Box's diagnosis should be carried out in practice but will not be adopted here, because it would carry us too far afield from our main goal, which is to demonstrate the technical mechanics of making posterior inferences. In any case, one will benefit from reading the Box paper.

The Probability Model

How does one choose a probability model f(x|θ) to represent the data generating mechanism? In our context, how does one choose one of the many linear models which are described in the earlier parts of the chapter? This is a separate area of statistics and we will not deal with it here. It will be assumed a particular model is indeed appropriate and that it was chosen either from subject matter considerations or from diagnostic checks. In addition to the classical goodness-of-fit procedures to check model adequacy, there are Bayesian methods of choosing a model from a class of models. It is a direct application of Bayes theorem and given by Zellner (1971, page 292). It will be assumed here that one has correctly chosen a parametric family of densities $\{f(x|\theta); x \in S, \theta \in \Omega\}$ which correctly represents the sample data, and our problem is to find out more about the parameter θ. Or it may be the problem of choosing one particular density (or subset of densities, as in some hypothesis testing problems) from this class.

For example, suppose one is sure an AR(p) process is appropriate for some time series data $\{y(t): t = 1,2, \ldots, n\}$, then how should one choose p? This identification problem is analyzed in Chapter 5. Here we have assumed the AR(p) class of models is appropriate and one may describe the class by a parametric family of Gaussian densities. Our problem is to learn more about p (the order) as well as the other parameters (the autoregressive coefficients and noise variance) of the model and to forecast future observations.

Prior Information

Of all the various aspects of Bayesian inference, this is the most difficult and controversial. Many object to treating unobservable parameters θ as random variables and giving them subjective probability distributions or densities ξ(θ); however not many people object to giving observable quantities x a frequency-type probability distribution. This is probably so because one is so used to doing it this way one does not question this part of their model. To assume x is a random variable with a frequency distribution is to assume the existence of an unending sequence of repetitions of the experiment, and of course, since such repetitions in fact do not occur they must be imagined to occur; hence to assume x has a frequency type probability distribution is no more difficult (or is as difficult) than it is to assume θ has some (subjective) probability distribution. In any case to use the Bayesian approach, one must have a probability distribution for θ which expresses one's information about the parameters before the data are observed. Phrased another way, one can say to adopt a frequency probability distribution for x and a subjective probability law for θ are both quite subjective activities. However the probability law of x is more objective in the sense x can be observed and if observed its probability law can be checked against the data.

Choosing a prior density for the parameters has an interesting history. We have seen that Bayes chooses a prior density for the parameter θ of a Bernoulli sequence by constructing a "billiard" table and that ξ(θ) (uniform over (0,1)) had a frequency interpretation.

In those situations where θ cannot be given a frequency interpretation, how does one choose a prior for θ, if one knows nothing about θ? Stigler (1982) argues that Bayes would choose ξ(θ) to be uniform because the marginal distribution of x (the total number of successes in n trials) has a discrete uniform distribution over 0,1,2, …, n, and a continuous uniform prior density for θ implies x has a discrete uniform distribution. This justification for a uniform prior for θ is quite different than the usual interpretation of Bayes' choice of a uniform prior for θ. Many have used the principle of insufficient reason to put a uniform distribution on the parameters. The principle of insufficient reason according to Jeffreys (1961) is "If there is no reason to believe one hypothesis rather than another, the probabilities are equal" and one may choose a uniform prior density for the parameters, however, if one knows nothing about θ, nothing is known about any function of θ, thus any function of θ should have a uniform prior density (according to the principle of insufficient reason), hence one is led to a contradiction in using the principle. Jeffreys has formulated a theory of choosing prior densities based on rules of invariance in situations where "nothing" is known about the parameters. For example, if θ_1 and θ_2 are the mean and standard deviation of a normal population, then Jeffreys' improper prior is

$$\xi(\theta_1, \theta_2) \propto \frac{1}{\theta_2}, \quad \theta_1 \in R, \quad \theta_2 > 0,$$

which implies θ_1 and θ_2 are independent, that θ_1 has a constant marginal density over R and that the marginal density of is θ_2 is $\xi_2(\theta_2) \propto 1/\theta_2$, $\theta_2 > 0$. This means every value of θ_1 is equally likely and that smaller values of the standard deviation are more likely than larger values.

This type of prior was used in Chapter 1 and will be used many times in the following chapters. Often Jeffreys' prior leads to confidence intervals which are identical to those constructed from a non-Bayesian theory, but of course the interpretation of the two would be different. To some Bayesians this perhaps is an advantage of the Jeffreys' approach but to others a liability. Whatever the advantages or disadvantages, the Jeffreys improper prior has been used over and over again by Box and Tiao (1971), Zellner (1971), Lindley (1965) and many others, but Lindley now prefers to use proper prior densities instead of Jeffreys' improper prior distributions, because their use can lead to logical inconsistencies, which was demonstrated by Dawid et al. (1973). The issue is still being debated and Jaynes (1980) reportedly refutes the Dawid argument.

Conjugate densities are proper distributions which have the property that the prior and posterior densities both belong to the same parametric family, thus they have the desirable property of mathematical convenience. Raiffa and Schlaifer (1961) developed Bayesian inference and decision theory with conjugate prior densities and their work is a major contribution to statistical theory. DeGroot (1970) also gives an excellent account of conjugate prior distributions and shows how to use them with the usual (popular) population models such as normal, Bernoulli, and Poisson. Press (1982), Savage (1954), and Zellner (1971) also consider conjugate densities.

Considering Bayes' original paper about a Bernoulli population parameter θ, the conjugate class is the two-parameter beta family with density

$$\xi(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 \leqslant \theta \leqslant 1, \quad \alpha > 0, \quad \beta > 0$$

and one must choose values for α and β in order to express prior information. This class is quite flexible because the class contains a large number and variety of distributions, but this is not a general property of a conjugate class. The Wishart class is conjugate to the multinormal population with an unknown precision matrix and is not very flexible (as compared to the beta family). See Press (1982) for a discussion of the Wishart and related distributions, but the main problem with employing a conjugate class is that one must choose the parameters (hyperparameters) of the prior distribution. With the beta one must choose two scalar hyperparameters and with the p-dimensional normal distribution, one must find p + p(p + 1)/2 hyperparameters.

With the beta one must choose two scalar hyperparameters and with the p-dimensional normal distribution, one must find $p + p(p + 1)/2$ hyperparameters.

Still another principle in choosing prior distributions is that of precise measurement, which "is the kind of measurement we have when the data are so incisive as to overwhelm the initial opinion, thus bringing a great variety of realistic initial opinions to practically the same conclusion" (Savage, 1954, page 70). In choosing a prior on the basis of this principle the normalized likelihood function is approximately the same as the posterior distribution, and the prior does not have much of an effect on the posterior distribution.

After the brief history of prior information, we are still left with the problem of how to choose a prior distribution for the parameters of the model, that is, how does one express what one actually knows about θ?

Apparently it is easier to model the observations than it is to model the parameters of the model of the observations. As statisticians we are familiar with choosing various probability models to different experimental or observational conditions, and it becomes somewhat routine in some environments. For example, regression models and models for designed experiments are clearly appropriate in many situations, but when it comes to adopting a model for the parameters of the probability model (such as for the regression coefficients in an autoregressive process) it is another story.

Of course, one should not always use a Jeffreys' improper prior density or a conjugate class, or the principle of precise measurement, but try to accurately represent the prior information.

Improper prior densities or a conjugate class of densities are used throughout this book to represent prior information, although this does not mean one should always use these two types of prior information. They are used because they combine nicely with the likelihood function.

Why is it so difficult to choose prior densities for the parameters? Possibly because we are so used to thinking in terms of observable random variables we have not developed a feeling or intuition about what the parameters represent in our probability models.

The approach the author favors is to choose the prior density in terms of the marginal distribution of the observations (predicted or past data). This approach is referred to as the predictive method or the device of imaginary results (Stigler, 1982, page 254) and is exploited by Good (1950, page 35), Geisser (1980), Kadane (1980), Winkler (1977), and Kadane et al. (1980). In a recent paper, Stigler (1982) argues that this "device of imaginary results" was what Bayes did in his justification for the choice of a uniform prior density for θ, where θ is the parameter of a binomial experiment and θ "must" be given a subjective probability distribution.

The device of imaginary results is thoroughly explained by Stigler (1982) and what follows is his description. Consider the usual parametric model $f(x|\theta)$, the conditional density of an observation x when θ is the value of the parameter and Ω is the parameter space, then the marginal density of x is

$$p(x) = \int_{\Omega} f(x|\theta)\xi(\theta)\, d\theta, \quad x \in S$$

where S is the sample space and ξ the prior density of θ, $\theta \in \Omega$. The device of imaginary results implies that one should choose $\xi(\theta)$ in such a way that it is compatible with one's choice of $p(x)$, $x \in s$. How does one choose $p(x)$? One can either "fit" p to imaginary predictions of future values of the observation or fit p to past data, which were collected under the same circumstances as the future observations will be observed. As Winkler (1980) notes, it is much easier for the experimentor to predict future values of x than it is to think directly in terms of θ.

In the binomial case considered by Bayes (1963) a uniform discrete distribution of x implied $\xi(\theta)$ was uniform on (0,1), that is

$$p(x) = \frac{1}{n+1}, \quad x = 0,1,2,\ldots,n$$

implies $\xi(\theta) = 1$, $0 < \theta < 1$, where

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0,1,2,\ldots,n.$$

Of course if $\xi(\theta)$ is uniform then $p(x)$ is uniform, but the converse is not necessarily true.

The prior predictive density $p(x)$ allows us to assess prior information about θ from prior knowledge about x instead of contemplating directly the prior density $\xi(\theta)$. If ξ depends on hyperparameters $\alpha \in A$, then the predictive density is

$$p(x|\alpha) = \int_{\Omega} f(x|\theta)\, \xi(\theta|\alpha)\, d\alpha, \quad x \in S, \quad \alpha \in A$$

and one may use this integral equation to find values of α which support or fit predicted observations $x_1, x_2, \ldots, x_n$ (sampled from the population with density $p(x|\alpha)$) or to fit past observations which were sampled from a population with density $p(x|\alpha)$.

Suppose in the binomial example

$$\xi(\theta|\alpha) \propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}, \quad 0 \leqslant \theta \leqslant 1, \quad \alpha_1 > 0, \quad \alpha_2 > 0$$

then the prior density belongs to the conjugate class, and the prior predictive density of x is

$$p(x|\alpha_1, \alpha_2) \propto \Gamma(\alpha_1 + x)\Gamma(n + \alpha_2 - x), \quad x = 0,1,2,\ldots,n.$$