the French Revolution. Yet even these enfranchisements revealed telling fault lines in the revolutionary conception of citizenship; the National Assembly, for instance, insisted that the Jews must give up their particular identity and corporate privileges as Jews in order to become French citizens. Since discussions of free black and slave rights inevitably involved the slave colonies, and since France's colony of Saint Domingue (San Domingo) was home to the first successful slave revolt in history beginning in 1791, the interest in citizenship has also stimulated study of France's slave colonies. Because the French Revolution put both revolution and modernity on the agenda, it continually offers rich possibilities for historical and theoretical debate.

See also: Bourgeoisie/Middle Classes, History of; Citizenship, Historical Development of; Democracy; Democratic Theory; Enlightenment; Freedom/Liberty: Impact on the Social Sciences; Ideology: History of the Concept; Nationalism: General; Political Representation; Representation: History of the Problem; Rousseau, Jean-Jacques (1712–78); Social Science, the Idea of; Socialism: Historical Aspects; Tocqueville, Alexis de (1805–59)

Bibliography

Arendt H 1963 *On Revolution*. Viking Press, New York

Barnave A 1988 *Introduction à la Révolution française. De la Révolution et de la Constitution*. Presses Universitaires de Grenoble, Grenoble, France

Buonarotti P 1828 *Conspiration pour l'égalité dite de Babeuf suivie du procès auquel elle donna lieu, et des pièces justificatives*. A la Libraire Romantique, Brussels, 2 Vols.

Burke E 1790 *Reflections on the Revolution in France, and on the Proceedings in Certain Societies in London Relative to that Event*, 3rd edn. J. Dodsley, London

Furet F 1978 *Penser la Révolution française*. Gallimard, Paris

Furet F 1986 *Marx et la Révolution française*. Flammarion, Paris

Godineau D 1988 *Citoyennes tricoteuses: les femmes du peuple à Paris pendant la Révolution française*. Alinéa, Aix-en-Provence, France

Hunt L A 1984 *Politics, Culture, and Class in the French Revolution*. University of California Press, Berkeley, CA

James C L R 1938 *The Black Jacobins: Toussaint L'Ouverture and the San Domingo Revolution*. Secker & Warburg, London

Landes J B 1988 *Women and the Public Sphere in the Age of the French Revolution*. Cornell University Press, Ithaca, NY

Taine H 1885–88 *La Révolution*. Hachette, Paris, 3 Vols.

Tocqueville A de 1856 *L'ancien régime et la Révolution*, 2nd edn. Michel Lévy frères, Paris

Tocqueville A de 1977 Oeuvres complètes. In: Jardin A, Lesourd J-A (eds.) *Correspondance Tocqueville-Louis de Kergolay*. Gallimard, Paris, Vol. 13

L. Hunt

Frequentist Inference

1. *Data, Models, and Inference*

The starting point of a statistical analysis is a set of data, for example, of counts or measurements. One aim may be simply to study what these data have to tell us. If they consist of a set of real numbers we might want to see, for example, whether they are small or large; tightly concentrated or spread out, whether they are stable or tend to increase with time, etc. If they are points in the plane we can get an idea of the shape of this set of points, for example, whether they cluster about a line. The branch of statistics dealing with this kind of investigation used to be called descriptive statistics, but now goes by the name data analysis or, more precisely, exploratory data analysis (EDA), a term introduced by Tukey (see *Exploratory Data Analysis: Univariate Methods*. Instead we consider how we can quantify the conclusions or decisions drawn from an analysis. Frequentist inference requires that any quantifying measure be interpretable in terms of frequentist probability of events (see *Frequentist Interpretation of Probability*) . That is, we assume the data are random quantities produced by some probability distribution and that something is known about this distribution. For example, if we have a set of n measurements of some quantity, we may assume that these measurements are independently and identically distributed. We may stop there or may go further and make some assumptions about this common distribution.

Statistical analyses based on such assumptions were common in the nineteenth century, and isolated instances can be found in the eighteenth century; but the first general framework was proposed by R. A. Fisher (1922) (see *Fisher, Ronald A (1890–1962)*). He states:

> The object of statistical methods is the reduction of data. A quantity of data which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole.
>
> This object is accomplished by constructing a hypothetical infinite population of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all quantities under discussion.

In modern terminology we might paraphrase this proposal by saying that we construct a mathematical model, according to which the data are produced by a probability distribution assumed to belong to some specified parametric family of distributions.

What are the statistical methods based on these models supposed to achieve? Fisher called the aim 'inductive inference' or 'inductive reasoning,' and described it variously as 'learning by experience' and

J. Rojo (ed.), *Selected Works of E. L. Lehmann*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1412-4_91, © Springer Science+Business Media, LLC 2012

as 'drawing inferences from the particular to the general, from consequences to causes' or—using statistical language—'from the sample to the population.' 'The purpose of inductive reasoning based on empirical observations,' he wrote 'is to improve our understanding of the systems from which these observations are drawn.'

This interpretation was criticized by Neyman (1961) who objected that 'After a conscientious effort to find the exact meaning of this term [inductive reasoning] I came to the conclusion that, at least in the sense of Fisher, the term is empty.' Neyman proposed instead that the purpose of statistical analysis is to serve as guide to appropriate action; and, in contrast to Fisher's inductive reasoning, named it inductive behavior (see also *Neyman, Jerzy (1894–1981)*).

Neyman's point of view was taken up by Wald who used it to construct a comprehensive framework for statistical decision making (Wald 1950). It involved three principal elements

(a) A family of probability distributions representing the various possible true situations.

(b) A set of actions from which it is the statistician's task to choose the most appropriate one.

(c) A loss function which measures the loss resulting from any action taken for any of the possible true situations.

Wald's formulation was castigated by Fisher as completely inappropriate for scientific work. 'It is important,' he wrote (Fisher 1973, p. 106) 'that the scientific worker introduces no cost functions for faulty decisions. To do so would imply that the purposes to which new knowledge was to be put were known and capable of evaluation. As workers in Science we aim at methods of inference which shall be equally convincing to all freely reasoning minds, entirely independently of any intentions that might be furthered by utilizing the knowledge inferred.'

While it is difficult not to sympathize with Fisher's stress on scientific aims as an alternative to immediate utilitarian use, his statements are distressingly vague. An effort to be more specific, which still seems rather vague, was made by Tukey (1960). For the outcome of a nonaction-oriented statistical analysis he proposed the term 'conclusion.' Concerning the difference he writes: 'Conclusions are established with careful regard to evidence but without regard to consequences of specific actions in specific circumstances. Conclusions are withheld until adequate evidence has been accumulated.'

The distinction between Fisher's point of view and that of Neyman and Wald, between conclusions and decisions, concerns the interpretation of the results of a statistical analysis. How those results are reached from the given model by means of frequentist inference f orms the subject of this article. (For the frequentist meaning of probability see *Frequentist Interpretation of Probability*.) It includes the derivation of appropriate procedures, the comparison of different procedures, investigating the sensitivity of a procedure to departures from the assumed model, and the checking of the suitability of the model. Another topic of interest, considered in Sect. 3.6, is the selection of an appropriate model.

There is an alternative approach to inference which can embody both the decision theory aspects of the Neyman–Wald approach and Fisher's views on learning by experience. This is the Bayes approach which requires not only the frequentist's specification of a family of possible distributions for the observation, but also that the distribution of the observation has itself been drawn from the specified model according to a known probability mechanism. This is coupled with the subjective point of view of probability in which the chance of an event represents a quantitative measure of the observers' odds on the occurrence or nonoccurrence of the event. This point of view, though attractive, puts a burden on the observer to specify his or her state of mind with possibly unreasonable precision and, of course, leaves open the issue of the generalizability of the conclusions by scientists at large (see also *Bayesian Statistics*).

The problems of inference have been the subject of much discourse among philosophers as well as scientists whose primary interest was not statistics. We refer to Keynes (1921), von Mises (1928) and Jeffreys (1939) among others.

A problem in discussing frequentist inference is that, while distinctions such as those between Fisher, Neyman, and Wald are required in order to formulate the goals of theory, they are artificial in practice. Even in scientific inquiry the cost of initiating extensive further investigation on the basis of weak evidence has to be kept in mind. On the other hand, the most interesting consequences may not be foreseen and the costs of actions difficult to quantify. Is foisting a slightly better-than-average drug with serious side effects on most patients better than doing nothing? What if the drug turns out to have major benefits for a small group of patients not represented in the original clinical trial of the drug? A chemist, E. B. Wilson (1952), after considering the work of the thinkers we mention, pleads eloquently, 'There is a great need for further work on the subject of scientific inference. To be fruitful it should be carried out by critical original minds who are not only well-versed in philosophy but also familiar with the way scientists actually work (and not just with the way some of them say they work).'

Wilson concludes pessimistically: 'Unfortunately the practical nonexistence of such people almost suggests that the qualities of mind required by a good philosopher and those needed by a working scientist are incompatible.'

In what follows we

(a) begin, for illustrative purposes, by describing a few simple models that have arisen from a frequentist point of view (Sect. 2), and

(b) discuss by example the principal frequentist inference procedures (hypothesis testing, point estimation, confidence regions, prediction, and model selection) (Sect. 3).

We use the term frequentist inference if all probabilities involved (such as significance level, power, and coefficient) are interpreted as frequentist probabilities in the sense of the article (*Frequentist Interpretation of Probability*). In particular, the assertion that the value of such a probability is p then means that in a large number of independent cases with probability p of some outcome, we expect the outcome to occur with a frequency close to p. In this connection it is important to realize that there is no need for these situations to be at all like each other, as long as the probability p is the same in each (see Neyman 1937).

2. *Some Simple Models*

Frequentist models are characterized by two principal features.

(a) The variability of the data. If the observations were repeated, different values would be obtained. This variability is represented in the model by postulating that the data are generated by a probability distribution.

(b) The unknown aspect of the situation which the statistical investigation is to elucidate. This unknown feature is represented in the model by the fact that the probability distribution of the observations is assumed to be only partially known.

These two features are seen clearly in the following model, which was mentioned briefly in the preceding section.

Example 1. (Error-measurement model) If $X_1, \ldots, X_n$ denote n measurements of an unknown quantity θ, this model can be written as

$$X_i = \theta + \varepsilon_i. \tag{1}$$

Here θ represents the unknown quantity we wish to estimate or test. The errors ε_i are the source of the variability of the observations. This frequency model assumes that, in a replication of the situation, the unknown value of θ would be the same but that the ε's would take on different values. The most common assumption concerning their variability is that they are independent random variables with a common distribution with mean zero. This distribution may be assumed to be known. More usually it too embodies some unknown features; for example, its form may be known but it may contain an unknown scale parameter.

Simpson and others toward the middle of the eighteenth century considered various possible forms for this distribution; but by the early nineteenth century it was agreed generally that the most suitable distribution of such measurement errors, in most cases, was the normal. The basis for this belief was the hypothesis of 'elementary errors,' which assumes that observational errors are the sum of a large number of small, elementary errors; and, by the Central Limit Theorem, are therefore approximately normally distributed.

If the ε's in (1) are assumed to be independently normally distributed with the mean 0 and variance σ^2, the model (1) is equivalent to assuming that

$$X_1, \ldots, X_n \text{ are independent } N(\theta, \sigma^2). \tag{2}$$

Example 2 (Normal one-sample model) Model (2) arises also in contexts quite different from the measurement situation described above. Suppose X is some numerical characteristic of a person such as height, weight, blood pressure, intelligence, etc., or of an animal, plant or manufactured product. However, this time we are not taking several observations X_i on the same subject but one observation each on different subjects drawn at random from a population which, in the model, we shall assume to be infinitely large. In this setting θ is the mean value of X in the population, and (2) is called the normal one-sample model. (The measurement situation can be viewed as a special case if one considers the n measurements actually taken as a sample from an essentially infinite number of measurements that could be taken.)

In these applications the assumption of normality is frequently made even though it is often not suitable and each situation ought to be considered in its own right. Often normality can be brought closer by applying it not to the observations X_i themselves but to some transformation $T(X_i)$. In other cases the normal model (2) is replaced by the more general model

$$X_1, \ldots, X_n \text{ are independently distributed according to } F(x-\theta) \tag{3}$$

where F has mean zero and otherwise is arbitrary, or where it is assumed to be an arbitrary distribution that is symmetric about zero. Some smoothness conditions may also be imposed; for example, that F has a density f.

Example 3. (Linear Model) Large areas of statistical methodology are based on the following extension of (1), due to Gauss (1809) and known as the linear model

$$X_i = \sum_{j=1}^{p} \beta_j z_{ij} + \varepsilon_i \quad (i = 1, \ldots, n) \tag{4}$$

where the z's are known constants (sometimes called explanatory variables) and the β's unknown parameters. The ε's are the observational errors as before and are assumed to have expectation zero.

Examples of (4) abound. Here are a few.

(a) The measurement model (1) is the special case of (4) with $p = 1$, $z_{i2} = 1$ and $\beta_1 = \theta$.

(b) The two-sample problem with $n_1 + n_2$ independent variables $X_1, \ldots, X_{n_1} : N(\theta_1, \sigma^2)$ and $X_{n_1+1}, \ldots, X_{n_1+n_2} : N(\theta_2, \sigma^2)$ corresponds to $p = 2$,

$$z_{11} = \cdots = z_{1n_1} = 1, \quad z_{1n_1+1} = \cdots = z_{1n_1+n_2} = 0, \quad \beta_1 = \theta_1 \tag{5}$$

and

$$z_{21} = \cdots = z_{2n_1} = 0, \quad z_{2n_1+1} = \cdots = z_{2n_1+n_2} = 1, \quad \beta_2 = \theta_2 \tag{6}$$

(c) The k-sample model is defined analogously. The two-way layout without interactions assumes that the variables X_{ijk} $(k = 1, \ldots, n_{ij}; i = 1, \ldots, a; j = 1, \ldots, b)$ are independent normal with common variance σ^2 and with means

$$E(X_{ijk}) = \mu + \alpha_i + \beta_j (\Sigma\alpha_i = \Sigma\beta_j = 0). \tag{7}$$

The k-sample model and the two-way layout are the simplest cases of analysis of variance models.

Another important class of examples of (4) are models for regression. The simplest case is

(d) Simple linear regression, given by

$$X_i = \alpha + \beta t_i + \varepsilon_i \tag{8}$$

where the t's are known constants, and α and β unknown parameters.

This is a special case of

(e) Polynomial regression, with

$$X_i = f(t_i) + \varepsilon_i \tag{9}$$

where f is a polynomial (for example, quadratic) of known degree and with unknown coefficients.

Other functions f, of course, are also possible.

A modification of (4) that is sometimes suitable is to have some of the z's random. For example (7) in some situations is replaced by

$$X_{ijk} = \mu + A_i + \beta_j + \varepsilon_{ijk} (\Sigma\beta_j = 0) \tag{10}$$

where the A's are unobservable random variables assumed to be independent $N(0, \sigma_A^2)$ and independent of ε. For more, see *Linear Hypothesis*.

When the distribution of the observational errors the ε's in (1) and (4) is specified, the models considered so far are parametric, that is, they can be smoothly parameterized by Euclidean labels. Models such as (3) with the distributional form of errors unspecified up to symmetry are semiparametric. Nonparametric models are ones in which as little is assumed as possible.

Example 4. (Nonparametric regression) This model is suitable in situations where, on each member of a sample of size n from some population, we observe not only a response Y but also observe characteristics (covariates) $\mathbf{Z} = (Z_1, \ldots, Z_k)$. For example, Y might be the income of a randomly sampled individual, Z_1 might be age, Z_2 educational level, etc. Of principal interest here is the average relation between Y and $\mathbf{Z}$

$$E(Y|Z_1 = z_1, \ldots, Z_k = z_k) \equiv m(z_1, \ldots, z_k). \tag{11}$$

If we assume nothing about the form of m, then subject to the assumptions that $E|Y| < \infty$ and that $(Y_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{Z}_n)$ are independent and identically distributed, this model is completely general since we can always write

$$Y_i = m(Z_{i1}, \ldots, Z_{ik}) + \varepsilon_i \tag{12}$$

where $E(\varepsilon_i | Z_i) = 0$. Note that if we assume

$$m(z_1, \ldots, z_k) = \sum_{j=1}^{k} \beta_j z_j,$$

then, given $\mathbf{Z}_i = z_i$ $(i = 1, \ldots, n)$ we are back in the linear model (4).

In order to identify m in (12) we need to make some assumptions, for instance that m is continuous.

An important method for constructing more complex models from simpler ones is that of hierarchical models. As exemplified in (10), such models often contain unobservable random variables which are themselves of interest.

From this point of view Bayes models can be considered as two-stage hierarchical models in which the first stage is a frequentist model involving unobservable parameters for which the Bayes model specifies a distribution as a second stage. If the frequentist model is itself hierarchical, a last stage is added when the model is Bayesianized. To this difference—that the unknown parameters of the frequentist model become unobservable random variables in the corresponding Bayes model of course, is added the difference in interpretation of the probabilities involved as frequentist or subjective.

3. *Inference Methods*

As indicated in Sect. 1 linked closely with the frequentist paradigm is that of decision theory. The principal types of problems considered in that framework are

(a) hypothesis testing,

(b) point estimation,

(c) confidence regions,

(d) prediction,
(e) model selection.
We shall discuss these in the above order.

3.1 Hypothesis Testing

Wald's decision theory grew out of the Neyman–Pearson theory of hypothesis testing (Neyman and Pearson 1933) which the authors formulated as choosing between two decisions, whether a prespecified statement (the hypothesis H) about the distribution P generating the data is or is not correct. If $\theta \in \Theta$ parameterizes the model, this is equivalent to determining whether $H: \theta \in \Theta_0$ or the alternative $K: \theta \in \Theta_1$ is true.

This formulation is satisfactory in many practical contexts such as occur, for example, in medicine, agriculture, or industry. For instance, in Example 3(b) the two samples may correspond to control and treatment groups of patients. In Example 3(b) the assumption of normality is made, and implicitly that of an additive treatment effect measured by the difference $\theta_2 - \theta_1$. The hypothesis to be tested is usually formulated as $H: \theta_2 \leqslant \theta_1$; i.e., that the treatment has no beneficial effect. The alternative $K: \theta_2 > \theta_1$ claims the existence of a beneficial effect.

The test for this problem is a test function δ which takes on the values $\delta(x) = 0$ or $\delta(x) = 1$ for the sample points x for which we decide to accept or reject H. The performance of such a test is measured by the probabilities of

(a) Type I errors—rejecting H when H is true, and
(b) Type II errors—accepting H when H is false,
that is, by the probabilities

$$P_\theta[\delta(x) = 1] \text{ when } \theta \in \Theta_0 \quad \text{and}$$
$$P_\theta[\delta(x) = 0] \text{ when } \theta \in \Theta_1. \tag{13}$$

The test δ is required to satisfy the condition

$$P_\theta(\text{Type I error}) \leqslant \alpha \quad \text{for all } \theta \in \Theta_0. \tag{14}$$

where the level α of the test is preassigned (usually as 0.05 or 0.01) as the maximum probability of false rejection to be tolerated. Subject to (14) we look for tests δ which make P_θ (Type II error) as small as possible.

This formulation is meaningful if it is possible to distinguish sharply between hypothesis and alternatives, between those θ-values for which rejection of H is clearly desirable or undesirable. In Example 3(b), for instance, we can specify that only beneficial effects exceeding $\varepsilon\sigma$ (for some given $\varepsilon > 0$) are of interest. Such a distinction is often possible in the contexts we have mentioned.

The Neyman–Pearson formulation of hypothesis testing in terms of the two types of error works well when we have clear definitions of both the hypothesis and the alternatives. In scientific contexts it is often the case that the hypothesis is clear but the alternatives are vague. It is then the unlikelihood of observations that would be surprising if the hypothesis were true (but less so if the vaguely entertained alternatives were valid) which takes on particular significance. It is still necessary to define a test statistic T and to suppose the hypothesis is rejected when its values are sufficiently large, since such values would be surprising if H were true but not if H were false. Then the probability of observing values greater than or equal to the observed value of T is the p-value or significance probability of the hypothesis. If it is sufficiently small, it draws attention to anomalies not expected under H. The calculation of such probabilities is central to science. However, whether such attention should be followed by action as assumed in the Neyman–Pearson theory depends on extraneous considerations.

The Bayesian approach to testing assigns prior probabilities to hypothesis and alternative and calculates the ratio of posterior probabilities, the Bayes factor, as a measure of evidence in favor of or against the hypothesis rather than the significant probability. This approach is very attractive in providing exactly what everyone wants: the probability of the hypothesis being true, given the data. However, this probability depends on the prior probability distribution assigned to the parameters; and these priors can overwhelm strong opposite data evidence. Disparate priors can result in conflicting conclusions based on the same data by analysts holding different prior opinions (see also *Hypothesis Testing in Statistics*).

3.2 Point Estimation

The decision-oriented point of view of Neyman and Pearson led to the full-fledged decision theory of Wald, based on assigning losses to the consequences of inappropriate decisions. Perhaps the most successful example of this approach is point estimation (for which the idea of loss functions had already been discussed by Gauss and Laplace).

The prototypical estimation problem is that of estimating an unknown physical constant subject to measurement error discussed in Example 1. The loss function most commonly considered is squared error leading to its expected value, the mean squared error (MSE), as its risk. The standard assumption of independent normal errors with mean zero and common variance implies that the sample mean $\bar{X}$ is the appropriate estimator according to a number of different criteria.

(a) Minimax. This is a worst case analysis in which the maximum risk is minimized.

(b) Uniform minimum variance unbiased. An estimator of θ is said to be unbiased if its expectation is equal to θ. For unbiased estimators the MSE is equal to the variance; and, in the present case, $\bar{X}$ minimizes the variance among all unbiased estimators.

(c) Maximum likelihood (ML). This is a non-decision-theoretic approach based on the likelihood the density of the data viewed as a function of the parameter. Fisher considered the likelihood of an observation as a measure of support for the different θ-values that might have generated it and proposed the θ-value maximizing it as an appropriate estimator of θ. When a uniform prior distribution can be placed on the parameter, the ML estimator is just the mode of the posterior distribution. However, when θ ranges over the infinite line (as is the case in the measurement model and many others), no uniform probability distribution for θ exists; and only an 'improper Bayes' interpretation is then possible.

The ML principle is popular for a number of reasons

(a) general applicability,

(b) invariance under reparameterization,

(c) in cases such as Example 2, where meaning can be attached to 'large sample sizes,' the method is approximately optimal,

(d) in many standard cases it leads to explicit solutions.

However, there are examples in which the resulting estimators are completely misleading, (LeCam 1990, Ferguson 1996). In particular, ML is either not applicable or can lead to very poor results in many non- and semiparametric models such as Example 4. For a general discussion of different approaches to estimation see, for example, Bickel and Doksum (2000), Lechmann and Casella (1998).

The assumption of joint normality of the data, for instance in Examples 1–3, is often seriously in doubt. In such cases robustness criteria are invoked and lead to alternatives to the linear procedures appropriate for normally distributed data. For example, the median rather than the mean (or some intermediate trimmed mean) might be used in Example 1, and minimum L_1 (least absolute values) rather than L_2 (least squares) estimators in Example 3. For an extended discussion of robustness see Hampel et al. (1986) and Staudte and Sheather (1990).

3.3 *Bayes Estimation*

Still another principle is that of Bayes estimation, i.e., minimizing the expected risk under a prior distribution π of the parameter $\boldsymbol{\theta}$. Since this is the same as minimizing the average risk, averaged with respect to the weight-function π, such estimators are of interest both from a Bayesian and a frequentist point of view. For estimating a real-valued parameter $g(\boldsymbol{\theta})$ with data X and squared error loss, the Bayes estimator is

$$E[g(\boldsymbol{\theta})|X] \tag{15}$$

the mean of the posterior distribution of $g(\boldsymbol{\theta})$ given X.

Bayes estimators are essentially never unbiased (Girshick and Savage 1951) since they naturally 'pull' the estimator toward one's prior opinion, the prior mean of $g(\boldsymbol{\theta})$. For example, in the one-sample model of Example 2. If π is normal (μ, τ^2), the Bayes estimator is

$$\left(\frac{n\bar{X}}{\sigma^2}+\frac{\mu}{\tau^2}\right) \Big/ \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right). \tag{16}$$

Frequentist and Bayes inference generally merge for large sample sizes since the data then wash out the influence of the prior distribution (provided the prior does not exclude parts of the sample space); see, for example, Blackwell and Dubins (1962).

In estimation even more can be said. If the prior distribution is locally uniform, Bayes estimators are asymptotically optimal in a frequentist sense by the Bernstein–von Mises theorem (see, for example, Lehmann and Casella 1998, Bickel and Doksum 2000). This is, however, a limit result; in practice, the effect of a strong prior often persists even for substantial sample sizes.

Both aspects can be illustrated by the Bayes estimator (16). For fixed μ, the difference of this estimator from $\bar{X}$ is of the order $1/n\tau^2$. For fixed τ^2, it is accordingly of the order $1/n$, while the difference of both estimators from θ is of the order $1/\sqrt{n}$. On the other hand, the convergence of (16) can become arbitrarily slow if $\tau^2 \to 0$ as $n \to \infty$. The basic message is that optimal frequentist estimators are appropriate for Bayesian problems and vice versa.

For the sake of simplicity, we have here focused attention to the estimation of real-valued parameters. Of course, estimation problems arise also for multivariate parameters (for instance, in Example 3) and for function-valued parameters such as the function m in Example 4 or in the estimation of an unknown density. Squared error loss is then replaced by summed or integrated MSEs. However, the univariate and multivariate situations present unexpected differences. For example, the typically unique minimax estimator of the univariate case is replaced by an infinite set of minimax estimators; and among these the natural extension of the classical univariate estimator is no longer optimal (Stein 1956).

Point estimates by themselves are hardly ever sufficient. Typically one also requires an idea of the error committed in using the estimator; that is, an estimate of the MSE. This is again the problem of estimating a parameter, and there is little to add. However, interpretation of the estimated measure of error is often difficult. It is clarified by the concept of confidence estimation, which we take up next.

3.4 *Confidence Regions*

When estimating θ in the measurement error model of Example 1, it is customary to indicate the reliability of

the estimate $\bar{X}$ by attaching to it the error estimate $\pm S/\sqrt{n}$ with $S^2 = \Sigma(X_i - \bar{X})^2/(n-1)$ as a proxy for $\pm\sigma/\sqrt{n}$. In the case of normal errors with mean zero.

$$P_\theta\left[\bar{X} - \frac{\sigma}{\sqrt{n}} \leqslant \theta \leqslant \bar{X} + \frac{\sigma}{\sqrt{n}}\right] = 0.67 \quad \text{for all } \theta. \tag{17}$$

If σ is known, (17) provides intervals of fixed length (and with random midpoints) which, in many repetitions of the experiment of taking n measurements, will 'cover' the true θ 67 percent of the time whatever are the true values of θ and σ. The intervals (17) are confidence intervals for θ with confidence coefficient 0.67.

In practice, σ is usually unknown. We can then obtain confidence intervals for θ by replacing σ by its estimator S. If t_α is the $100(1-\alpha/2)$ percent point of the t-distribution with $n-1$ degrees of freedom, the intervals

$$\bar{X} - t_\alpha \frac{S}{\sqrt{n}} \leqslant \theta \leqslant \bar{X} + t_\alpha \frac{S}{\sqrt{n}} \tag{18}$$

cover the true θ with probability $1-\alpha$ so that the statement (18) is correct about $100(1-\alpha)$ percent of the time. The length of the intervals is no longer fixed but random, its length tending to increase as the accuracy of the observations (the inverse of which is measured by σ and estimated by S) decreases.

In general, in estimating a parameter θ taking values in an arbitrary space Θ, a $100\ (1-\alpha)$ percent confidence region for θ is a random subset $C(X)$ of Θ, depending on the data with the property that

$$P[\theta \in C(X)] \geqslant 1-\alpha \tag{19}$$

for all probability distribution P in the model. Here the region $C(X)$ can, for example, be an interval as above, an ellipse as in the Scheffé regions for the parameters of the linear model in Example 3, or a 'band' about an unknown distribution function.

Confidence regions can be viewed as simultaneously making statements about the acceptability of a family of hypotheses (see Lehmann 1986). In particular, they provide a measure of the acceptability of the alternatives. A 95 percent confidence interval for the treatment effect of Example 3(b), for instance, which includes zero, tells us not only that the null hypothesis is accepted at the 5 percent level but also specifies all alternatives to the null hypothesis which are accepted at that level. An interval which does not contain zero but to which zero is close suggests that although we have seen something that is surprising under the null hypothesis, the data are consistent with alternatives that do not differ materially from zero.

The interpretation of confidence regions is conceptually difficult because the probability statements refer not to any random variation of the parameter, which, although unknown, is considered fixed, but to that of the data-dependent region. In contrast, a Bayesian analysis views θ as random and constructs credible regions $C^*(X)$ such that the posterior probability of θ falling into $C^*(X)$ is $\geqslant 1-\alpha$. This is just the kind of conclusion one would like to make; but, as in the case of testing, tends to be strongly influenced by assumptions concerning the prior.

Again, as in point estimation, Bayes regions derived on the basis of optimality criteria of size (length, volume, etc.) for large sample sizes agree approximately with frequentist confidence regions satisfying minimax size properties subject to fixed probability of coverage (see also *Estimation: Point and Interval*).

3.5 *Prediction*

This is the part of frequentist inference that is closest to the Bayesian approach and that fits least well into the decision-theoretic framework. Typically we are given a sample from a population of the form $(\mathbf{Z}_i, Y_i)$, $i = 1, \ldots, n$ where $\mathbf{Z}_i$ is a vector of predictors. Using this sample (which provides information concerning the relationship of $\mathbf{Z}$ and Y) and a new $\mathbf{Z}_{n+1}$ we wish to predict the unobserved Y_{n+1}. The predictor of Y_{n+1} is a function

$$\delta(\mathbf{Z}; \mathbf{Z}_1, Y_1, \ldots, \mathbf{Z}_n, Y_n) = \hat{\delta}(\mathbf{Z})$$

such that $\hat{\delta}(\mathbf{Z}_{n+1})$ is used to predict Y_{n+1}. In the continuous case (regression), the classical measure of loss resulting from incorrect prediction is squared error,

$$[Y_{n+1} - \hat{\delta}(Z_{n+1})]^2.$$

On the other hand, in the case of categorical variables taking on values $\{1, \ldots, k\}$ (classification), the standard loss function takes on the value 0 if $Y_{n+1} = \hat{\delta}(\mathbf{Z}_{n+1})$ and 1 otherwise. Prediction has the attractive feature that it allows nonparametric estimates of error since for any procedure we can compute its performance by applying it to the training sample on which it was built, for instance, $(1/n)\Sigma_{i=1}^n(\hat{\delta}(Z_i) - Y_i)^2$. This measure typically underestimates error, but techniques such as cross validation can be used to adjust it.

3.6 *Model Selection*

An important aspect of a frequentist analysis is the specification of a model. As an illustration, consider Example 3(e) where the regression function is assumed to be well approximated by a polynomial. Suppose the errors are normally distributed with zero mean and common variance. To complete the specification of the model, it is then necessary to decide on the degree d of the polynomial. With classical (least squares) prediction methods, the variance of the prediction error increases and its bias decreases with d. On a frequentist