?

Attempts to develop conventional “default,” “uninformative,” or “reference” priors are deliberately designed to prevent prior opinions and beliefs from influencing the posterior probabilities, thereby attaining an “objective” or impersonal Bayesian formulation. The goal is to retain the benefits of the Bayesian approach while avoiding the problems posed by introducing subjective opinions into scientific inference. A classic conundrum, however, is that no unique “noninformative” flat prior exists that would be appropriate for all inference problems within a given model. (To assume one exists leads to inconsistencies in calculating posterior marginal probabilities.) Any representation of ignorance or lack of information that succeeds for one parameterization will, under a different parameterization, appear to entail having knowledge; so that special properties of particular parameterizations have to be appealed to (Dawid, A.P., Stone, M.M., and Zidek, J.V., 1973).

Rather than seek uninformative priors, the majority of contemporary reference Bayesian research is directed to finding priors that are to be regarded as *conventions* for obtaining reference posteriors. The priors are not to be considered expressions of uncertainty, ignorance, or degree of belief. Conventional priors may not even be probabilities in that a constant or flat prior for a parameter may not sum to 1 (improper prior). They are conventions intended to allow the data to be “dominant” in some sense.

The most elaborately developed versions of this are the reference priors chosen to maximize the contribution of the data to the resulting inference (Bernardo, 2005). However, if priors are not probabilities, what then is the interpretation of a posterior? It may be stipulated, by definition, that the posteriors based on a reference prior *are* objective degrees of belief in the parameters of interest. More is required to show that the calculated posteriors succeed in measuring a warranted strength of evidence afforded

by data in the approximate truth or correctness of the various parameter values. Even if the reference prior research program succeeds in identifying priors that satisfy its own desiderata (Bernardo), it is necessary to show they satisfy this epistemic goal. Otherwise, it is not clear how to evaluate critically the adequacy of reference Bayesian computations for their intended epistemological measurement, in contrast to possibly regarding them as convenient mathematical procedures for deriving methods with good frequentist properties (Cox, 2006).

11.2 Problems with Nuisance Parameters

To ensure that unknown nuisance parameters exert minimal threats to the validity of p -value and other frequentist calculations, we saw how techniques for conditioning on sufficient statistics for such parameters are employed. By contrast, the Bayesian requires a joint distribution for all these unknowns, and the posterior will depend on how this is assigned. Not only may the calculation of a reference prior be relatively complicated but the prior for a particular parameter may depend on whether it is a parameter “of interest” or if it is a nuisance parameter, and even on the “order of importance” in which nuisance parameters are arranged. For example, if a problem has two nuisance parameters, the appropriate reference prior may differ according to which is considered the more important. The dependency on such apparently arbitrary choices tends to diminish the central goal of maximizing the contribution of the data to the resulting inference. The problem is not so much that different researchers can arrive at different posterior degrees with the same data; it is that such choices would appear to be inextricably bound up with the reported posteriors. As such, it would not be apparent which parts of the final inference were due to the data and which to the particular choice of ordering parameters.

11.3 Priors Depend on the Sampling Rule

Reference priors differ according to the sampling distribution associated with the model formulation. The result is to forfeit what is often considered a benefit of the Bayesian approach and to violate the strong likelihood principle (SLP), despite it often being regarded as the cornerstone of Bayesian coherency. Now the sampling distribution and the consideration of relevant hypothetical repetitions are at the heart of the frequentist objective assessment of reliability and precision, but violation of the SLP introduces incoherency into the reference Bayesian account. Reference Bayesians

increasingly look upon the violation of the SLP as the “price” that has to be paid for objectivity. We agree. Violating the SLP is necessary for controlling error probabilities, but this alone is not sufficient for objectivity in our sense.

Granted, as some (e.g., Berger, 2004) have noted in practice, arriving at subjective priors, especially in complex cases, also produces coherency violations.⁴ But there would seem to be an important difference between falling short of a formal principle (e.g., due to human limitations) and having its violation be required in principle to obtain the recommended priors.

11.4 Reference Posteriors with Good Frequentist Properties

Reference priors yield inferences with some good frequentist properties, at least in one-dimensional problems – a feature usually called *matching*. Although welcome, it falls short of showing their success as objective methods. First, as is generally true in science, the fact that a theory can be made to match known successes does not redound as strongly to that theory as did the successes that emanated from first principles or basic foundations. This must be especially so where achieving the matches seems to impose swallowing violations of its initial basic theories or principles.

Even if there are some cases where good frequentist solutions are more neatly generated through Bayesian machinery, it would show only their technical value for goals that differ fundamentally from their own. But producing identical numbers could only be taken as performing the tasks of frequentist inference by reinterpreting them to mean confidence levels and significance levels, not posteriors.

What some Bayesians seem to have in mind when pointing, as evidence of the success of reference priors, is that in some cases it is possible to match reference posteriors, construed as degrees of rational belief, with frequentist error probabilities. That is, the ability to match numbers helps to justify construing reference posteriors as objective degrees of belief in hypotheses. It is hard to know how to assess this, even in the very special cases where it

⁴ If the prior were intended to represent external knowledge, a Bayesian might justify using different priors in the cases described in Example 5 – binomial versus negative binomial – by considering that the latter is often used when the probability of success is small.

holds. Frequentist performance, we have shown, may, if correctly specified, be used to obtain measures of consistency with hypothesized parameter values and to assess sensitivity and precision of inferences about the system or phenomena at hand. It is not clear how the reference Bayesian's stated aim – objective degree of belief assignments – is attained through long-run error rates or coverage probabilities, even where these are achieved.

It is important not to confuse two kinds of “error probabilities”: Frequentist error probabilities relate to the sampling distribution, where we consider hypothetically different outcomes that could have occurred in investigating this one system of interest. The Bayesian allusion to frequentist “matching” refers to the fixed data and considers frequencies over different systems (that could be investigated by a model like the one at hand). Something would need to be said as to why it is relevant to consider other hypotheses, perhaps even in different fields, in reasoning about this particular H . The situation with frequentist priors considered as generating empirical Bayesian methods is distinct.

The ability to arrive at numbers that agree approximately and asymptotically with frequency-based measures in certain special cases does not seem sufficient grounds for the assurances often given that frequentist goals are being well achieved with reference priors, considering the cases of disagreement and the differences in interpretation in general. There is also the problem that distinct approaches housed under the impersonal Bayesian banner do not always agree with each other as to what method is to be recommended, even in fairly ordinary cases. Many seem to regard reference Bayesian theory to be a resting point until satisfactory subjective or informative priors are available. It is hard to see how this gives strong support to the reference prior research program.

12 Testing Model Assumptions

An important part of frequentist theory is its ability to check model assumptions. The use of statistics whose distribution does not depend on the model assumption to be checked lets the frequentist split off this task from the primary inference of interest.

Testing model adequacy formally within Bayesian formulations is not straightforward unless the model being used in initial analysis is itself embedded in some bigger family, typically putting fairly high prior probability on the initial model being close to the truth. That presupposes a reasonably clear notion of the possible departures that might arise.

13 Concluding Remarks

A statistical account, to be fairly assessed and usefully explicated, requires a clear understanding of both its aims and its methods. Frequentist methods, as here conceived, aim to learn about aspects of actual phenomena by testing the statistical consistency of hypotheses framed within statistical models of the phenomena. This approach has different tools, just as in science in general, for different questions and various stages of inquiries within its domain. Yet the frequentist standpoint supplies a unified argument and interconnected strategies that relate to achieving the goal of objective learning about the world. It achieves this goal by means of calibrations afforded by checkable standards. While there are certainly roles for other approaches, notably the use of personalistic Bayesian ideas in the context of personal decision making, we consider that a frequentist formulation provides a secure basis for the aims on which we are focusing: the analysis of data and the communication of conclusions in scientific research.

In our discussion of a frequency-based approach, the information in the data about the model is split into two parts: one captures all the information, assuming the model to be correct, and the other allows for checking the adequacy of the model. At least in simple problems this splitting is unique and unambiguous. The only other step in the development is the need in some, but not all, contexts to condition the probability calculations to achieve separation from nuisance parameters and to ensure their relevance to the issue under study. The appropriate choice for the statistic on which conditioning takes place depends on the particular aim of conditioning in the case at hand, taking account of the type of null hypothesis of interest. In cases with nuisance parameters, conditioning on a sufficient statistic V enables assessments of p -values and confidence levels that are free from threats to validity from these unknowns. In the most familiar class of cases, the strategies for constructing appropriately sensitive tests, separate from nuisance parameters, produces the same tests entailed by Neyman–Pearson theory, albeit with a difference in rationale.

In conditioning to induce relevance, the aim is to ensure inferences are constrained to reflect the actual precision of our test as an instrument for probing the underlying data-generation mechanisms. Tests that emerge to ensure relevance for the particular inference may differ from those developed based solely on Neyman–Pearson long-run behavior goals. The conditional perspective grows out of the desire that p -values (and confidence levels) reflect the relevant precision of particular inferences. It allows us to avoid well-known counterintuitive examples while remaining within the

frequentist, sampling theory framework. Understanding the sampling properties of statistical tools, as well as attention to the data-collection process, is the key to inductive inference from data to underlying values of θ . The sampling account deliberately leaves a clear trail regarding the basis for inferences drawn, providing a valuable framework for an objective scrutiny and improvement of results.

There is a historical and philosophical basis for a different notion of “objectivity” – one that is satisfied by automatic, conventional, “a priori” measures. In the contemporary “impersonal Bayesian” accounts, much as in earlier conventional approaches, the stumbling block remains one of showing appropriateness for achieving empirical scientific goals. Although it is among the most promising accounts currently on offer, we have also seen that the impersonal Bayesian paradigm is at odds with two fundamental goals of much Bayesian discussion: incorporating background information via priors, and adhering to the SLP (and the associated freedom from having to consider sampling distributions and stopping rules). Taking stock of the implications for foundations of statistics is therefore especially needful.

References

- Berger, J. (2004), “The Case for Objective Bayesian Analysis,” *Bayesian Analysis*, 1: 1–17.
- Bernardo, J.M. (2005), “Reference Analysis,” *Handbook of Statistics*, vol. 35, Elsevier, Amsterdam.
- Birnbaum, A. (1962), “On the Foundations of Statistical Inference,” *Journal of the American Statistical Association*, 57: 269–306.
- Cox, D.R. (1958), “Some Problems Connected with Statistical Inference,” *Annals of Mathematical Statistics*, 29: 357–72.
- Cox, D.R. (1990), “Role of Models in Statistical Analysis,” *Statistical Science*, 5: 169–74.
- Cox, D.R. (2006), *Principles of Statistical Inference*, Cambridge University Press, Cambridge.
- Cox, D.R., and Brandwood, L. (1959), “On a Discriminatory Problem Connected with the Works of Plato,” *Journal of the Royal Statistical Society B*, 21: 195–200.
- Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, Chapman and Hall, London.
- Dawid, A.P., Stone, M. and Zidek, J.V. (1973), “Marginalization Paradoxes in Bayesian and Structural Inference,” (with discussion), *Journal of the Royal Statistical Society B*, 35: 189–233.
- Fisher, R.A. (1935), *Design of Experiments*, Oliver and Boyd, Edinburgh.
- Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press, Cambridge.
- Lehmann, E.L. (1986), *Testing Statistical Hypotheses*, 2nd ed., Wiley, New York.
- Mayo, D.G. and Cox, D.R. (2006), “Frequentist Statistics as a Theory of Inductive Inference,” in J. Rojo (ed.), *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph Series, Institute of Mathematical Statistics (IMS), 49: 77–97.