1
Preliminaries

Summary. Key ideas about probability models and the objectives of statistical analysis are introduced. The differences between frequentist and Bayesian analyses are illustrated in a very special case. Some slightly more complicated models are introduced as reference points for the following discussion.

1.1 Starting point

We typically start with a subject-matter question. Data are or become available to address this question. After preliminary screening, checks of data quality and simple tabulations and graphs, more formal analysis starts with a provisional model. The data are typically split in two parts $(y : z)$, where y is regarded as the observed value of a vector random variable Y and z is treated as fixed. Sometimes the components of y are direct measurements of relevant properties on study individuals and sometimes they are themselves the outcome of some preliminary analysis, such as means, measures of variability, regression coefficients and so on. The set of variables z typically specifies aspects of the system under study that are best treated as purely explanatory and whose observed values are not usefully represented by random variables. That is, we are interested solely in the distribution of outcome or response variables conditionally on the variables z; a particular example is where z represents treatments in a randomized experiment.

We use throughout the notation that observable random variables are represented by capital letters and observations by the corresponding lower case letters.

A model, or strictly a family of models, specifies the density of Y to be

$$f_Y(y : z; \theta), \tag{1.1}$$

where $\theta \subset \Omega_\theta$ is unknown. The distribution may depend also on design features of the study that generated the data. We typically simplify the notation to $f_Y(y;\theta)$, although the explanatory variables z are frequently essential in specific applications.

To choose the model appropriately is crucial to fruitful application.

We follow the very convenient, although deplorable, practice of using the term *density* both for continuous random variables and for the probability function of discrete random variables. The deplorability comes from the functions being dimensionally different, probabilities per unit of measurement in continuous problems and pure numbers in discrete problems. In line with this convention in what follows integrals are to be interpreted as sums where necessary. Thus we write

$$E(Y) = E(Y;\theta) = \int y f_Y(y;\theta) dy \tag{1.2}$$

for the expectation of Y, showing the dependence on θ only when relevant. The integral is interpreted as a sum over the points of support in a purely discrete case. Next, for each aspect of the research question we partition θ as (ψ, λ), where ψ is called the *parameter of interest* and λ is included to complete the specification and commonly called a *nuisance parameter*. Usually, but not necessarily, ψ and λ are *variation independent* in that Ω_θ is the Cartesian product $\Omega_\psi \times \Omega_\lambda$. That is, any value of ψ may occur in connection with any value of λ. The choice of ψ is a subject-matter question. In many applications it is best to arrange that ψ is a scalar parameter, i.e., to break the research question of interest into simple components corresponding to strongly focused and incisive research questions, but this is not necessary for the theoretical discussion.

It is often helpful to distinguish between the primary features of a model and the secondary features. If the former are changed the research questions of interest have either been changed or at least formulated in an importantly different way, whereas if the secondary features are changed the research questions are essentially unaltered. This does not mean that the secondary features are unimportant but rather that their influence is typically on the method of estimation to be used and on the assessment of precision, whereas misformulation of the primary features leads to the wrong question being addressed.

We concentrate on problems where Ω_θ is a subset of R^d, i.e., d-dimensional real space. These are so-called *fully parametric* problems. Other possibilities are to have semiparametric problems or fully nonparametric problems. These typically involve fewer assumptions of structure and distributional form but usually contain strong assumptions about independencies. To an appreciable

extent the formal theory of semiparametric models aims to parallel that of parametric models.

The probability model and the choice of ψ serve to translate a subject-matter question into a mathematical and statistical one and clearly the faithfulness of the translation is crucial. To check on the appropriateness of a new type of model to represent a data-generating process it is sometimes helpful to consider how the model could be used to generate synthetic data. This is especially the case for stochastic process models. Understanding of new or unfamiliar models can be obtained both by mathematical analysis and by simulation, exploiting the power of modern computational techniques to assess the kind of data generated by a specific kind of model.

1.2 Role of formal theory of inference

The formal theory of inference initially takes the family of models as given and the objective as being to answer questions about the model in the light of the data. Choice of the family of models is, as already remarked, obviously crucial but outside the scope of the present discussion. More than one choice may be needed to answer different questions.

A second and complementary phase of the theory concerns what is sometimes called *model criticism*, addressing whether the data suggest minor or major modification of the model or in extreme cases whether the whole focus of the analysis should be changed. While model criticism is often done rather informally in practice, it is important for any formal theory of inference that it embraces the issues involved in such checking.

1.3 Some simple models

General notation is often not best suited to special cases and so we use more conventional notation where appropriate.

Example 1.1. *The normal mean.* Whenever it is required to illustrate some point in simplest form it is almost inevitable to return to the most hackneyed of examples, which is therefore given first. Suppose that $Y_1, \ldots, Y_n$ are independently normally distributed with unknown mean μ and known variance σ_0^2. Here μ plays the role of the unknown parameter θ in the general formulation. In one of many possible generalizations, the variance σ^2 also is unknown. The parameter vector is then (μ, σ^2). The component of interest ψ would often be μ

but could be, for example, σ^2 or μ/σ, depending on the focus of subject-matter interest.

Example 1.2. *Linear regression.* Here the data are n pairs $(y_1, z_1), \ldots, (y_n, z_n)$ and the model is that $Y_1, \ldots, Y_n$ are independently normally distributed with variance σ^2 and with

$$E(Y_k) = \alpha + \beta z_k. \tag{1.3}$$

Here typically, but not necessarily, the parameter of interest is $\psi = \beta$ and the nuisance parameter is $\lambda = (\alpha, \sigma^2)$. Other possible parameters of interest include the intercept at $z = 0$, namely α, and $-\alpha/\beta$, the intercept of the regression line on the z-axis.

Example 1.3. *Linear regression in semiparametric form.* In Example 1.2 replace the assumption of normality by an assumption that the Y_k are uncorrelated with constant variance. This is semiparametric in that the systematic part of the variation, the linear dependence on z_k, is specified parametrically and the random part is specified only via its covariance matrix, leaving the functional form of its distribution open. A complementary form would leave the systematic part of the variation a largely arbitrary function and specify the distribution of error parametrically, possibly of the same normal form as in Example 1.2. This would lead to a discussion of smoothing techniques.

Example 1.4. *Linear model.* We have an $n \times 1$ vector Y and an $n \times q$ matrix z of fixed constants such that

$$E(Y) = z\beta, \quad \text{cov}(Y) = \sigma^2 I, \tag{1.4}$$

where β is a $q \times 1$ vector of unknown parameters, I is the $n \times n$ identity matrix and with, in the analogue of Example 1.2, the components independently normally distributed. Here z is, in initial discussion at least, assumed of full rank $q < n$. A relatively simple but important generalization has $\text{cov}(Y) = \sigma^2 V$, where V is a given positive definite matrix. There is a corresponding semiparametric version generalizing Example 1.3.

Both Examples 1.1 and 1.2 are special cases, in the former the matrix z consisting of a column of 1s.

Example 1.5. *Normal-theory nonlinear regression.* Of the many generalizations of Examples 1.2 and 1.4, one important possibility is that the dependence on the parameters specifying the systematic part of the structure is nonlinear. For example, instead of the linear regression of Example 1.2 we might wish to consider

$$E(Y_k) = \alpha + \beta \exp(\gamma z_k), \tag{1.5}$$

where from the viewpoint of statistical theory the important nonlinearity is not in the dependence on the variable z but rather that on the parameter γ.

More generally the equation $E(Y) = z\beta$ in (1.4) may be replaced by

$$E(Y) = \mu(\beta), \tag{1.6}$$

where the $n \times 1$ vector $\mu(\beta)$ is in general a nonlinear function of the unknown parameter β and also of the explanatory variables.

Example 1.6. *Exponential distribution.* Here the data are $(y_1, \ldots, y_n)$ and the model takes $Y_1, \ldots, Y_n$ to be independently exponentially distributed with density $\rho e^{-\rho y}$, for $y > 0$, where $\rho > 0$ is an unknown rate parameter. Note that possible parameters of interest are ρ, $\log \rho$ and $1/\rho$ and the issue will arise of possible invariance or equivariance of the inference under reparameterization, i.e., shifts from, say, ρ to $1/\rho$. The observations might be intervals between successive points in a Poisson process of rate ρ. The interpretation of $1/\rho$ is then as a mean interval between successive points in the Poisson process. The use of $\log \rho$ would be natural were ρ to be decomposed into a product of effects of different explanatory variables and in particular if the ratio of two rates were of interest.

Example 1.7. *Comparison of binomial probabilities.* Suppose that the data are (r_0, n_0) and (r_1, n_1), where r_k denotes the number of successes in n_k binary trials under condition k. The simplest model is that the trials are mutually independent with probabilities of success π_0 and π_1. Then the random variables R_0 and R_1 have independent binomial distributions. We want to compare the probabilities and for this may take various forms for the parameter of interest, for example

$$\psi = \log\{\pi_1/(1-\pi_1)\} - \log\{\pi_0/(1-\pi_0)\}, \quad \text{or} \quad \psi = \pi_1 - \pi_0, \tag{1.7}$$

and so on. For many purposes it is immaterial how we define the complementary parameter λ. Interest in the nonlinear function $\log\{\pi/(1-\pi)\}$ of a probability π stems partly from the interpretation as a log odds, partly because it maps the parameter space $(0, 1)$ onto the real line and partly from the simplicity of some resulting mathematical models of more complicated dependences, for example on a number of explanatory variables.

Example 1.8. *Location and related problems.* A different generalization of Example 1.1 is to suppose that $Y_1, \ldots, Y_n$ are independently distributed all with the density $g(y - \mu)$, where $g(y)$ is a given probability density. We call μ

a *location parameter*; often it may by convention be taken to be the mean or median of the density.

A further generalization is to densities of the form $\tau^{-1}g\{(y-\mu)/\tau\}$, where τ is a positive parameter called a *scale parameter* and the family of distributions is called a location and scale family.

Central to the general discussion of such models is the notion of a family of transformations of the underlying random variable and the parameters. In the location and scale family if Y_k is transformed to aY_k+b, where $a>0$ and b are arbitrary, then the new random variable has a distribution of the original form with transformed parameter values

$$a\mu+b,\ a\tau. \tag{1.8}$$

The implication for most purposes is that any method of analysis should obey the same transformation properties. That is, if the limits of uncertainty for say μ, based on the original data, are centred on $\tilde{y}$, then the limits of uncertainty for the corresponding parameter after transformation are centred on $a\tilde{y}+b$.

Typically this represents, in particular, the notion that conclusions should not depend on the units of measurement. Of course, some care is needed with this idea. If the observations are temperatures, for some purposes arbitrary changes of scale and location, i.e., of the nominal zero of temperature, are allowable, whereas for others recognition of the absolute zero of temperature is essential. In the latter case only transformations from kelvins to some multiple of kelvins would be acceptable.

It is sometimes important to distinguish invariance that springs from some subject-matter convention, such as the choice of units of measurement from invariance arising out of some mathematical formalism.

The idea underlying the above example can be expressed in much more general form involving two groups of transformations, one on the sample space and one on the parameter space. Data recorded as directions of vectors on a circle or sphere provide one example. Another example is that some of the techniques of normal-theory multivariate analysis are invariant under arbitrary nonsingular linear transformations of the observed vector, whereas other methods, notably principal component analysis, are invariant only under orthogonal transformations.

The object of the study of a theory of statistical inference is to provide a set of ideas that deal systematically with the above relatively simple situations and, more importantly still, enable us to deal with new models that arise in new applications.

1.4 Formulation of objectives

We can, as already noted, formulate possible objectives in two parts as follows.

Part I takes the family of models as given and aims to:

- give intervals or in general sets of values within which ψ is in some sense likely to lie;
- assess the consistency of the data with a particular parameter value ψ_0;
- predict as yet unobserved random variables from the same random system that generated the data;
- use the data to choose one of a given set of decisions $\mathcal{D}$, requiring the specification of the consequences of various decisions.

Part II uses the data to examine the family of models via a process of model criticism. We return to this issue in Section 3.2.

We shall concentrate in this book largely but not entirely on the first two of the objectives in Part I, interval estimation and measuring consistency with specified values of ψ.

To an appreciable extent the theory of inference is concerned with generalizing to a wide class of models two approaches to these issues which will be outlined in the next section and with a critical assessment of these approaches.

1.5 Two broad approaches to statistical inference

1.5.1 General remarks

Consider the first objective above, that of providing intervals or sets of values likely in some sense to contain the parameter of interest, ψ.

There are two broad approaches, called *frequentist* and *Bayesian*, respectively, both with variants. Alternatively the former approach may be said to be based on *sampling theory* and an older term for the latter is that it uses *inverse probability*. Much of the rest of the book is concerned with the similarities and differences between these two approaches. As a prelude to the general development we show a very simple example of the arguments involved.

We take for illustration Example 1.1, which concerns a normal distribution with unknown mean μ and known variance. In the formulation probability is used to model variability as experienced in the phenomenon under study and its meaning is as a long-run frequency in repetitions, possibly, or indeed often, hypothetical, of that phenomenon.

What can reasonably be said about μ on the basis of observations $y_1, \ldots, y_n$ and the assumptions about the model?

1.5.2 Frequentist discussion

In the first approach we make no further probabilistic assumptions. In particular we treat μ as an unknown constant. Strong arguments can be produced for reducing the data to their mean $\bar{y} = \Sigma y_k/n$, which is the observed value of the corresponding random variable $\bar{Y}$. This random variable has under the assumptions of the model a normal distribution of mean μ and variance σ_0^2/n, so that in particular

$$P(\bar{Y} > \mu - k_c^*\sigma_0/\sqrt{n}) = 1 - c, \tag{1.9}$$

where, with $\Phi(.)$ denoting the standard normal integral, $\Phi(k_c^*) = 1 - c$. For example with $c = 0.025, k_c^* = 1.96$. For a sketch of the proof, see Note 1.5.

Thus the statement equivalent to (1.9) that

$$P(\mu < \bar{Y} + k_c^*\sigma_0/\sqrt{n}) = 1 - c, \tag{1.10}$$

can be interpreted as specifying a hypothetical long run of statements about μ a proportion $1 - c$ of which are correct. We have observed the value $\bar{y}$ of the random variable $\bar{Y}$ and the statement

$$\mu < \bar{y} + k_c^*\sigma_0/\sqrt{n} \tag{1.11}$$

is thus one of this long run of statements, a specified proportion of which are correct. In the most direct formulation of this μ is fixed and the statements vary and this distinguishes the statement from a probability distribution for μ. In fact a similar interpretation holds if the repetitions concern an arbitrary sequence of fixed values of the mean.

There are a large number of generalizations of this result, many underpinning standard elementary statistical techniques. For instance, if the variance σ^2 is unknown and estimated by $\Sigma(y_k - \bar{y})^2/(n - 1)$ in (1.9), then k_c^* is replaced by the corresponding point in the Student t distribution with $n - 1$ degrees of freedom.

There is no need to restrict the analysis to a single level c and provided concordant procedures are used at the different c a formal distribution is built up.

Arguments involving probability only via its (hypothetical) long-run frequency interpretation are called *frequentist*. That is, we define procedures for assessing evidence that are calibrated by how they would perform were they used repeatedly. In that sense they do not differ from other measuring instruments. We intend, of course, that this long-run behaviour is some assurance that with our particular data currently under analysis sound conclusions are drawn. This raises important issues of ensuring, as far as is feasible, the relevance of the long run to the specific instance.

1.5.3 Bayesian discussion

In the second approach to the problem we treat μ as having a probability distribution both with and without the data. This raises two questions: what is the meaning of probability in such a context, some extended or modified notion of probability usually being involved, and how do we obtain numerical values for the relevant probabilities? This is discussed further later, especially in Chapter 5. For the moment we assume some such notion of probability concerned with measuring uncertainty is available.

If indeed we can treat μ as the realized but unobserved value of a random variable M, all is in principle straightforward. By Bayes' theorem, i.e., by simple laws of probability,

$$f_{M|Y}(\mu \mid y) = f_{Y|M}(y \mid \mu) f_M(\mu) \Big/ \int f_{Y|M}(y \mid \phi) f_M(\phi) d\phi. \tag{1.12}$$

The left-hand side is called the *posterior density* of M and of the two terms in the numerator the first is determined by the model and the other, $f_M(\mu)$, forms the *prior distribution* summarizing information about M not arising from y. Any method of inference treating the unknown parameter as having a probability distribution is called *Bayesian* or, in an older terminology, an argument of *inverse probability*. The latter name arises from the inversion of the order of target and conditioning events as between the model and the posterior density.

The intuitive idea is that in such cases all relevant information about μ is then contained in the conditional distribution of the parameter given the data, that this is determined by the elementary formulae of probability theory and that remaining problems are solely computational.

In our example suppose that the prior for μ is normal with known mean m and variance v. Then the posterior density for μ is proportional to

$$\exp\{-\Sigma(y_k - \mu)^2/(2\sigma_0^2) - (\mu - m)^2/(2v)\} \tag{1.13}$$

considered as a function of μ. On completing the square as a function of μ, there results a normal distribution of mean and variance respectively

$$\frac{\bar{y}/(\sigma_0^2/n) + m/v}{1/(\sigma_0^2/n) + 1/v}, \tag{1.14}$$

$$\frac{1}{1/(\sigma_0^2/n) + 1/v}; \tag{1.15}$$

for more details of the argument, see Note 1.5. Thus an upper limit for μ satisfied with posterior probability $1 - c$ is

$$\frac{\bar{y}/(\sigma_0^2/n) + m/v}{1/(\sigma_0^2/n) + 1/v} + k_c^* \sqrt{\frac{1}{1/(\sigma_0^2/n) + 1/v}}. \tag{1.16}$$

If v is large compared with σ_0^2/n and m is not very different from $\bar{y}$ these limits agree closely with those obtained by the frequentist method. If there is a serious discrepancy between $\bar{y}$ and m this indicates either a flaw in the data or a misspecification of the prior distribution.

This broad parallel between the two types of analysis is in no way specific to the normal distribution.

1.6 Some further discussion

We now give some more detailed discussion especially of Example 1.4 and outline a number of special models that illustrate important issues.

The linear model of Example 1.4 and methods of analysis of it stemming from the method of least squares are of much direct importance and also are the base of many generalizations. The central results can be expressed in matrix form centring on the least squares estimating equations

$$z^T z\hat{\beta} = z^T Y, \tag{1.17}$$

the vector of fitted values

$$\hat{Y} = z\hat{\beta}, \tag{1.18}$$

and the residual sum of squares

$$\text{RSS} = (Y - \hat{Y})^T (Y - \hat{Y}) = Y^T Y - \hat{\beta}^T (z^T z)\hat{\beta}. \tag{1.19}$$

Insight into the form of these results is obtained by noting that were it not for random error the vector Y would lie in the space spanned by the columns of z, that $\hat{Y}$ is the orthogonal projection of Y onto that space, defined thus by

$$z^T(Y - \hat{Y}) = z^T(Y - z\hat{\beta}) = 0 \tag{1.20}$$

and that the residual sum of squares is the squared norm of the component of Y orthogonal to the columns of z. See Figure 1.1.

There is a fairly direct generalization of these results to the nonlinear regression model of Example 1.5. Here if there were no error the observations would lie on the surface defined by the vector $\mu(\beta)$ as β varies. Orthogonal projection involves finding the point $\mu(\hat{\beta})$ closest to Y in the least squares sense, i.e., minimizing the sum of squares of deviations $\{Y - \mu(\beta)\}^T\{Y - \mu(\beta)\}$. The resulting equations defining $\hat{\beta}$ are best expressed by defining

$$z^T(\beta) = \nabla\mu^T(\beta), \tag{1.21}$$

where ∇ is the $q \times 1$ gradient operator with respect to β, i.e., $\nabla^T = (\partial/\partial\beta_1, \ldots, \partial/\partial\beta_q)$. Thus $z(\beta)$ is an $n \times q$ matrix, reducing to the previous z