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

If one observes values $x_1, x_2,\ldots, x_n$ from this distribution, one may find values of α_1 and α_2 compatible with these predictions by the method of moments or maximum likelihood or some other principle of estimation. One might put a known prior density on α_1 and α_2 and estimate α_1 and α_2 from the conditional distribution of α_1 and α_2 given the predicted observations.

Winkler (1980) shows how this method works with the general linear model of Chapters 1 and 3 and develops the conjugate prior density for the parameters of a multiple linear regression model

$$y = \sum_{i=1}^{k} \theta_i x_i + e,$$

where y is the dependent variable, θ_i is an unknown regression coefficient, x_i is the i-th independent variable, and e is $n(0,\sigma^{-2})$. If $\sigma^2$is known the conjugate class is normal with mean say b and covariance matrix σ^2 V, where b is k × 1 and V is k × k, then the predictive distribution of y is normal with mean

$$E(y|\sigma^2,x) = b'x,$$
$$Var(y|\sigma^2,x)=\sigma^2(x'Vx + 1),$$

where $x' = (x_1,x_2,\ldots, x_k)$, $b = (b_1,b_2,\ldots,b_k)'$. Winkler shows how one may assess this predictive distribution, thus choosing values for b and V, the hyperparameters of the prior distribution of $\theta =(\theta_1\ \theta_2,\ldots,\theta_k)'$ Given x, one predicts the mean and variance of y for k(k + l)/2 values of x, where k of the equations (assessments) allows one to find b, and all k(k + 1)/2 allows one to find V. Winkler also shows how to deal with the case when σ^2 is unknown and Kadane et al. (1980) develop a sophisticated but complicated method of fitting the hyperparameters of prior distributions.

The advantage of the predictive approach is that the experimenter is better able to think about an observable random variable than he is (directly) about an unobservable parameter θ. A disadvantage is that one must usually restrict prior information to a parametric family such as a conjugate class of distributions, and the class may not be as flexible as one desires. Perhaps this can be avoided by a mixture of densities from the conjugate class.

The principle of imaginary results appears to be a very promising way to assess prior information and it should be used in practice to see if experimenters and other users of statistical techniques are willing to adopt it.

One of the latest developments assessing prior information is the idea that the parameters of the model are exchangeable. The idea of exchangeability was introduced by deFinetti (1937) (see Kyburg and Smokier, 1964) and has been applied by Lindley and Smith (1972) and Leonard (1972) for the Bayesian analysis of linear models and binomial probability laws.

Let us consider an example given by Lindley and Smith (1972) where $y_1, y_2,\ldots, y_n$ are independent, normally distributed, have a known variance and $E(y_i) = \theta_i$, i = 1,2,…,n, where the means $\theta_1, \theta_2, \ldots, \theta_n$ are unknown real parameters and very little is known about the means. Since very little is known about the means it perhaps is reasonable to assume the means are exchangeable, that is, one's prior information about θ_1 is the same as that about θ_2, or any other θ_i and that in general one's prior knowledge about $\theta_1, \theta_2,\ldots,\theta_n$ is invariant under any permutation of the indices. An exchangeable sequence can arise as a mixture of independent identically distributed random variables, given some hyperparameter μ, say, that is the density of $\theta = (\theta_1,\theta_2,\ldots,\theta_n)'$ is

$$g(\theta) = \int \prod_{i=1}^{n} p(\theta_i|\mu)\, dQ(\mu),$$

where $p(\theta_i|\mu)$ is the conditional density of θ_i given μ, Q(μ) the distribution function of μ, and p and Q are arbitrary. Thus $\theta_1, \theta_2,\ldots,\theta_n$ (given μ) constitute a random sample of size n from a population with density p, and μ is a random variable with distribution function Q(μ). Thus, for example, we might let p be normal with mean μ and known variance τ^{-1} and give μ some known probability distribution, or we could have the prior distribution of μ depend on unknown hyper-parameters and put a known distribution on them and continue in a hierarchical fashion.

Lindley and Smith next consider a randomized block layout of b blocks in each of which t treatments are assigned, then the model is

$$y_{ij} = \mu+\alpha_i+\beta_j + e_{ij},$$

where y_{ij} is the observation on the experimental unit of block j to which treatment i was applied, the errors are independent $n(0,\sigma^{-2})$ and one must assign a prior distribution to μ, $\alpha_1,\alpha_2,\ldots,\alpha_t$, $\beta_1, \beta_2,\ldots, \beta_b$ If the treatment constants are exchangeable and also the block parameters, and they are independent, then one perhaps would assign $\alpha_i \sim n(0,\sigma_\alpha^{-2})$, $\beta_j \sim n(0,\sigma_\beta^{-2})$ and $\mu \sim n(0,\sigma_\mu^{-2})$, where these distributions are independent. If the variances σ_α^2, σ_β^2, and σ_μ^2 are known, one stops here at this stage and proceeds with the posterior analysis. Also, one might consider these variances as unknown and give them some known distribution, and this is also considered by Lindley and Smith.

The idea of exchangeability has played an important role in the Bayesian analysis of linear models because deFinetti's work provides a mathematical justification for selecting a prior distribution when a priori symmetry is present about the parameters. It imposes a reasonable restriction on the prior probability distribution when our knowledge of the parameters exhibits this peculiar symmetric property, but it usually doesn't tell us what particular parametric family is the appropriate one to use. For example, in the above two problems exchangeability of the model parameters does not imply a normal prior density for them, although these parameters are indeed exchangeable.

This section has reviewed some of the ways one may express prior information about the parameters. We have looked at vague prior information, conjugate prior densities, the principle of insufficient reason, the principle of imaginary results, and exchangeable prior information. Also, empirical methods of assessing prior information were discussed, thus it is easy to see that much has been developed in the area of prior information but that also much remains to be accomplished.

The Posterior Analysis

Suppose one is now satisfied that the probability model, $f(x|\theta)$, for observations and the prior density, $\xi(\theta)$, for the parameters are appropriate and one must determine the plausible values of θ. One now enters the posterior analysis stage of a Bayesian statistical analysis, where one perhaps might want to estimate θ (or some of its components), test hypotheses about θ, predict future values of the observations, or possibly control the observations at predetermined levels.

Regardless of what particular activity is contemplated, one must first find the posterior density of θ,

$$\xi(\theta|x) \propto f(x|\theta)\,\xi(\theta), \quad \theta \in \Omega \subseteq R^p$$

which is found by Bayes theorem. Often all the components of θ are of interest, but sometimes some of these components θ_1 are regarded as nuisance parameters and the remaining $\theta_2(q \times 1)$ are of primary interest. How should one make inferences about θ_2? The Bayesian will use

$$\xi_2(\theta_2|x) = \int_{\Omega_1} \xi(\theta_1,\theta_2|x)\, d\theta_1, \quad \theta_2 \in \Omega_2, \quad \Omega = \Omega_1 \times \Omega_2,$$

where θ_1 is $(p - q) \times 1$. ξ_2 is called the marginal posterior density of θ_2, and as with any posterior density function it may be used to estimate and test hypotheses concerning θ_2. To estimate θ_2, either the marginal mean, median, or mode, or any other typical or representative value of θ_2 can be computed from the marginal posterior density of θ_2. Of course these estimates are summary characteristics of $\xi(\theta_2|x)$ and only tell us part of the story about the parameter and they are not a substitute for the entire distribution.

Consider an experiment which has been conducted according to a randomized block design (see Chapter 3), then the primary parameters are the treatment effects but the block effects and error variance are regarded as nuisance parameters. Then the Bayesian would test for significance of treatment effects by referring to the marginal posterior distribution of those parameters.

To test hypotheses about the parameters θ_2, one perhaps would find a $1 - \gamma$ HPD, $0 \leqslant \gamma \leqslant 1$ (highest posterior density) region $R_\gamma(\theta_2)$ from $\xi(\theta|x)$. Such a region has the property that if $\theta'_2 \in R_\gamma(\theta_2)$ and $\theta''_2 \notin R_\gamma(\theta_2)$, then $\xi(\theta'_2|x) > \xi(\theta''_2|x)$, that is parameter values inside the region have larger posterior probability density than those excluded from the region which must satisfy

$$1 - \gamma = \int_{R_\gamma(\theta_2) \subseteq \Omega_2} \xi(\theta_2|x)\, d\theta_2,$$

that is the HPD region has posterior probability content $1 - \gamma$. Box and Tiao (1971) and Chapter 1 give other properties of HPD regions and they are used in Chapter 3. To test the hypothesis H_0: $\theta_2 = \theta_{20}$(known) versus H_a: $\theta_2 \neq \theta_{20}$ one rejects H_0 if θ_{20} (the hypothesized value) is excluded from the region. In the case of a randomized block design, one tests for treatment effects by finding an HPD region for the treatment contrasts.

The commonly used other way to test hypotheses is based on the posterior odds ratio but is not implemented in this book. DeGroot (1970), Zellner (1971), and others use the odds ratio, however Box and Tiao (1971) do not. Suppose θ_2 is scalar, that Ω_2 is some non-degenerate interval of the real line and that H_0: $\theta_2 \in (a,b) \subset \Omega_2$ while H_a: $\theta_2 \notin (a,b)$, $\theta_2 \in \Omega_2$, then the posterior odds ratio is

$$O(H_0|H_a,x) = \frac{p_\gamma\{\theta_2 \in (a,b)|x\}}{p_\gamma\{\theta_2 \in \Omega_2 - (a,b)|x\}}$$

where

$$p_\gamma\{\theta_2 \in (a,b)|x\} = \int_a^b \xi(\theta_2|x)\, d\theta,$$

and the larger the ratio the more the indication the null hypothesis is true.

If the null hypothesis is H_0: $\theta_2 = \theta_{20}$ versus H_a: $\theta_2 \neq \theta_{20}$, Jeffreys (1939/1961) expresses prior information about θ in terms of a mixed (part discrete, part continuous) distribution made up of a discrete prior probability of $\xi(\theta_{20})$ at $\theta_2 = \theta_{20}$ and a Probability density $\xi(\theta)$, $\theta \neq \theta_{20}$ for the other values of θ, then the posterior odds ratio for the null versus the alternative is

$$\xi(\theta_{20}|x) \Big/ \int_{\Omega_2 - \{\theta_{20}\}} \xi(\theta|x)\, dx = O(H_0|H_a,x).$$

For an up-to-date account of Bayesian hypothesis testing see Bernardo (1980) and Barnett (1982). Since the posterior odds ratio is still being developed, it will not be used to test hypotheses but instead the HPD region method which was formulated by Lindley (1965) will be employed. Lindley advocates the HPD region only when a vague prior density is appropriate. However, it is not an unreasonable procedure with any prior probability density function, such as a member of a conjugate class and it is done this way, sometimes (see Chapter 3) in this book.

For almost all the linear models of this book, estimation is accomplished by formulas for the posterior mean and variance of the parameters of the model. For example, Chapter 4 examines the random and mixed linear models where the marginal distribution of each variance component is found (approximately) and formulas for the mean and variance are derived. In this way each component can be estimated and an idea of the uncertainty in the estimate is given by the marginal posterior variance. Unfortunately, it was not possible to find estimates for the fixed effects of the model because their marginal posterior distribution cannot be determined analytically and must be determined numerically as was done by Box and Tiao (1973).

Another example of estimation is the Kalman filter, which is the mean of the marginal normal posterior distribution of the successive states of a linear

Another example of estimation is the Kalman filter, which is the mean of the marginal normal posterior distribution of the successive states of a linear dynamic system. This particular method of estimation is explained in Chapter 6.

Sometimes all of the parameters of the model are nuisance parameters, in the sense one is mainly interested in predicting future observations or controlling the future values of the dependent variable of a regression analysis or time series.

Bayes Forecasting

Suppose one observes a series of observations y(l), y(2),…, y(n) and one wants to forecast the next observation y(n + 1) and that the generating mechanism of the series is known and consequently one "knows" the joint density f[y(l),y(2),…, y(n)|θ], $\theta \in \Omega$, of the observations and the prior density for θ, say ξ(θ). Then how does the Bayesian predict the next observation y(n + 1)? The standard way is to find the conditional distribution of y(n + 1) given y(l), y(2),…, y(n) and this is called the Bayesian predictive distribution of y(n + 1) and is

$$\begin{aligned} f[y(n+1)|y(1),\ldots,y(n)] = \int_{\Omega} & f[y(n+1)|y(1),y(2),\ldots,y(n),\theta] \\ & \times \xi(\theta|y(1),\ldots,y(n)]\, d\theta, \end{aligned}$$