STATISTICS 601

Advanced Statistical Methods

Mark S. Kaiser

Department of Statistics

Iowa State University

Spring 2025

Contents

List of Figures

Part I

Advanced Model Construction

Part II

Advanced Model Construction

Chapter 1

Extending Traditional Models

We assume the reader is familiar with non-normal group models (e.g., two samples from gamma distributions), basic generalized linear models, models with additive error (both linear and nonlinear), finite mixture models, and basic mixture and simple hierarchical models. The construction of these classes of models involved several concepts that can be extended to formulate more complex models.

1.1 Parameterized Link Functions

We may formulate a generalized linear model in the usual manner except that, rather than specifying a completely known link function such as $\log(\mu)$, we specify only that a link function be contained in some family of link functions that are indexed by a parameter, λ say. The link function then becomes $g(\mu|\lambda)$ and we may wish to estimate λ along with the other parameters of the model.

1.1.1 Historical Note

To the best of my knowledge, the concept of embedding a link function into an entire family of functions was introduced by Pregibon (1980). Pregibon used this idea primarily to develop a score test for a hypothesized link function. The idea was that we will typically have a some link function we're thinking about using and would like to assess that hypothesized link against a range of alternatives. Suppose that the hypothesized link function can be embedded in a parameterized family of link functions $g(\mu|\lambda)$ that contains both the hypothesized link and what we think of as the true link. For example, suppose that the parameterized family of link functions is the power family,

$$g(\mu|\lambda) = \begin{cases} \mu^\lambda & \lambda \neq 0 \\ \log(\mu) & \lambda = 0 \end{cases}$$

and the true (but unknown) link function also belongs to this family for a particular parameter value, say λ_*. We may be contemplating using a model with a link function in the above family with a particular value of the parameter, $\lambda_0 = 1$ for example would give an identity link and $\lambda_0 = 2$ would give the canonical link for a gamma random component. We might like a test for the hypothesis that $\lambda_* = \lambda_0$, and this is the problem Pregibon addressed.

Pregibon's solution made use of a first order Taylor series for the true link function expanded about the hypothesized link function, resulting in a model that could be "fitted" for one step using the hypothesized link (meaning no new software was needed, which was more important in 1980 than it is now). Pregibon also noted that this procedure was the first step of what could become an iterative solution for maximum likelihood estimation of λ which, although true, was perhaps unfortunate, because Pregibon's procedure has been used in an improper manner under the assumption that maximum like-

lihood estimates resulted (Kaiser 1997). But, there is an easy way to do it right, which we now give.

1.1.2 Maximum Likelihood Estimation

Let $Y_1, \ldots, Y_n$ represent independent response variables from exponential dispersion family distributions with density or mass functions,

$$f(y|\theta_i, \phi) = \exp\left[a(\phi)\{y\theta_i - b(\theta_i)\} + c(y, \phi)\right], \tag{1.1}$$

where $a(\phi) = \phi\, m_i$ for a known set of weights $\{m_i : i = 1, \ldots, n\}$. We have modified our standard form from class just a bit by using this function $a(\phi)$ in place of the simple ϕ; this will be useful in considering binomial random components, in which case the m_i become the binomial sample sizes. To complete the model, let the systematic model component be written as,

$$g(\mu_i|\lambda) = \boldsymbol{x}_i^T \beta = \eta_i \tag{1.2}$$

where $g(\mu|\lambda)$ is some family of link functions specified up to an unknown parameter λ, which may be either a scalar or vector. The i^{th} contribution to the log likelihood now is, for $i = 1, \ldots, n$,

$$\ell_i = a(\phi)\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi), \tag{1.3}$$

and the complete log likelihood is $\ell = \sum_{i=1}^n \ell_i$.

Here, we suppose that $\boldsymbol{x}_i^T = (x_{1,i}, \ldots, x_{p,i})$, $\beta = (\beta_1, \ldots, \beta_p)^T$, and $\lambda = (\lambda_1, \ldots, \lambda_q)^T$. Following the same type of progression we used in Stat 520 for developing a Fisher Scoring algorithm to locate maximum likelihood estimates of the regression parameters β, we have

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\ \frac{\partial \ell_i}{\partial \lambda_k} &= \frac{\partial \ell_i}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}\frac{\partial \mu_i}{\partial \lambda_k}. \end{aligned} \tag{1.4}$$

Note that in (1.4) the third right hand side terms are now partial derivatives rather than the ratio of differentials that we had for basic generalized linear models with fixed link functions.

In a manner similar to what was done for basic generalized linear models, note that

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \theta_i} &= a(\phi)\{y_i - b'(\theta_i)\} = a(\phi)(y_i - \mu_i) \\
\frac{d\theta_i}{d\mu_i} &= \left(\frac{d\mu_i}{d\theta_i}\right)^{-1} = \frac{1}{b''(\theta_i)} = V^{-1}(\mu_i) \\
\frac{\partial \mu_i}{\partial \eta_i} &= \left(\frac{\partial g(\mu_i|\lambda)}{\partial \mu_i}\right)^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} \\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{i,j} \\
\frac{\partial \mu_i}{\partial \lambda_k} &= \left(\frac{-\partial g(\mu_i|\lambda)}{\partial \lambda_k}\right)\left(\frac{\partial g(\mu_i|\lambda)}{\partial \mu_i}\right)^{-1} = \left(\frac{-\partial \eta_i}{\partial \lambda_k}\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1}
\end{aligned} \tag{1.5}$$

Notice that in the third and fifth lines of expression (1.5) we have applied an implicit function theorem as follows. For a basic generalized linear model with fixed link it is easy to write $\mu_i = g^{-1}(\eta_i)$ where $g^{-1}(\cdot)$ is the inverse of the link function $g(\cdot)$. When $g(\cdot)$ is a simple function of one argument this is not difficult (e.g., if $g(x) = \log(x)$ then $g^{-1}(x) = \exp(x)$). But when the link function is parameterized as $g(\cdot|\lambda)$ this is often not possible. Nevertheless, it remains true that $g(\mu_i|\lambda) - \eta_i = 0$ and then implicit functions immediately give lines three and five.

From the expressions (1.4) and (1.5) we can now write the i^{th} contribution to the first derivatives as,

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_j} &= \phi\, m_i(y_i - \mu_i)\, V^{-1}(\mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} x_{i,j} \\
\frac{\partial \ell_i}{\partial \lambda_k} &= \phi\, m_i(y_i - \mu_i)\, V^{-1}(\mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} \left(\frac{-\partial \eta_i}{\partial \lambda_k}\right)
\end{aligned}$$

Now define the terms

$$w_i = m_i \left[\left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V(\mu_i) \right]^{-1},$$

which allows (1.6) to be written as,

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \phi \, (y_i - \mu_i) \, w_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{i,j} \\ \frac{\partial \ell_i}{\partial \lambda_k} &= \phi \, (y_i - \mu_i) \, w_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \left(\frac{-\partial \eta_i}{\partial \lambda_k} \right) \end{aligned} \tag{1.6}$$

The point here is that the contribution of individual terms to the score functions (first derivatives) have been written in the same form for derivatives with respect to the link function parameters as for the regression parameters. That is, the only difference between the first and second lines of (1.6) is the final term on the right hand side.

Now, following the same progression for second derivatives and taking expected values to simplify the expressions, which is laid out in some detail for fixed link models in the Stat 520 notes, we end up with the following expressions.

$$\begin{aligned} -E \left(\frac{\partial^2 \ell_i}{\partial \beta_j \, \partial \beta_l} \right) &= \phi \, w_i x_{i,j} \, x_{i,l} \\ -E \left(\frac{\partial^2 \ell_i}{\partial \beta_j \, \partial \lambda_k} \right) &= \phi \, w_i \left(\frac{-\partial \eta_i}{\partial \lambda_k} \right) x_{i,j} \\ -E \left(\frac{\partial^2 \ell_i}{\partial \lambda_k \, \partial \lambda_h} \right) &= \phi \, w_i \left(\frac{-\partial \eta_i}{\partial \lambda_k} \right) \left(\frac{-\partial \eta_i}{\partial \lambda_h} \right) \end{aligned} \tag{1.7}$$

Let $\xi \equiv (\beta^T, \lambda^T)^T$ be the complete $(p+q)$ vector of systematic model component parameters. Summing over the expressions in (1.6) and collecting the score functions into a vector results in the gradient,

$$\nabla L = \left(\sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_1}, \ldots, \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_p}, \sum_{i=1}^n \frac{\partial \ell_i}{\partial \lambda_1}, \ldots, \sum_{i=1}^n \frac{\partial \ell_i}{\partial \lambda_q} \right)^T \tag{1.8}$$

The (negative) expected second derivatives may be collected into a matrix H as

$$H = \begin{pmatrix} H_1 & H_2 \\ H_2^T & H_3 \end{pmatrix}, \tag{1.9}$$

where

$$\begin{aligned} H_1 \quad &\text{is} \quad p \times p \text{ with } jl^{th} \text{ element } \sum_{i=1}^{n} \phi w_i x_{i,j}\, x_{i,l} \\ H_2 \quad &\text{is} \quad p \times q \text{ with } jk^{th} \text{ element } \sum_{i=1}^{n} \phi w_i \left(\frac{-\partial \eta_i}{\partial \lambda_k}\right) x_{i,j} \\ H_3 \quad &\text{is} \quad q \times q \text{ with } kh^{th} \text{ element } \sum_{i=1}^{n} \phi w_i \left(\frac{-\partial \eta_i}{\partial \lambda_k}\right)\left(\frac{-\partial \eta_i}{\partial \lambda_h}\right) \end{aligned} \tag{1.10}$$

A Fisher Scoring algorithm may then be defined to update a current estimate $\xi^{(m)}$ to a new estimate $\xi^{(m+1)}$ as,

$$\xi^{(m+1)} = \xi^{(m)} + \left(H^{-1}\, \nabla L\right)|_{\xi=\xi^{(m)}}. \tag{1.11}$$

This is really all we need to locate maximum likelihood estimates of the elements of ξ but, as with basic generalized linear models we can perform some further manipulations to arrive at the form of an iteratively weighted least squares algorithm as follows.

Let X_A denote a matrix formed by augmenting the usual X matrix with q additional columns having elements given by the derivatives

$$\frac{-\partial \eta_i}{\partial \lambda_k}; \quad k = 1, \ldots, q$$

Then X_A is an $n \times (p+q)$ matrix with i^{th} row

$$\boldsymbol{x}_{A,i}^T = \left(x_{1,i}, \ldots, x_{p,i}, \frac{-\partial \eta_i}{\partial \lambda_1}, \ldots, \frac{-\partial \eta_i}{\partial \lambda_q}\right)$$

Let W be an $n \times n$ diagonal matrix with elements w_i as given immediately prior to expression (1.6), and let $\boldsymbol{z}^T = (z_1, \ldots, z_n)$ where

$$z_i = (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}.$$

Then we have that

$$\nabla L = \phi X_A^T W \boldsymbol{z} \quad \text{and} \quad H = \phi X_A^T W X_A, \tag{1.12}$$

and the Fisher Scoring algorithm of expression (1.11) becomes

$$\begin{aligned} \xi^{(m+1)} &= \xi^{(m)} + \left[\left(X_A^T W X_A \right)^{-1} X_A^T W \boldsymbol{z} \right] |_{\xi=\xi^{(m)}} \\ &= \left[\left(X_A^T W X_A \right)^{-1} \left(X_A^T W X_A \xi + X_A^T W \boldsymbol{z} \right) \right] |_{\xi=\xi^{(m)}} \\ &= \left[\left(X_A^T W X_A \right)^{-1} X_A^T W \boldsymbol{z}^* \right] |_{\xi=\xi^{(m)}}, \end{aligned} \tag{1.13}$$

where, $\boldsymbol{z}^* = (z_1^*, \ldots, z_n^*)^T$ with

$$z_i^* = \boldsymbol{x}_{A,i}^T \xi + z_i$$

The last line of (1.13) is in the form of an iteratively re-weighted least squares algorithm. Note here that the matrix X_A does not remain fixed in this algorithm as it needs to be evaluated at the current estimate $\xi^{(m)}$ at each iteration.

Useful Families of Link Functions

If we would like to estimate parameters of link functions we need useful families of such functions, and developing these is not a trivial task. Consider, for example, the power family given earlier,

$$g(\mu|\lambda) = \begin{cases} \mu^\lambda & \lambda \neq 0 \\ \log(\mu) & \lambda = 0 \end{cases} \tag{1.14}$$

This is a fine way to write a family of functions if our use is to select a power for a fixed link, but it is not so useful for estimation (of λ) if our desire is to separate a log link from some other power. For this, we might consider a family of link functions given by Pregibon (1980) as,

$$g(\mu|\lambda) = \frac{1}{\lambda_2} \left[(\mu + \lambda_1)^{\lambda_2} - 1 \right]. \tag{1.15}$$

This family includes, for example, the identity link which results from taking $\lambda_1 = \lambda_2 = 1$. It also includes the log link if we take $\lambda_1 = 0$ and let $\lambda_2 \to 0$,

$$\lim_{\lambda_2 \to 0} g(\mu|\lambda_1 = 0, \lambda_2) = \lim_{\lambda_2 \to 0} \frac{\mu^{\lambda_2} - 1}{\lambda_2} = \log(\mu).$$

Note that the difference with the power family (1.14) is that in (1.14) the link was *defined* as log for $\lambda = 0$ while in (1.15) we get the log link as the value of λ_2 goes to zero. This makes a difference if our objective is to estimate λ. What about other powers? Consider using $\lambda_1 = 0$ and $\lambda_2 = 2$ in (1.15). This gives

$$g(\mu|\lambda) = \frac{\mu^2 - 1}{2}.$$

Now, suppose that the linear predictor is $\eta_i = \beta_0 + \beta_1 x_i$. Then we would have the systematic model component,

$$g(\mu_i|\lambda) = \beta_0 + \beta_1 x_i \Rightarrow \frac{\mu^2 - 1}{2} = \beta_0 + \beta_1 x_i \Rightarrow \mu_i^2 = (2\,\beta_0 + 1) + 2\,\beta_1 x_i$$

or,

$$\mu_i^2 = \gamma_0 + \gamma_1 x_i$$

so that our model is one with an ordinary squared link function.

Another useful family of link functions that we will use in analysis of a short-term toxicity test is,

$$g(\mu|\lambda) = \log\left[\frac{(1-\mu)^{-\lambda} - 1}{\lambda}\right].$$

This family includes the logit link for $\lambda = 1$,

$$g(\mu|\lambda = 1) = \log\left[\frac{1}{(1-\mu)} - 1\right] = \log\left(\frac{\mu}{1-\mu}\right),$$

and the complementary log-log link as $\lambda \to 0$,

$$\begin{aligned}
\lim_{\lambda\to 0} g(\mu|\lambda) &= \lim_{\lambda\to 0} \log\left[\frac{(1-\mu)^{-\lambda}-1}{\lambda}\right] \\
&= \log\left[\lim_{\lambda\to 0}\frac{1}{\lambda}\left\{\frac{1}{(1-\mu)^{\lambda}}-1\right\}\right] \\
&= \log\left[\lim_{\lambda\to 0}\frac{1-(1-\mu)^{\lambda}}{\lambda(1-\mu)^{\lambda}}\right] \\
&= \log\left[\lim_{\lambda\to 0}\frac{-\log(1-\mu)}{1+\lambda\log(1-\mu)}\right] \\
&= \log\left[-\log(1-\mu)\right].
\end{aligned}$$

1.1.3 Analysis of Short-Term Toxicity Test Data

The design of a short-term toxicity test is quite simple. We have k concentrations of some potentially toxic substance and we expose groups of $n_1, n_2, \ldots, n_k$ organisms to these concentrations for a fixed period of time. At the end of that time the number of organisms in each group that have "responded" (usually died) is recorded. Such data are often called quantal response data.

The theoretical basis for the analysis of quantal response data is somewhat more complex than the experimental design. The fundamental elements of this theory are as follows.

1. It is supposed that for each individual organism there is a concentration of the toxicant, x say, such that the organism will respond for any con-

centration greater than or equal to x and the organism will not respond for any concentration less than x. This value is called the *tolerance* of the organism. That is, if R_j denotes the response of organism j, x_j its tolerance, and d the concentration to which it is exposed,

$$Pr(R_j = 1|d < x_j) = 0 \qquad Pr(R_j = 1|d \geq x_j) = 1$$

2. It is also supposed that, in the population of organisms, the tolerances x_j follow some distribution that is a location-scale family with distribution function G, mean μ_x and variance σ_x^2. Then the standardized tolerances are such that,

$$\tilde{x}_j = \frac{x_j - \mu_x}{\sigma_x} \sim iidG(0, 1)$$

3. Assume that at a given experimental concentrations (dose) d_i; $i = 1, \ldots, k$, there is a certain probability p_i; $i = 1, \ldots, k$ that the dose will exceed the tolerance of a randomly chosen organism so that, if Y_i is defined as the number of responses out of n_i organisms at dose i, the probability mass function of Y_i is,

$$f(y_i|p_i) = \frac{n_i!}{y_i!(n_i - y_i)!}\, p_i^{y_i}\, (1 - p_i)^{n_i - y_i}; \quad y_i = 0, 1, \ldots, n_i \qquad (1.16)$$

4. From the above we have that, at a given dose d_i,

$$p_i = \int_{-\infty}^{\delta_i} d\, G, \quad \text{where} \quad \delta_i = \frac{d_i - \mu_x}{\sigma_x}. \qquad (1.17)$$

That is, δ_i is the $p(100)\%-$tile of $G(0, 1)$, the distribution of standardized tolerances.

5. The objective is, given fixed $d_1, \ldots, d_k$, fixed $n_1, \ldots, n_k$ and observed $y_1, \ldots, y_k$, estimate μ_x and σ_x^2, the parameters of the tolerance distribution $G(\mu_x, \sigma_x^2)$. In particular, if the response is mortality and G is

assumed or chosen to be symmetric, μ_x is often called the "median effective dose" (MED), or the "lethal concentration that kills 50%" (LC_{50}).

Formulation as a Generalized Linear Model

To formulate this problem as a standard glm, take the response variables to be expressed as observed proportions rather than the counts of expression (1.16). The mass functions of these Y_i may be written in exponential dispersion family form as,

$$f(y_i|\theta_i) = \exp[a(\phi)\{y_i\theta_i - b(\theta_i)\} + c(y_i)], \tag{1.18}$$

where

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right); \quad b(\theta_i) = \log\{1+\exp(\theta_i)\}; \quad \text{and} \quad \phi \equiv 1; \;\; a(\phi) = n_i$$

Now, from (1.17) $G(\delta_i) = p_i$ so then,

$$G^{-1}(p_i) = \delta_i = \frac{d_i - \mu_x}{\sigma_x} = \frac{-\mu_x}{\sigma_x} + \frac{1}{\sigma_x}d_i, \tag{1.19}$$

which completes a standard generalized linear model with binomial random component, link function G^{-1}, and regression parameters $\beta_0 = -\mu_x/\sigma_x$ and $\beta_1 = 1/\sigma_x$.

Notice from this development that there is a one-to-one relation between distinct tolerance distributions $G(\mu_x, \sigma_x^2)$ and link functions. In particular some of the typical links and tolerance distributions are:

- Normal tolerance distribution, probit link
- Logistic tolerance distribution, logit link

- Extreme value tolerance distribution, complementary log-log link

Now, we would like to fit a model with random component (1.18), the linear predictor (1.19) and using the family of link functions,

$$g(\mu_i|\lambda) = \log\left[\frac{(1-\mu)^{-\lambda}-1}{\lambda}\right]. \tag{1.20}$$

Notice here that we are using μ_i for the expected value of Y_i. We have also used μ_x for the expected value of the tolerance distribution. It will be important to maintain this distinction in what is to come. Continuing to use $d_i : \ i = 1, \ldots, k$ as the experimental doses (i.e., the covariates in the glm), our systematic model component is,

$$g(\mu_i|\lambda) = \eta_i = \beta_0 + \beta_1 d_i; \quad i = 1, \ldots, k$$

The maximum likelihood algorithm presented earlier can be used to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. To do this, however, requires computing a number of quantities in a different manner than would be the case for a model with fixed link. I will not present derivations here, but will list the quantities that would need to be calculated in order to implement the algorithm given in expression (1.13).

1. Linear Predictor
$$\eta_i = \beta_0 + \beta_1 d_i$$
2. Means (of Y_i)
$$\mu_i = 1 - \frac{1}{\{1+\lambda\exp(\eta_i)\}^{1/\lambda}}$$
3. Derivative of η_i wrt μ_i
$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{\lambda}{(1-\mu_i)\{1-(1-\mu_i)^{\lambda}\}}$$

4. Weights

$$w_i = n_i \left[\left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V(\mu_i) \right]^{-1}$$

where $V(\mu_i) = \mu_i(1 - \mu_i)$ as for any model with random component (16)

5. Derivative of η_i wrt λ

$$\frac{\partial \eta_i}{\partial \lambda} = \frac{-\log(1 - \mu_i)}{1 - (1 - \mu_i)^\lambda} - \frac{1}{\lambda}$$

Estimation of Tolerance Distributions

With maximum likelihood estimates of β_0 and β_1 in hand we have, from (1.19) and the invariance property of maximum likelihood, that maximum likelihood estimates of the tolerance function parameters are,

$$\hat{\mu}_x = \frac{-\hat{\beta}_0}{\hat{\beta}_1}; \qquad \hat{\sigma}_x = \frac{1}{\hat{\beta}_1} \tag{1.21}$$

Now, under the theory developed, link function is the inverse distribution function for standardized tolerances (denoted as $\tilde{x}_j$ previously). The distribution function and density function of the tolerances x_j may then be found by inverting expression (1.20) as follows. We wish to find $G(\tilde{x}) = p$ for some $0 < p < 1$ given that

$$\tilde{x} = G^{-1}(p|\lambda) = \log \left[\frac{1}{(1 - p)^\lambda} - 1 \right] - \log(\lambda)$$

Then,

$$\begin{aligned} & \frac{1-(1-p)^{\lambda}}{(1-p)^{\lambda}\,\lambda} = \exp(\tilde{x}) \\ \Rightarrow \quad & \exp(\tilde{x})\,\lambda(1-p)^{\lambda} = 1-(1-p)^{\lambda} \\ \Rightarrow \quad & (1-p)^{\lambda}\,\{\lambda\,\exp(\tilde{x})+1\} = 1 \\ \Rightarrow \quad & p = 1-\frac{1}{\{\lambda\,\exp(\tilde{x})+1\}^{1/\lambda}}. \end{aligned}$$

So the distribution function for standardized tolerances is then,

$$G(\tilde{x}|\lambda) = 1-\frac{1}{\{\lambda\,\exp(\tilde{x})+1\}^{1/\lambda}}. \tag{1.22}$$

Since tolerance x_j corresponds to standardized tolerance $\tilde{x}_j$ as $\tilde{x}_j = (x_j - \mu_x)/\sigma_x$, to obtain the actual distribution function of tolerances, we simply make the appropriate location and scale transformations as

$$x_j = \sigma_x\,\tilde{x} + \mu_x$$

and then,

$$G(x|\lambda,\mu_x,\sigma_x) = \left[1-\frac{1}{\{\lambda\,\exp\{(x-\mu_x)/\sigma_x\}+1\}^{1/\lambda}}\right] \tag{1.23}$$

which has density

$$g_t(x|\lambda,\mu_x,\sigma_x) = \frac{\exp\{(x-\mu_x)/\sigma_x\}}{\sigma_x\,[\lambda\exp\{(x-\mu_x)/\sigma_x\}+1]^{(1+1/\lambda)}}. \tag{1.24}$$

An estimated tolerance density then results from substitution of the estimates $\hat{\mu}_x$ and $\hat{\sigma}_x$ from (1.21) into (1.24).

Concentration (log)	No. Mortalities	No. Exposed
3.893	6	59
3.970	13	60
4.042	18	62
4.108	28	56
4.171	52	63
4.230	53	59
4.285	61	62
4.338	60	60

Table 1.1: Bliss Beetle Data

Example 1.1 – Bliss Beetle Data

One version of the famous Bliss beetle data is presented in Table 1.1. These data have been presented in a number of forms since there were two replicates that are sometimes combined and sometimes not, and concentration is reported on various scales. The data arose from a short-term toxicity test conducted with flour beetles and gaseous carbon disulphide exposure for 5 hours.

Estimates of the parameters in the systematic model component were located using a model with a fixed logit link, and a model using the family of links in expression (1.20). The results are given Table 1.2. Estimates of μ_x and σ_x were obtained from expression (1.21).

Because the model with a fixed logit link function is nested within the model having an estimated link function (by taking $\lambda = 1$) we may conduct a likelihood ratio test to compare a reduced model (logit link) with a full model (estimated link). The maximized log likelihood (sans constant terms) of the

	Estimates Under Model With	
Parameter	Logit Link	Link Family
β_0	-60.640	-39.352
β_1	14.865	9.518
λ	NA	-0.006
μ_x	4.079	4.134
σ_x	0.067	0.105

Table 1.2: Point Estimates for Bliss Beetle Data

logit link model was -186.1993, while that for the model with estimated link was -182.3464. This results in a likelihood ratio test statistic $T = -2(-186.1003 + 182.3464) = 7.7058$ and an associated p-value of $p = 0.0055$ (from comparison with a χ^2 distribution having 1 degree of freedom). Thus, we would prefer the model with an estimated link in this example. A plot of the observed responses and fitted curves from both the model with a logit link and the model with an estimated link is presented in the upper panel of Figure 1.1. A plot of the estimated tolerance density functions is presented in the lower panel of Figure 1.1, in which one can see the difference between the symmetric tolerance density dictated by the logit link (dashed curve) and the asymmetric density that results from the estimated link function.

Example 1.2 – Sub-lethal Exposure of Trout to Petroleum Hydrocarbons

We will apply the same methods used with the Bliss beetle data to another example, this involving the effect of sub-lethal exposure of Rainbow Trout to a petroleum hydrocarbon on response to lethal concentrations of the same

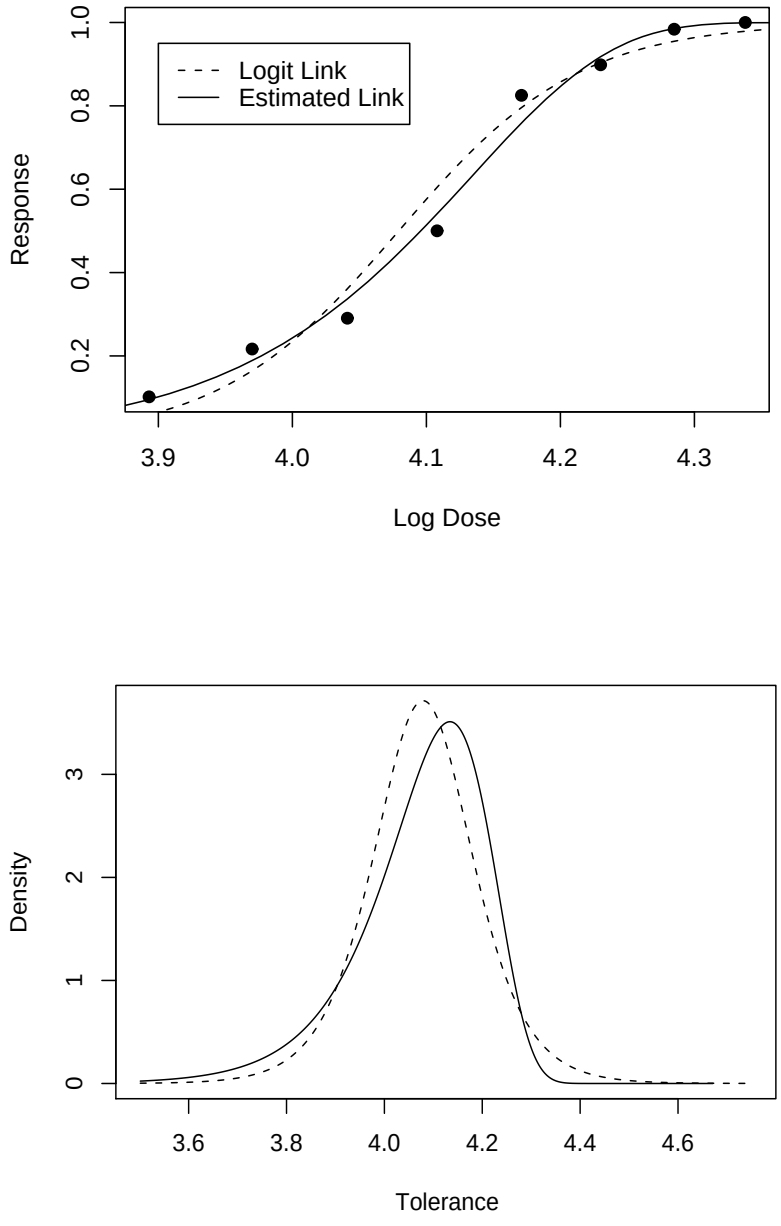

Figure 1.1: Estimated response curves (upper panel) and tolerance distributions (lower panel) for the Bliss beetle data.

substance. The scientific underpinnings of this study involved the fact that petroleum hydrocarbons (e.g., gas and fuel oils) are frequently spilled into aquatic environments. Once released into the environment, petroleum hydrocarbons bind to sediments and are slowly released back into the water column. Thus, in areas with natural hydrocarbon seeps (near natural deposits), around processing facilities (e.g., offshore oil platforms), or in areas that have been previously polluted, petroleum hydrocarbons are found at the level of μg/L, which are generally not fatal to fish. On the other hand, "major spills" do also occur with resulting concentrations in the range of mg/L (much higher). The question being investigated was whether a low level of "pre-exposure" makes Rainbow Trout more or less susceptible to higher levels of exposure, should they occur. There is some controversy about whether a low level of chronic exposure makes organisms *more resistant* or *less resistant* to particular toxicants.

In the study, test organisms (immature rainbow trout) were exposed to low levels of number 2 fuel oil (2FO) for 21 days. The concentrations used in this "pre-exposure period included 0, 25, and 50 mg/L of 2FO. After 21 days, fish were transferred to other tanks and portions of each pre-exposure group were exposed to 2FO at concentrations of 28.7, 57.4, 114.8, 229.6, 459.1 and 918.2 mg/L for a total of 335 hours. In the actual study, mortality was recorded every 4 or 8 hours and time to death was one of the response variables examined. Here, we will simply consider the data at 258 hours within the context of a typical dose-response analysis. The overall objective is to determine whether the pre-exposure groups differ in their response during the main toxicity test (at least at the time point of 258 hours). The data are presented in Table 1.3. We will analyze the data in the same way as for the Bliss beetle data, with a generalized linear model having binomial

	Pre-exposure concentration					
	0 μg/L		25 μg/L		50 μg/L	
Exposure (mg/L)	Y	N	Y	N	Y	N
28.7	0	10	0	12	1	9
57.4	0	10	1	10	0	10
114.8	2	10	1	10	2	10
229.6	4	10	4	10	5	10
459.1	8	9	8	10	7	10
918.2	10	10	9	9	10	10

Table 1.3: Mortalities (Y) and number exposed (N) in a toxicity test with 2FO.

random component and either logit or estimated link functions. Note that all of the results presented will come from models using log concentration as the covariate.

Parameter estimates for the control treatment (e.g., 0 μg/L of pre-exposure) are presented in Table 1.4 for models with fixed logit and estimated link functions. Parameter estimates for the pre-exposure treatment group of 25 μg/L are presented in Table 1.5 for models with fixed logit and estimated link functions. Parameter estimates for the pre-exposure treatment group of 50 μg/L are presented in Table 1.6 for models with fixed logit and estimated link functions.

Because a model with a fixed logit link is nested within a model with link family (1.20) we may conduct likelihood ratio tests between logit link and estimated link models for each pre-exposure group. Maximized log likelihoods and these likelihood ratio tests are presented in Table 1.7

	Model Link			
	Logit		Estimated	
Parameter	Estimate	Std. Error	Estimate	Std. Error
β_0	-15.820	4.172	-12.164	4.662
β_1	2.905	0.765	2.140	0.907
λ	NA	NA	0.129	0.811

Table 1.4: Parameter estimates for pre-exposure group 0 μg/L.

	Model Link			
	Logit		Estimated	
Parameter	Estimate	Std. Error	Estimate	Std. Error
β_0	-12.896	3.194	-9.422	2.943
β_1	2.338	0.580	1.591	0.553
λ	NA	NA	-0.199	0.567

Table 1.5: Parameter estimates for pre-exposure group 25 μg/L.

	Model Link			
	Logit		Estimated	
Parameter	Estimate	Std. Error	Estimate	Std. Error
β_0	-9.897	2.409	-6.580	2.132
β_1	1.816	0.442	1.069	0.404
λ	NA	NA	-0.482	0.386

Table 1.6: Parameter estimates for pre-exposure group 50 μg/L.

	Maximized Likelihoods			
Pre-exposure	Logit	Estimated	T	p-value
0 μg/L	−15.7426	−15.5222	1.0412	0.3075
25 μg/L	−19.3685	−18.7538	1.229	0.2675
50 μg/L	−23.7589	−22.7459	2.026	0.1546

Table 1.7: Likelihoods and LRT tests for logit versus estimated link models.

Pre-exposure	$\hat{\mu}_x$	95% Interval
0 μg/L	5.685	(5.049, 6.319)
25 μg/L	5.923	(5.304, 6.543)
50 μg/L	6.154	(5.303, 7.004)

Table 1.8: Estimates of tolerance distribution means.

Based on these results we would conclude that there is not sufficient evidence in the data to say that the tolerance distributions in any of the pre-exposure groups differ from a logistic. There is, however, an interesting pattern that suggests itself. The estimated values of the link function parameter λ appear to be decreasing as one moves from 0 to 25 to 50 μg/L pre-exposure (Tables 1.4, 1.5, 1.6). Concomitantly, the $p-$values for likelihood ratio tests are becoming smaller as well (Table 1.7). This might peek our curiosity as to whether anything is being "suggested" by the data in terms of a systematic pattern in the tolerance distributions.

Using results from the models with estimated link functions, values for μ_x and σ_x were arrived at through the use of expression (1.21) and their standard errors were computed using the delta method in the usual manner. Point and 95% interval estimates of μ_x are given in Table 1.8

These intervals certainly overlap to a great extent, and the same is true for intervals computed under the model with a fixed logit link (not shown). Thus, we are led to the belief that the data do not provide sufficient support for claiming any difference at all between the pre-exposure groups. There is insufficient evidence in the data that pre-exposure of trout to sublethal levels of 2FO changes the response to lethal concentrations at all. What happens if we plot the estimated tolerance distributions for the pre-exposure groups? The estimated densities are presented in Figure 1.2

This figure does suggest a systematic change in the tolerance distributions as the level of pre-exposure to 2FO increases from 0 to 25 to 50 μg/L, but is that suggestion one of sensitization (i.e., becoming more susceptible) or acclimation (i.e., becoming less susceptible) to the toxicant. Our eye is drawn to the left tails of these estimated densities, which are apparently becoming heavier as the level of pre-exposure increases. But note also the upper portions of the densities which also contain more and more probability as pre-exposure increases. Recall that examination of cumulative densities can often aid in interpretation. The cumulative densities corresponding to the distributions of Figure 3 are presented in Figure 1.3.

These estimated cumulative densities suggest an effect of acclimation since the rate at which probability (of mortality) is accumulating in these distributions is slower for pre-exposure of 50 μg/L than it is for pre-exposure of 25 μg/L which is in turn slower than for no pre-exposure to 2FO. For example, the cumulative probability in these distributions at a log concentration of 6.0 is 0.862 for the 0 μg/L group, 0.642 for the 25 μg/L group, and 0.432 for the 50 μg/L group.

The overall suggestion from fitting these models is that the effect of pre-exposure, if there in fact is one, is an acclimation effect. Recall that we are

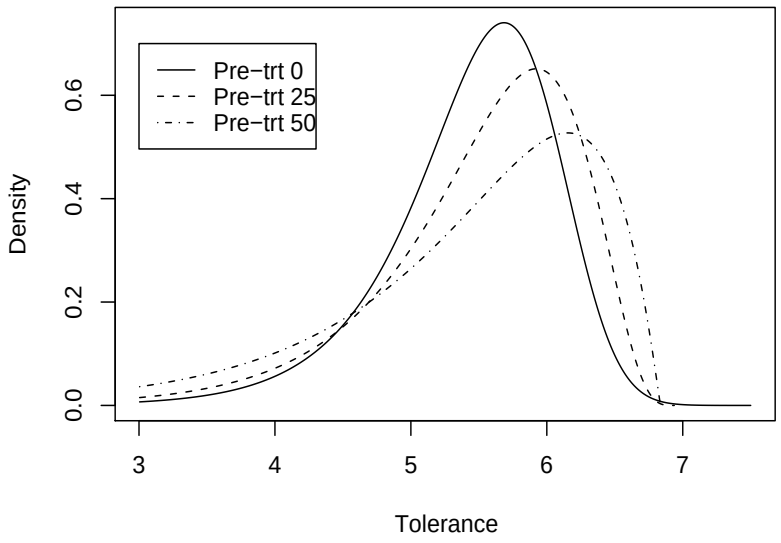

Figure 1.2: Estimated tolerance densities for Pre-exposure groups.

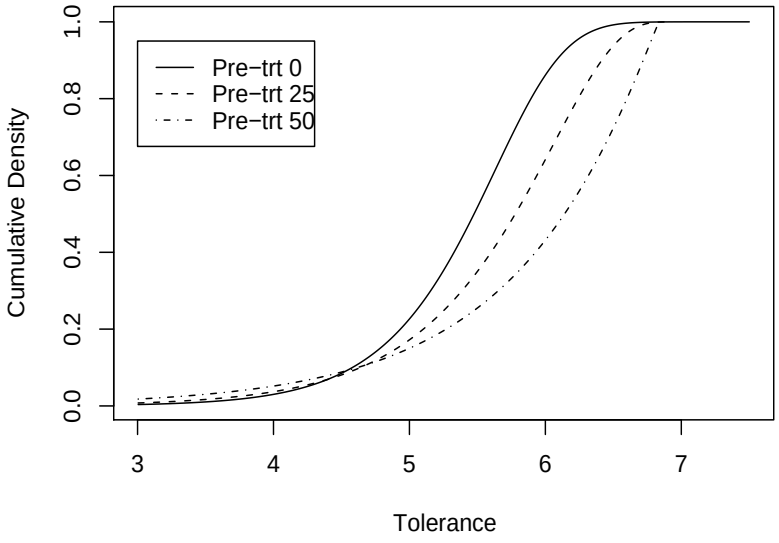

Figure 1.3: Estimated cumulative tolerance densities for Pre-exposure groups.

not able to conclude there is sufficient evidence in the data to proclaim any effect at all. But consider whether a conclusion that there is, in fact, no effect of pre-exposure is warranted. The study contained small sample sizes and we might also suggest that in order to detect this type of difference, if indeed one exists, one needs increased information. The point is that, although we cannot conclude from this study that there is any effect of pre-exposure, we are able to suggest (1) if such an effect is real it is likely to be acclimation rather than sensitization, and (2) in order to detect such differences a larger sample size is needed.

1.2 Lessons from Generalized Linear Models

While the unified algorithm for estimation of basic generalized linear models is nice, of greater value are the lessons for model construction that result from the development of this class of models. In particular, recall that writing distributions in the form of exponential dispersion families allowed us to isolate a parameter that governs expected values. That parameter, or a simple function of it, was allowed to vary across a set of random variables while the other parameter, if present, was held constant. Along the way, each distribution dictated a certain relation between means and variances, which assisted in selection of the model random component. There are distributions that cannot be put into the form of exponential dispersion families but to which we can apply these same ideas.

1.2.1 Example 1.3 – Storage Time of Meat

The production of what is called "case-ready" meat by processing plants has become quite popular, especially in grocery stores and supermarkets that are part of large national or regional chains. Such case-ready meats are delivered to the store in pre-packaged containers and reduce or remove the need for the store to have a butcher on staff. A major concern with this process is shelf-life of the packaged meats, particularly with respect to red meats such as beef. The aging process of beef results in color changing from pink to brown, and this occurs prior to spoilage. And, color is one of the primary characteristics used by consumers in choosing whether or not to buy beef. As a result, many grocers end up discarding large amounts of discolored beef, even when that beef is perfectly safe and healthy to eat.

It has been discovered that a packaging process that uses vacuum packaging with a tiny amount of carbon monoxide inserted into the package helps retard discoloration of beef, making the product look fresher for longer, and thus extending the effective shelf-life of case-ready beef products. This practice has, of course, become quite controversial, with any number of consumer advocate groups claiming it amounts to false advertising and is unsafe (even if a product is safe when purchased, if it looks fresher than it is, a consumer might store it for longer than they otherwise would before eating it, increasing the risk that it is spoiled when consumed). In March 2006, the City of Chicago considered a ban on carbon monoxide packaging (I don't know what they decided) and in mid-2006 the FDA was petitioned by a number of groups to ban the practice nationwide (I think that's still pending).

In a major study to examine the effect of carbon monoxide packaging on the appeal of beef steak to consumers, a food science department conducted

the following study. Over a period of 25 days, fresh beef was obtained every 12 hours from a packing plant. Two steaks were chosen at each point in time, chosen to be the same "grade" by meat experts; grade involves the amount of fat marbling, quality of trim to remove excess fat around the edges, and so forth. One steak was packaged in the traditional manner while the other was packaged using the carbon monoxide process. The total of 100 steaks were then judged in terms of how "desirable" each was on a scale of 1 to 10 by a large panel of 2000 consumers. The data reported are in the form of a proportional score; the total score for each steak was the total score for that steak divided by the score of the "most desirable" steak (the steak receiving the highest total score).

The objective in a statistical analysis of these data is to relate proportional score for desirability to storage time, and to determine if there is a systematic difference between traditional packaging and packaging with carbon monoxide. The relation at 14 days is of particular interest, because 14 days is typically taken as the "effective" shelf-life for steak packaged with the traditional method (i.e., after 14 days, traditionally packaged steak is often marked as "reduced for quick sale" or discarded by grocers).

The example just described involves a situation in which response variables might be taken as independent, the objective being to formulate a regression of responses on covariates (time in the example) and in which the possible values of response variables would most correctly be taken as the unit interval for each response. One could certainly take a transformation of the responses to make a model with additive errors more palatable (e.g., the so-called angular transformation) but doing so would certainly complicate making inference on the original scale. And in the example we already know that desirability score is going to decrease with time, so finding that there is

an inverse relation between the freshness of meat and time in package is not going to earn a paycheck. The natural random component to choose in the example would be to take the responses as having beta distributions but this places us solidly outside the realm of basic generalized linear models because the beta density cannot be coerced into the form of an exponential dispersion family.

1.2.2 A Model with Beta Response Distributions

To determine whether we can formulate a useful regression model with beta response distributions we can begin by attempting to achieve the same ends that result from exponential dispersion family random components. These are the isolation of expected values that can vary over observations, another parameter that can be assumed constant over observations, and the identification of a relation between expected values and variances. A first step is to determine a mean value parameterization of a beta density. A standard form for a beta density is, for $\alpha > 0$ and $\beta > 0$,

$$f(y|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}; \quad 0 < y < 1. \tag{1.25}$$

For a random variable Y that follows this density, the expected value is

$$\mu = E(Y) = \frac{\alpha}{\alpha+\beta}.$$

In the parameterization of (1.25) the variance becomes

$$var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)} = \mu(1-\mu)\frac{1}{\alpha+\beta+1}.$$

So now let $\phi = 1/(\alpha+\beta+1)$.

To formulate a regression then let $Y_1, \ldots, Y_n$ be random variables with beta distributions, expected values $E(Y_i) = \mu_i$ and constant dispersion parameter ϕ. The variances of these variables are $var(Y_i) = \phi\,\mu_i\,(1-\mu_i)$. The

model would be completed by modeling the μ_i in terms of covariates $\boldsymbol{x}_i$ and parameters $\boldsymbol{\gamma}$, say

$$\mu_i = h(\boldsymbol{x}_i, \boldsymbol{\gamma}) \tag{1.26}$$

If we wanted to mimic generalized linear models we might take $h(\cdot)$ to correspond to a simple link function such as $h(\boldsymbol{x}_i, \boldsymbol{\beta}) = \exp(\gamma_0 + \gamma_1 x_i)/[1 + \exp(\gamma_0 + \gamma_1 x_i)]$ but there is really no need to do so.

A question about this model is how much the assumption that ϕ is constant over observations restricts distributional shapes across levels of the covariate. Note that both α and β in (1.25) can vary across covariate values, but only in a manner such that ϕ remains constant.

Estimation and Inference

Estimation and inference can proceed according to a likelihood or Bayesian approach. In either case, it is probably more convenient to write the likelihood in terms of α_i and β_i and use the relations

$$\begin{aligned} \alpha_i &= \left(\frac{1}{\phi} - 1\right)\mu_i \\ \beta_i &= \left(\frac{1}{\phi} - 1\right)(1 - \mu_i) \end{aligned} \tag{1.27}$$

to translate between (μ_i, ϕ) and (α_i, β_i). The expected values $\{\mu_i : i = 1, \ldots, n\}$ are of course written as functions of the elements of $\boldsymbol{\gamma}$ as in (1.26) so the focus of estimation is actually $\boldsymbol{\gamma}$ and ϕ, and derivatives for a likelihood analysis are most easily derived using the chain rule. The parameter space of ϕ is the unit interval $\phi \in (0, 1)$ so some type of beta prior is natural to use for this parameter in a Bayesian analysis. Assuming that the systematic model component given by $h(\cdot)$ in (1.26) allows elements of $\boldsymbol{\gamma}$ to assume values on

the entire line, diffuse normal priors would be a naive choice for nearly any particular model.

1.3 Focusing on Random Model Components

A device we relied on in selecting random components in basic generalized linear models, and powers for power of the mean models in additive error regressions, was to relate response variances to expected values. In formulating models that are not members of these classes, mean-variance relations can still be valuable, although what they tell us about a problem may be different than what we are familiar with.

1.3.1 Example 1.4 – Soil Respiration and Temperature

In this era of climate change there is great interest in "sources and sinks" of carbon in the environment due to the role of carbon compounds in global warming. For example, the Amazon rain forest is estimated to have somewhere around 100 billion tons (or 75 billion tonnes) of carbon stored in trees. Trees do respire, and this releases carbon on a regular basis, as well as isoprene, a catalyst in the production of ozone (this is believed to be the reason for the "gloriously stupid concept" espoused by President Ronald Reagan in 1981 that trees 'cause more pollution than automobiles do')

`https://rationalwiki.org/wiki/Trees_cause_pollution.`

Certainly, when the forest is cleared (often by burning) huge amounts of carbon dioxide are released into the atmosphere, so destruction of forest causes substantial pollution and reduces the capacity of the environment to capture additional carbon. Although this role for the Amazon forest is impressive, it

pales in comparison with the amount of carbon in soils, estimated at about 2,500 gigatons. Soil, too, respires. More correctly, roots, bacteria, fungi, and subterranean animals respire. So soils also contribute to the carbon dioxide load of the atmosphere, as well as being a major carbon sink. It is generally accepted that soil respiration increases with soil temperature. This has apparently caused some to wonder if there might not be a "positive feed-back loop" in which a warming climate leads to greater soil respiration which, in turn, leads to more carbon being released to the atmosphere and, therefore, more warming. In a study that investigated the relation between soil respiration, soil temperature, soil moisture, and other factors, Raich *et al.* (2021) used observations from four locations that included temperate grasslands, northern forests, and tropical forest plantations. A scatterplot of soil respiration versus soil temperature from one of those locations is reproduced in Figure 1.4. The variable Rsoil is soil respiration and has units of grams of carbon per square meter per day. Soil temperature is measured in degrees celcius at a depth of 5 cm.

It is possible, but quite rare, for soil respiration to be negative at temperatures of less than 0 degrees C. There are a few such values evident in the scatterplot of Figure 1.4 and it would not be unreasonable to delete these values and assume that response variables are strictly positive. Even if we would use response distributions that accommodate negative values, we would want the left tail of those distributions to be 'short' relative to the right tail. The scatterplot also hints at right skew response distributions for larger values of the covariate of soil temperature. Taking all of this into consideration, we might decide that a regression having gamma random component would be a reasonable place to start in developing a model for these data after dropping negative responses. Figure 1.4 also suggests that a log link might not be a

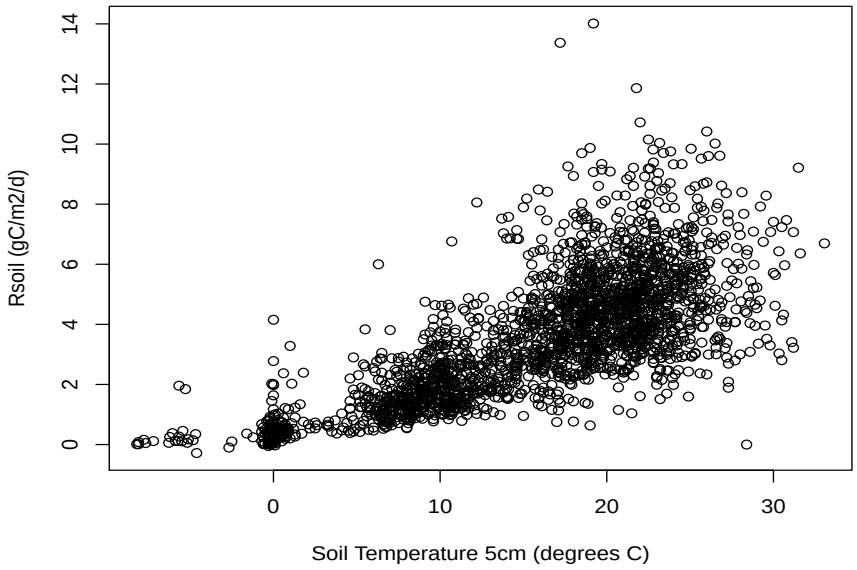

Figure 1.4: A scatterplot of Rsoil against Tsoil.

bad choice (the leftmost "tail" in the scatterplot is what argues against an identity link). So, as a fist try to model these data we might fit a basic generalized linear model with log link and gamma random component. If we do so, we find that the deviance residual plot in Figure 1.5 indicates that the model implies that response variances are given as too large a power of the expected values. That is, the data do not support a model in which $var(Y) \propto \mu^2$, as implied by the basic glm with gamma random component. Box-Cox plots for these data (not shown) suggest that there is a relation between variances and expected values in the form of $var(Y) \propto \mu^d$, but indicate that d should be chosen somewhere around 1, not 2. We are now in a situation for which we would like to (at least initially) maintain a gamma random component, but would also like to model variances as proportional to the expected values.

1.3.2 A Gamma Model

A basic glm with gamma random component is no longer an attractive possibility. In the development of a basic glm with a gamma random component, expected values and variances are related as follows. Begin with potentially different gamma parameters, α_i and β_i for each response random variable Y_i; $i = 1, \ldots, n$. To obtain variances as proportional to some power of the expected values we need,

$$\frac{\alpha_i}{\beta_i^2} \propto \left(\frac{\alpha_i}{\beta_i}\right)^c.$$

A basic glm takes the factor of proportionality to be $1/\alpha_i$ and then makes this constant across values of Y_i, resulting in,

$$\frac{\alpha}{\beta_i^2} = \frac{1}{\alpha}\left(\frac{\alpha}{\beta_i}\right)^2.$$

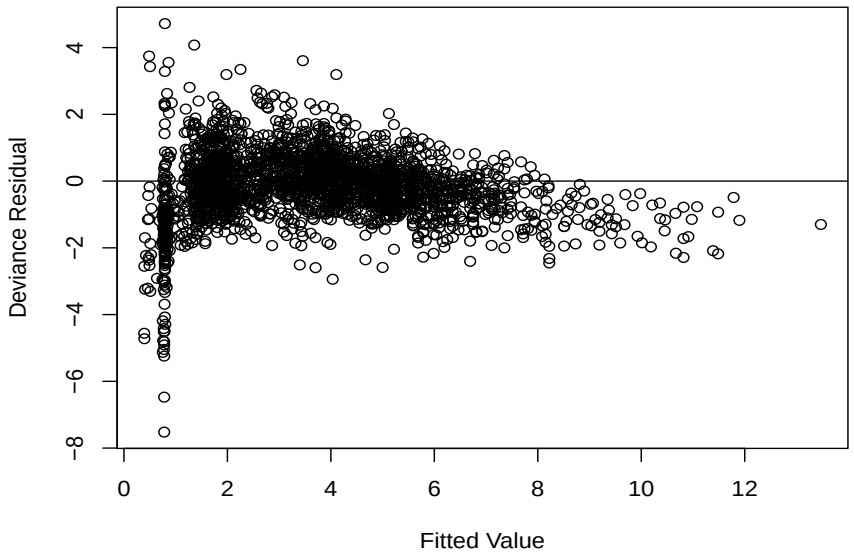

Figure 1.5: Deviance residuals for a standard glm with gamma random component and log link.

But if we would take the proportionality factor to be $1/\beta_i$, and then make this constant across values of Y_i we would arrive at,

$$\frac{\alpha_i}{\beta^2} = \frac{1}{\beta}\left(\frac{\alpha_i}{\beta}\right).$$

We could then formulate a model with gamma response distributions such that

$$\begin{aligned} E(Y_i) &= \mu_i = \frac{\alpha_i}{\beta} \\ var(Y_i) &= \phi V(\mu_i) = \frac{1}{\beta}\mu_i \\ \log(\mu_i) &= \gamma_0 + \gamma_1 x_i \end{aligned}$$

Note that this is no longer a basic generalized linear model, as gamma densities with parameters α_i and constant β can no longer be written in the form of an exponential dispersion family. This is however, of no great consequence unless one is dependent on software packages that only will deal with certain classes of models, such as glms.

1.3.3 An Extreme Value Model

Consider, again, the decision to eliminate negative values of soil respiration from the data before developing a regression model. The whole point of such a modeling exercise is to arrive at a model that captures the *distributions* of responses (soil respiration) as soil temperatures vary. It is, after all, well accepted that respiration will increase with increasing temperature, so determining that there is a positive relation between these variables is of no consequence; actually, this is well accepted for temperatures up to *some point* after which it is not clear that respiration continues to increase. We might want to see if we can determine a model that basically captures the

same features of the data as the gamma model developed previously, but that can also accommodate negative responses. One response distribution that suggests itself is that of a right-skew extreme value random variable. This extreme value distribution is a location-scale family, so an additive error model formulation would be natural. At the same time, we also would like to model response variances as proportional to expected values. While this is not easily achieved, we can do something quite similar.

The right-skew version of the extreme value distribution has probability density function, for parameters $-\infty < \xi < \infty$ and $\theta > 0$,

$$f(y|\xi,\theta) = \frac{1}{\theta}\,\exp\left(-\left\{\frac{y-\xi}{\theta}\right\}\right)\,\exp\left[-\exp\left(-\left\{\frac{y-\xi}{\theta}\right\}\right)\right];\quad -\infty < y < \infty. \tag{1.28}$$

The density (1.28) defines a location-scale family of distributions in which the location parameter ξ is equal to the *mode* of the distribution (rather than the expected value) and the variance is given by $(\pi^2/6)\theta^2$. We could then formulate a 'power of the mode' model as,

$$Y_i = \xi_i + \sigma\,\xi_i^{\phi}\,\epsilon_i, \tag{1.29}$$

where

$$\log(\xi_i) = \beta_0 + \beta_1 x_i,$$

and the ϵ_i are assumed to be independent and identically distributed with densities

$$f(\epsilon) = \exp(-\epsilon)\,\exp[-\exp(-\epsilon)];\quad -\infty < \epsilon < \infty.$$

Chapter 2

Models With Latent Variables

While all statistical models contain random variables associated with observable effects, some models may also incorporate random variables that represent "unobservable" effects in a model. The title of *latent variables* is, in this context, a broad concept. In many applications connected with the social sciences, the term latent variable model implies what are known as "structural equations" models. While structural equation models certainly contain latent variables, latent variables occur in many other contexts as well. We will use the term latent variable to refer to any portion of a model that corresponds to a phenomenon (or a collection of phenomena) that cannot be measured or observed. There are few guiding principles as to how such models are formulated. Thus, our presentation will consist of examples intended to illustrate the flexibility of such models.

2.1 Limiting Factors in Ecology

Quantities in nature are sometimes related in the form of regressions, but other times do not seem to be related in this simple manner, presumably because nearly all mechanisms that produce bivariate patterns depend on more factors than only the two under consideration. And, multiple regressions are often (usually) incapable of adequately modeling the manner in which highly complex sets of interacting variables express themselves. We present here two examples of situations in which a response of interest is related to an environmental covariate, but not in a way that allows traditional regression to be used for analysis.

Example 2.1 – Aquatic Primary Productivity

A important question for limnologists, who study the chemistry and quality of bodies of water, is what factors control the level of primary productivity in those waters. Primary productivity is largely a matter of how many algae are present in the photic zone and is typically assessed by measuring the concentration of chlorophyll in the water. If primary productivity is too low, a lake or reservoir will fail to support a healthy ecological community and may be essentially barren of life. If primary productivity is too high, eutrophication is accelerated, anoxic zones can occur (e.g., the Gulf of Mexico) and water drawn from the lake or reservoir may be rendered useless for its intended purpose such as irrigation or municipal drinking water. In the 1960s several published regressions of chlorophyll on the concentration of phosphorus (actually log chlorophyll on log phosphorus) resulted in the opinion that in the temperate zone phosphorus was the primary regulator of algal growth. International treaties were signed by the United States and Canada to lower

levels of phosphorus entering Lake Erie. For about a decade nearly every limnological study conducted in the United States recorded chlorophyll and phosphorus concentrations. Sometimes there were reasonable regressions, many times there were not. Scatterplots of chlorophyll on phosphorus most often resulted in the shape of a wedge or a right triangle filled in with observed points.

Example 2.2 – Abundance of *Microcystin*

There are certain types of algae that produce toxins. Ingestion of these toxins can lead to adverse health effects, including death in rare instances (and yes, there have been deaths in the US attributed to this cause, as well as elsewhere in the world). In North America, the most predominant such toxin-producing algae is a genus called *Microcystin*. It is of interest to public health officials to determine factors that may be related to the concentration of these algae in lakes and reservoirs. A study was conducted in which the concentration of *Microcystin spp.* and various water chemistry variables were measured. Exploratory analyses suggested that a model of *Microcystin* concentration versus the nitrogen concentration of waterbodies might be useful in describing the situation in the Midwestern US, and could potentially lead to prediction of possible problem waters. But little is known about the manner in which various water chemistry variables may be related to *Microcystin* abundance (i.e., concentration) in lakes and reservoirs, and it is clear that there is not a simple regression-like relation between nitrogen and *Microcystin*. Scatterplots of Microcystin concentration against nitrogen concentration look like unimodal curves but, as with chlorophyll-phosphorus in Example 2.1, having observed points filling in under the curve, not following

it.

2.1.1 Ecological Basis for Model Development

What the problems of Examples 2.1 and 2.2 share in common are data that look like they fall below some type of systematic relation between responses and covariates, not along such a relation. We might question whether there is any aspect of ecological theory that would indicate why this might be, and there is. There is a very basic ecological concept known as the *Law of Limiting Factors* that explains why such data patterns might arise, and that might be useful in formulating a model for these situations. The fundamental ideas of limiting factors are captured in the two portions of this theory, called *Leibig's Law of the Minimum* and *Shelford's Law of Tolerance*. Very briefly (and in reduced technical form) these two laws are as follows:

1. Leibig's Law of the Minimum.

   When a biological process (such as algal growth) depends on a number of necessary inputs (such as nutrients) that are consumed during the process, the process is limited or stopped by the input factor that is used up the most quickly. For example, if the growth of a plant depends on the primary plant nutrients of phosphorus and nitrogen, radiant energy (as sunlight), and carbon, whichever of these factors are in shortest supply will stop the growth process when it runs out. This concept is employed, by the way, in agriculture when a farmer decides what type of fertilizer to apply to a crop, and was the basis for interpreting the 1960s regressions of log chlorophyll on log phosphorus – phosphorus was limiting algal growth. But there may not be one common active limiting factor in all situations. In one lake at one time

phosphorus may be limiting, but in another lake or in the same lake at a different time, something else may be limiting. A regression line should really represent the potential amount of chlorophyll for a given level of phosphorus (a maximum) rather than the expected value.

2. Shelford's Law of Tolerance.

   The ecological fitness of an organism is often reflected in the abundance of that organism (e.g., species or genus). The environment in which organisms exist consist largely of a set of environmental gradients such as altitude, temperature, salinity, oxygen availability, etc. Along a given gradient, a type of organism will have a range of tolerance outside of which it cannot exist. But, even within that range, there will be a preferred level along the gradient at which the organism is most "comfortable". This is often represented as a unimodal curve for which the vertical axis is abundance and the horizontal axis is the environmental gradient. The mode of the curve is interpreted as the optimal level of the gradient for the given organism. But such a unimodal curve only exists along one environmental gradient without controlling for other potentially important factors. Thus, such a curve should represent an optimal level at any point along the gradient, that is, a potential if all other factors are at their most favorable. When this is not true, which is probably much of the time, observations should be scattered below the curve.

2.1.2 Developing Statistical Models

How do we reflect the ideas of limiting factors in statistical models? Leibig's law of the minimum is reflected, for a single factor, in the type of model

introduced by Kaiser, Speckman and Jones (1994), a simple version of which has the form

$$Y_i = x_i \gamma \, U_i + \sigma \, \epsilon_i, \tag{2.1}$$

in which Y_i is the response variable (e.g., chlorophyll), x_i is the potential limiting factor of concern (e.g., phosphorus), and U_i is an unobservable (i.e., latent) random variable with possible values on the interval (0, 1). If $U_i < 1$, then some factor other than x_i must be limiting for the observation connected with the response Y_i. Kaiser, Speckman and Jones (1994) took $U_i \sim iid$ beta (α, β) and $\epsilon_i \sim iid \, N(0, 1)$ for $i = 1, \ldots, n$, and U_i independent of ϵ_i. This model leads to an expected data pattern of a triangular array of data scattered below a straight line through the origin, as illustrated by a simulated data set in Figure 2.1. These data were simulated using a beta $(2, 3)$ distribution for the latent U_i variables, which gives an expected value of 0.40 times the line $x_i \, \gamma$. There are a host of questions regarding response function form that we are brushing aside at the moment. The point for us here is that a latent random variable U_i has been used to model departures from the mechanism of interest. Various data patterns can be generated by this model, depending on the values of the beta parameters and σ that are used. For analyses of actual data sets containing chlorophyll and phosphorus see Kaiser, Speckman and Jones (1994).

Now, combine this modeling idea with Shelford's law of tolerance. Consider a unimodal curve that describes the tolerance or affinity of an organism for various levels of a single environmental gradient (this is Shelford's law of tolerance). Certainly, the abundance of that organism depends on a number of factors besides the one gradient in question, and abundance may not reach the potential level it could for a given level of the gradient. The inescapable

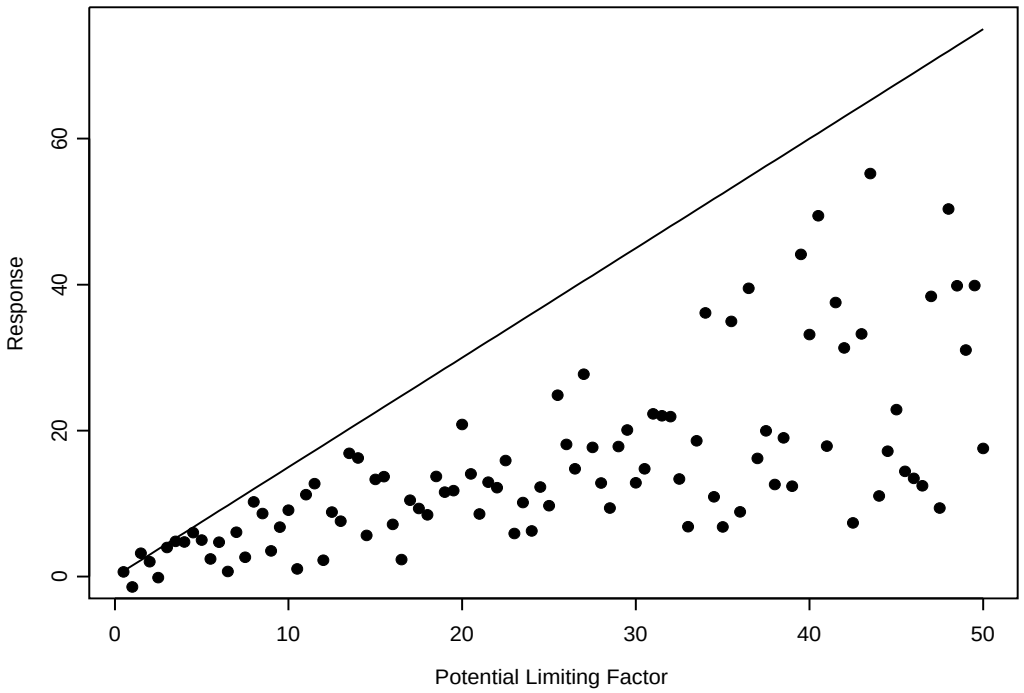

Figure 2.1: Simulated data showing model for limiting factors based on Leibig's law of the minimum.

conclusion is that the entire unimodal curve in question represents an optimum, given that all other factors are at their most favorable levels. Thus, in observed data what should we expect? We should expect observed values of the abundance of an organism to be scattered below a unimodal curve, because it is not always true that all other factors are at their most favorable levels.

What is needed to formulate a model for this situation? Essentially, replacing the linear limit function $x_i\,\gamma$ of model (2.1) with a unimodal curve. There are any number of possible choices for unimodal curves, many of which have restrictions on shape other than unimodality (e.g., a normal pdf is a unimodal curve, but must always be symmetric). One flexible possibility is the function

$$f(x_i,\,\boldsymbol{\theta}) = \frac{\theta_1}{\Gamma(\theta_2 + \theta_3 x_i + \theta_4 x_i^2)}. \tag{2.2}$$

The function in expression (2.2) is quite flexible, with fairly nice connections between its behavior and the values of θ_1, θ_2, θ_3 and θ_4; θ_1 governs height, θ_2 governs whether both or only one tail is seen, θ_3 governs rate of increase, and θ_4 governs rate of decrease. A model for the situation we are trying to capture may then be formulated as,

$$Y_i = f(x_i,\,\boldsymbol{\theta})U_i + \sigma\epsilon_i, \tag{2.3}$$

where $f(\cdot)$ is given in (2.2), $U_i \sim iid\, G$ for some distribution on the interval $(0,\,1)$, and $\epsilon \sim iid\, F_\epsilon$ for a location-scale family F_ϵ with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = 1$.

Model (2.3) has been applied to data on Microcystin abundance (Y_i) with the covariate or potential limiting factor of nitrogen concentration (x_i). In this application, F_ϵ was taken to be logistic, and G was a histogram model formed by dividing the unit interval into a partition (e.g., 0 to 0.25, 0.25

to 0.5, 0.5 to 0.75, and 0.75 to 1.0). A simulated data set is presented in Figure 2.2, along with the true limit function $f(x_i, \boldsymbol{\theta})$ (as the solid curve) and an estimated limit function based on maximum likelihood estimates of the components of $\boldsymbol{\theta}$ (as the dashed curve).

To fully specify models (2.1) or (2.3) we need to write down forms of the various distributions involved so that we can derive the marginal distribution of the Y_i given parameters involved in the distributions of the U_i and ϵ_i.

Assuming continuous Y_i and continuous U_i, the general form implied by model (2.3) is as follows. First, for given U_i, the conditional density of Y_i is a location-scale transformation of F_ϵ,

$$f(y_i|u_i, \boldsymbol{\theta}, \sigma) = \frac{1}{\sigma} f_\epsilon \left(\frac{y_i - f(x_i, \boldsymbol{\theta})\, u_i}{\sigma} \right). \tag{2.4}$$

The marginal distribution of the U_i are (iid) with pdf $g(u_i, \boldsymbol{\eta})$ which may depend on the parameter (vector) $\boldsymbol{\eta}$. Then the joint of Y_i and U_i is,

$$p(y_i,\, u_i|\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) = f(y_i|u_i, \boldsymbol{\theta}, \sigma)\, g(u_i, \boldsymbol{\eta}), \tag{2.5}$$

and the marginal of Y_i is given by the mixture distribution,

$$h(y_i|\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) = \int f(y_i|u_i, \boldsymbol{\theta}, \sigma)\, g(u_i, \boldsymbol{\eta})\, du_i. \tag{2.6}$$

For model (2.1), which is a special case of (2.3), $f(x_i, \boldsymbol{\theta}) = x_i\gamma$, $\boldsymbol{\theta} \equiv \gamma$, $f_\epsilon(\cdot)$ is standard normal and $g(\cdot)$ is beta with $\boldsymbol{\eta} \equiv (\alpha, \beta)$ so that

$$p(y_i|\gamma, \sigma, \alpha, \beta) =$$

$$\frac{\Gamma(\alpha + \beta)}{(2\pi\sigma^2)^{1/2}\Gamma(\alpha)\Gamma(\beta)}$$

$$\times \quad \int_0^1 \exp\left\{ \frac{-1}{2\sigma^2}(y_i - \gamma x_i u_i)^2 \right\} u_i^{\alpha-1}(1 - u_i)^{\beta-1}\, d\, u_i.$$

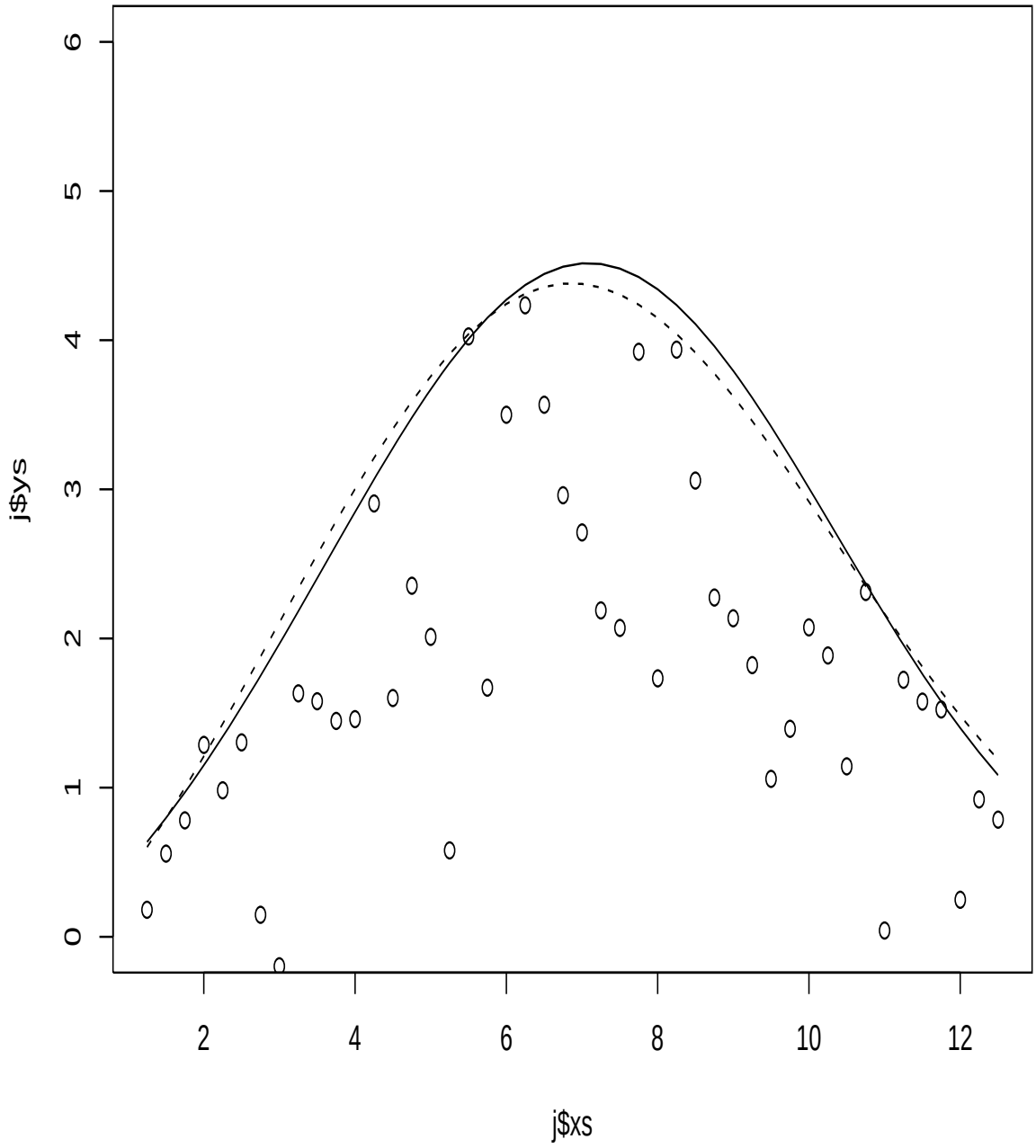

Figure 2.2: Data simulated from model (2.3) with true limit function given as the solid curve and estimated limit function as the dashed curve.

In the application to Microcystin data, we defined $\lambda_j \equiv 0.25\, j$, for $j = 0, 1, 2, 3, 4$, defined $\boldsymbol{\eta} \equiv (\eta_1, \eta_2, \eta_3, \eta_4)$, and took

$$g(u_i|\boldsymbol{\eta}) = \eta_j\, I(\lambda_{j-1} < u_i < \lambda_j); \quad j = 1, 2, 3, 4.$$

Note this is a histogram model with probabilities $0.25\eta_j$ so that we have imposed the restriction

$$\sum_{j=1}^{4} \eta_j = 1.$$

For the moment, leave $f_\epsilon(\cdot)$ unspecified. Then (2.5) becomes, for $j = 1, 2, 3, 4$,

$$p(y_i,\, u_i|\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) = f(y_i|u_i, \boldsymbol{\theta}, \sigma)\eta_j\, I(\lambda_{j-1} < u_i < \lambda_j),$$

which leads to the mixture distribution of (2.6) as,

$$h(y_i|\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) = \sum_{j=1}^{4} \eta_j \int_{\lambda_{j-1}}^{\lambda_j} f(y_i|u_i, \boldsymbol{\theta}, \sigma). \tag{2.7}$$

Now, let

$$w_i \equiv \frac{1}{\sigma}\, \{y_i - f(x_i, \boldsymbol{\theta})u_i\}\,,$$

or,

$$u_i = \frac{1}{f(x_i, \boldsymbol{\theta})}(y_i - \sigma w_i); \quad \frac{d\, u_i}{d\, w_i} = \frac{-\sigma}{f(x_i, \boldsymbol{\theta})}.$$

Then from (2.4) and (2.7),

$$\begin{aligned} h(y_i|\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) &= \sum_{j=1} 4\eta_j \int_{\xi_{j-1}}^{\xi_j} \frac{f_\epsilon(w_i)}{f(x_i, \boldsymbol{\theta})}\, d\, w_i \\ &= \sum_{j=1}^{4} \frac{\eta_j}{f(x_i, \boldsymbol{\theta})}\, [F_\epsilon(\xi_j) - F_\epsilon(\xi_{j-1})]\,, \end{aligned} \tag{2.8}$$

where

$$\xi_j \equiv \frac{1}{\sigma}\, \{y_i - f(x_i, \boldsymbol{\theta})\lambda_{j-1}\}\,; \quad \xi_{j-1} \equiv \frac{1}{\sigma}\, \{y_i - f(x_i, \boldsymbol{\theta})\lambda_j\}$$

The fact that ξ_j is a function of λ_{j-1} and ξ_{j-1} is a function of λ_j comes from $(d\,u_i/d\,w_i) < 0$. In this particular example, F_ϵ was taken to be a logistic distribution, $F_\epsilon(x) = (1 + \exp(x))^-1$.

Now, all of this effort has been expended to render the mixture (2.8) in a form that does not involve an unevaluated integral as is present in the mixture formulated from model (2.1). It would, of course, have been possible to use the same type of distributions for both U_i and ϵ_i in both models.
In any case, returning to the general notation of (2.4) through (2.6), the log likelihood for a set of observed $y_1, \ldots, y_n$ is,

$$L(\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}) = \sum_{i=1}^{n} h(y_i|\boldsymbol{\theta}, \sigma, \boldsymbol{\eta}). \tag{2.9}$$

An application of model (2.3) with $f(x_i, \boldsymbol{\theta})$ as in (2.2) and the resulting mixture in the form of (2.8) to the actual Microcystin data resulted in estimates presented in Figure 2.3. In this figure, the unimodal "tolerance" curve (2.2), evaluated at maximum likelihood estimates of the components of $\boldsymbol{\theta}$ is shown as the solid line, with point-wise 90% interval estimates given as dashed lines. Actual inference in this problem depends, of course, not only on the estimated value of $\boldsymbol{\theta}$ but also σ and, importantly, $\boldsymbol{\eta}$; these latter values determine the probability that responses reach various proportions of the tolerance curve.

2.2 Zero-Inflated Models

What are known as zero-inflated models have application in situations for which one might naturally consider using a binomial or a Poisson model, but in which the observed frequency of zeros in data are in excess of what would

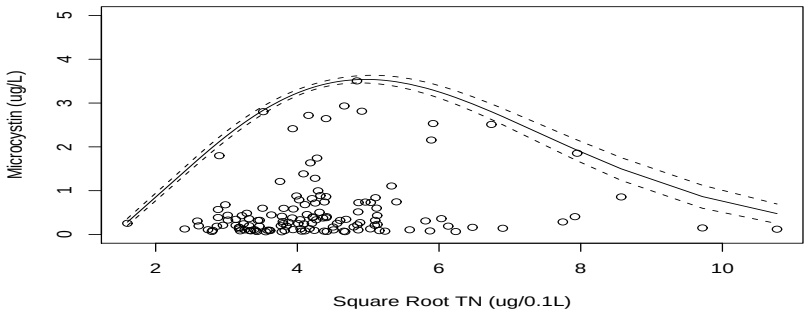

Figure 2.3: Actual data on abundance of Microcystin as a function of nitrogen concentration in midwestern lakes and reservoirs.

reasonably be expected in a model with positive mean and the specified distribution. Some statisticians are perfectly happy to simply specify a data model with an inflated probability of zero in such cases, just as they are to use a gamma-Poisson mixture and call it a negative binomial distribution. I greatly prefer if a reasonable conceptualization is available that can represent a (at least hypothetical) mechanism that could lead to increased frequency of zero values, just as I prefer a gamma-Poisson to be thought of as a hierarchical model in which the relative frequencies with which a mechanism (represented in a given situation by a Poisson parameter) manifests itself are reflected in the gamma mixing distribution. Several examples will illustrate.

Example 2.3 – Sales of Green Beans

Retail grocers are aware that effects of many factors determine the quantities of various goods that they will sell. Untangling these factors is a major problem that most grocers approach through simple analyses of available data. A part of the reason for this approach is that formulation of appropriate statistical models is difficult. The overall goal in analysis of the data of this example was to develop a model to relate number of units of green beans to price for 179 stores in a grocery chain in the Midwestern United States. Units were not provided with the data, but from the values in the data set one might guess that units were cans and prices were in dollars. Here, we will only consider the problem of selecting a suitable random model component, that is, selecting a distribution for number of units sold at given prices. Empirical probability functions for sales at particular prices, and in particular stores, are presented in Figure 2.4 as examples (with 179 stores and roughly $10-12$ prices at each store there are a large number of such plots that would be

possible). The most notable feature of these plots is the exaggerated relative frequency for values of 0 units sold. Aside from this elevated frequency, the distributions might be described by a unimodal right skewed distribution for discrete non-negative random variables. Although evidence has not been presented for other levels of price, it appears that at least for the realizations of Figure 2.4, a reasonable model might be a "zero inflated Poisson," which results from mixing a Poisson distribution with a binary distribution. A constructive development of this distribution is to define random variables $\{Y_i : i = 1, \ldots, n\}$ associated with the quantity of number of units of green beans sold on day i in an individual store at a fixed price level (i.e., i indexes days for which the store was selling green bean units for a fixed price x, say). Also define random variables $\{Z_i : i = 1, \ldots, n\}$ associated with the unobservable construct of we might think of as consumer interest in green beans. Take the Z_i to be independent and identically distributed with a binary probability mass function having parameter p. If $Z_i = 0$ then $Pr(Y_i = 0|Z_i = 0) = 1$. Take the Y_i to be independent with identical conditional distributions given $Z_i = 1$ that are Poisson with parameter λ. Note that Y_i can assume a value of 0 even if $Z_i = 1$.

The marginal distribution of Y_i, has probability mass function, for $\lambda > 0$ and $0 < p < 1$,

$$f(y|p, \lambda) = \begin{cases} (1-p) + p\exp(-\lambda) & y = 0 \\ \frac{1}{y!} p\lambda^y \exp(-\lambda) & y = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

The probability mass function (2.10) is often called a zero-inflated Poisson (or ZIP), and it could be used as the basis for the random component in a model to relate sales to price. As mentioned previously, some statisticians would

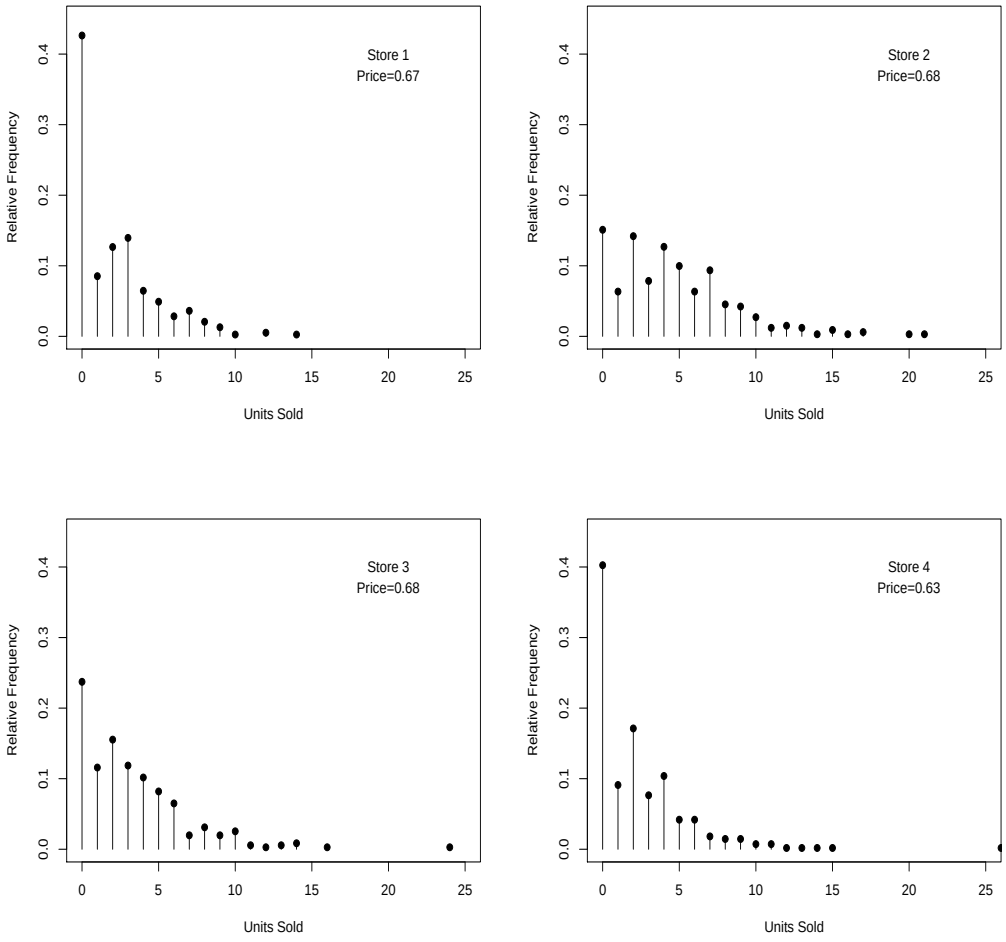

Figure 2.4: Empirical probability functions for sales at particular prices ranging from 0.63 to 0.68 for four individual stores.

be happy to examine exploratory plots such as those in Figure 2.4, make note of an elevated frequency for zero values, and conclude that a ZIP model might be appropriate to model the response variables $Y_{i,j}$ of units of green beans sold on day i at price x_j. Have we gained anything by embedding this problem in the conceptual mechanism of consumer interest and sales given interest, instead of just noting that we need a model with a greater frequency of zero values than can be handled by a Poisson? I believe we have. If we interpret the parameter p as the probability of consumer interest, we could model both the $p_{i,j}$ and the $E(Y_{i,j})$ as functions of price or we could model $p_{i,j}$ as functions of time of year, or other covariates such as an index of advertisement level.

Example 2.4 – Monitoring Habitat Occupancy

Natural resource agencies expend considerable time and resources on monitoring programs. Royle and Dorazio (2008, p. 137) report data from a survey of fisher (*Martes pennanti* abundance in California. A fisher is a medium-sized carnivore in the weasel family that lives in forests and eats everything from insects to rabbits and porcupines. A total of 464 sampling locations were visited 8 times each, and the detection of fisher sign (see Royle and Dorazio for a complete explanation) recorded as a binary variable at each visit. The number of times fisher were detected are given in Table 2.1 (fisher were detected zero times at 400 locations, once at 16 locations, etc.).

Times Detected	0	1	2	3	4	5	6	7	8
No. of Locations	400	16	12	12	5	10	3	4	2

Table 2.1: Frequency of detection of fisher at sampling locations in California.

A simple model for this problem is to define random variables Y_i associated with the number of times fishers were detected at location $i = 1, \ldots, L$, assume independence and identical distribution, and assign the Y_i a binomial probability mass function with parameter θ and binomial sample sizes n_i. For the data of Table 2.1, $L = 464$ and $n_i = n = 8$ for all i. We could claim to construct this model from sets of independent and identically distributed binary trials, but the assumption of identical distribution for the binary variables could be questioned. We certainly can, however, simply assign binomial distributions without construction. We would then have the log likelihood, modulo an additve constant,

$$\ell(\theta) = \sum_{i=1}^{L} \left[y_i \log(\theta) + (n_i - y_i) \log(1 - \theta) \right]. \tag{2.11}$$

Maximizing (2.11) in θ gives the estimate

$$\hat{\theta} = \frac{\sum_{i=1}^{L} y_i}{\sum_{i=1}^{L} n_i},$$

which, for the data of Table 2.1 yields $\hat{\theta} = 0.05603448$, from which our conclusion would be that the probability of detecting fisher use of an area is small. If we look at the values of Table 2.1 we may be struck by the rather large number of locations at which fishers were never detected (400) and we might wonder whether zero-inflated binomial distributions would be more appropriate than our model of independent and identically distributed binomials. We might compute a Chi-square test statistic for our fitted simple binomial model, which would result in a value of $T = 9.1 \times 10^7$. It would probably not be a good idea to determine the associated $p-$value because

any number of expected frequencies are quite small under the fitted model. If we wanted to determine a $p-$value for this statistic we could do so using a parametric bootstrap or Monte Carlo procedure. It would seem clear, however, that our model allows a great deal of room for improvement. If we look at the individual contributions to the Chi-square statistic, however, we find that the contribution for 0 detections is 39.1 while the contributions from 6 detections is 2.5×10^4 and that from 8 detections is 8.9×10^7. The lack of fit occurs at the larger values. So it is not the zero frequency that is the problem. Or is it? The large frequency of zero values forces the estimate from this simple model to be small, resulting in the fitted model reflecting the zero frequency perhaps better than the frequencies for larger values.

Even if the simple binomial model had proven able to reflect the observed data, there would remain a problem in the interpretation of that model. The single parameter θ can reflect only the probability that fisher use is detected, *assuming that fishers are present in all locations*. The large number of zero values in the data may be due to the fact that it is unlikely that fishers are present (occupying) at all of the 464 locations. And, if they are not using an area it will be impossible to detect them there. This suggests a strategy similar to what was used for green bean sales in Example 2.3. Define *occupancy* indicators as a binary random variables $\{Z_i : i = 1, \ldots, L\}$ such that $Z_i = 1$ if location i is occupied (or used) by fishers and $Z_i = 0$ if location i is unoccupied by fishers. The Z_i are a type of latent variable. While we can observe use of an area if fisher sign is found, the lack of positive evidence for use does not guarantee that an area is not being used. Once again, we will assign the Z_i independent and identical binary distributions with parameter p, take $Pr(Y_i = 0|Z_i = 0) = 1$ and, conditional on $Z_i = 1$ assign the Y_i binomial distributions with parameter θ. The resulting marginal probability

mass function for Y_i is

$$f(y|p,\theta) = \begin{cases} (1-p) + p(1-\theta)^{n_i} & y = 0 \\ p\frac{n_i!}{y!\,(n_i-y)!}\,\theta^y(1-\theta)^{n_i-y} & y = 1, 2, \ldots, n_i \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

Maximizing the log likelihood formed from these probability mass functions results in maximum likelihood estimates of $\hat{p} = 0.14$ and $\hat{\theta} = 0.40$. Expected frequencies from this fitted model produce a Chi-square statistic of $T = 35.5127$, which is quite a decrease from the simple binomial model, although the associated $p-$vlaue is still small at 2.16×10^{-5}. The observed frequencies of Table 2.1 along with expected frequencies (rounded to the nearest whole number) under the simple binomial and zero-inflated binomial models are given in Table 2.2.

Times Detected	0	1	2	3	4	5	6	7	8
Observed	400	16	12	12	5	10	3	4	2
Simple Binomial	293	139	29	3	0	0	0	0	0
Zero-Inflated Binomial	400	6	14	18	15	8	3	1	0

Table 2.2: Frequency of detection of fishers and expected values under two models.

Note that the expected frequencies for the zero-inflated model sum to 465 (rather than 464) due to rounding to whole numbers. Neither of our fitted models are entirely pleasing for representing these data, but the zero-inflated model is a vast improvement on the simple binomial model. if we were forced to choose between the two, the zero-inflated binomial seems quite a bit better. Importantly, the zero-inflated model again allows a rational scientific

conceptualization of the problem, rather than simply being a device to try to fix a problem with the original simple binomial model.

Chapter 3

Censored Observations and Truncated Variables

The need to formulate probability structures for censored observations or truncated variables can arise in a variety of ways. Two of the most common will be briefly described in this section. When continuous variables have been associated with observable quantities, but values beyond some point in the support (either to the right or to the left) are not observed exactly, those observations are censored. When we restrict the support of the distribution of random variables to avoid the possibility of modeling impossible realizations (e.g., negative values for something that can only be positive) we do so by truncating the distribution. There are a number of types of censoring that can occur and there are two main types of truncation that get used. We will discuss these in turn.

3.1 Censored Random Variables

Probably the most classic examples of censoring occurs in studies connected with reliability or survival analysis in which response random variables are associated with time to some event. Examples include time to failure for machine components or time to death for cancer patients. As already mentioned, there are a number of types of censoring that can occur. Divisions used to categorize types of censoring include Right, Left and Interval censoring, and Type I versus Type II censoring. The types of censoring that may arise in a study are largely determined by study design. The distinction between what are called Type I and Type II censoring is whether the censoring point is fixed in a study or varies across sampling units, perhaps due to some random mechanism. Unless we are trying to model a censoring mechanism along with a distribution of time-to-event, the difference between Type I and Type II censoring does not impact the analysis, and we will confine our discussion to right, left, and interval censoring.

3.1.1 Types of Censoring

1. Interval Censoring.

   Interval censoring occurs when a random variable Y_i is reported to have assumed value y_i, where all that is known is that the actual value of Y_i is in the interval $(y_i - \Delta_{1,i},\, y_i + \Delta_{2,i})$. Usually, $\Delta_{1,i} = \Delta_{2,i} = \Delta_i$ and often $\Delta_i = \Delta$ for all i in a set of random variables. In truth, all data are interval censored since even quantities modeled with continuous random variables are reported to some finite precision. Considering all data as interval censored is the easiest way to see that censoring does not necessarily have to occur with respect to variables that represent

time-to-event. Within that context, however, interval censoring can arise either from precision of data records or as a result of times of observation. Consider an acute toxicity test in which organisms are exposed to a potential toxicant, and time to death is to be recorded for each organism. The organisms may only be observed every 4 hours, or maybe even only once a day. This produces interval censoring as a result of study design or observational protocol.

2. Right Censoring.

   Right censoring is probably the most common type of censoring encountered in practice and occurs when some portion of a set of random variables are known only to have values greater than some *censoring times*. Consider a set of random variables $\{Y_i : i = 1, \ldots, n\}$ that are associated with time to some event for sampling units $i = 1, \ldots, n$. What is recorded is either the observation that Y_i has assumed the value y_i or the observation that $Y_i > c_i$ for some known values c_i; $i = 1, \ldots, n$. It is assumed that the censored observations can also be identified and we can let $\delta_i = 1$ if Y_i leads to a censored observation, and $\delta_i = 0$ if the observation of Y_i is not censored. What is recorded in the available data then are pairs of values $\{(min\{y_i, c_i\}, \delta_i) : i = 1, \ldots, n\}$. There is really no standard convention for reporting censored data, however, and one needs to exercise caution to determine what is contained in a given set of data. For example, some studies might be reported in the format $\{(min\{y_i, c_i\}, \delta_i) : i = 1, \ldots, n\}$ but with $\delta_i = 0$ indicating censoring (rather than $\delta_i = 1$).

   The censoring times may all be the same, which is the case in a study that is started at some fixed time that can be set to 0, and observation

is ended at some other fixed time T. This might occur, for example, if n automobile batteries are all connected to a device that produces constant discharge, and the time until battery death is recorded, but only up to 48 hours at which time the apparatus is dismantled. In other studies, the censoring times might vary, such as when patients drop out of a clinical trial.

3. Left Censoring.

   Although less common that right censoring, left censoring occurs when some portion of a set of random variables are known only to have values less than some set of censoring times. This is just the mirror image of right censoring, and data that are recorded have the form of pairs $\{(max\{y_i, c_i\}, \delta_i) : i = 1, \ldots, n\}$, where the δ_i are again indicators for observations that are censored. The most common situation I know of in which left censoring occurs is in chemical analysis in which there is a *limit of detection* below which the equipment being used cannot detect a compound. This occurs, for example, in the analysis of environmental samples (e.g., soil or water) for toxic substances such as PCBs (polychlorinated biphenyls) or PAHs (polycyclic aromatic hydrocarbons).

3.1.2 A Major Assumption

One way to think about censored observations is that they constitute missing values or involve missing information, and a crucial assumption about problems involving censored data is essentially the same as a crucial assumption about problems involving missing data. That assumption is that the censoring mechanism is independent of the mechanism producing observed values, which is similar to an assumption of missing at random or, in a situation

lacking covariates, missing completely at random.

The importance of this assumption is obvious if one is attempting to construct a model to represent a mechanism of time-to-event. Consider, for example, a clinical trial for a new treatment for althlete's foot (a fungal infection that usually occurs on the bottom of the feet and between the toes, but can spread to other parts of the body). The observable response might be time to complete cure. If people drop out of the trial because they believe the treatment has not been successful for them, and go back to using their old remedy, then the results of the study have been compromised (at least in terms of the type of analysis discussed here).

3.1.3 Likelihoods for Censored Variables

Regardless of the particular model used or whether estimation and inference are to proceed using likelihood or Bayesian techniques, we need the likelihood for any model containing censored observations. We will assume that the assumption of independence between censoring and time-to-event mechanisms is justified and that the observation of sampling units justifies an assumption of independence among random variables that are associated with observations on those units. Consider, then, a setting in which we have continuous random variables $\{Y_i : i = 1, \ldots, n\}$ associated with times to some event for sampling units, and we are willing to model these variables as independent and identically distributed with common probability density function $f(y_i|\boldsymbol{\theta})$ having support Ω and some unknown parameter $\boldsymbol{\theta}$. Recorded observations of these random variables may be interval, right, or left censored Our objective is to determine a likelihood for estimation of $\boldsymbol{\theta}$ using either likelihood or Bayesian methods.

It is most convenient here to develop likelihoods based entirely on probabilities, rather than densities, and this can be motivated by the recognition that all recorded data are actually censored in some manner, as mentioned previously. Let $\mathcal{I} = \{i : \text{ observation } i \text{ is interval censored}\}$. Let $\mathcal{R} = \{i : \text{ observation } i \text{ is right censored}\}$. Let $\mathcal{L} = \{i : \text{ observation } i \text{ is left censored}\}$. Rather than using the terms *right censoring time* and *left censoring time* we will consider *lower possible value* and *upper possible value* for each observation. Let $\{c_{i,L} : i = 1, \ldots, n\}$ and $\{c_{i,U} : i = 1, \ldots, n\}$ denote these values, respectively. Then each observation, regardless of whether it is left, right, or interval censored, must be such that the corresponding random variable satisfies $c_{i,L} < Y_i < c_{i,U}$. If an observation is left censored, we take $c_{i,L} = -\infty$ and if an observation is right censored we take $c_{i,U} = \infty$. In a data set for which we will consider all observations censored, we now only need values for $c_{i,L}$, $c_{i,U}$ and censoring type (with values $\mathcal{L}$, $\mathcal{R}$, or $\mathcal{I}$).

The likelihood for $\boldsymbol{\theta}$ based on observations $\{y_i : i = 1, \ldots, n\}$ may now be written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{n} Pr(c_{i,L} \leq Y_i \leq c_{i,U}). \tag{3.1}$$

In (3.1),

$$Pr(c_{i,L} \leq Y_i \leq c_{i,U}) = \begin{cases} Pr(-\infty \leq Y_i \leq c_{i,U}) = F(c_{i,U}|\boldsymbol{\theta}) & \text{if } i \in \mathcal{L} \\ Pr(c_{i,L} \leq Y_i \leq \infty) = 1 - F(c_{i,L}|\boldsymbol{\theta}) & \text{if } i \in \mathcal{R} \\ Pr(c_{i,L} \leq Y_i \leq c_{i,U}) = F(c_{i,U}|\boldsymbol{\theta}) - F(c_{i,L}|\boldsymbol{\theta}) & \text{if } i \in \mathcal{I}. \end{cases} \tag{3.2}$$

Often, observations that are not either right or left censored will be taken as exact values rather than interval censored, as we usually do for observations corresponding to continuous random variables when we are not considering censoring. The likelihood will now be written as a combination

of probabilities (for censored observations) and densities (for exact observations).

$$L(\boldsymbol{\theta}) = \prod_{i \in \mathcal{L}} F(c_{i,U}|\boldsymbol{\theta}) \prod_{i \in \mathcal{R}} [1 - F(c_{i,L}|\boldsymbol{\theta})] \prod_{i \in \mathcal{I}} f(y_i|\theta) \tag{3.3}$$

This amounts, essentially, to replacing $F(c_{i,U}|\boldsymbol{\theta}) - F(c_{i,L}|\boldsymbol{\theta})$ in the last line of (3.2) with the density value $f(y_i|\boldsymbol{\theta})$. If all data are recorded to the same precision, then the intermediate value theorem from integral calculus holds, and

$$\begin{aligned} F(c_{i,U}|\boldsymbol{\theta}) - F(c_{i,L}|\boldsymbol{\theta}) &= F(y_i + \Delta|\boldsymbol{\theta}) - F(y_i - \Delta|\boldsymbol{\theta}) \\ &= \int_{y_i - \Delta}^{y_i + \Delta} f(t|\boldsymbol{\theta})\, dt \\ &\approx 2\Delta f(y_i|\boldsymbol{\theta}). \end{aligned}$$

The result is that the likelihood obtained by using density values for observations assumed to be exact is proportional to the likelihood given in (3.1) and (3.2) and thus has the same shape and mode.

Example 3.1 – Remission in Leukemia Patients

Lawless (1982, p. 136) presents data on length of a remission period in leukemia patients given a particular therapy. After therapy is given there is a period of remission for each patient that we will model as random. In particular, let $\{Y_i : i = 1, \ldots, n\}$ be random variables associated with the time to the end of the remission period for n patients. Of course, in a study like this not all patients were given the therapy at the same time, so when the study ended some patients had been under observation longer than some other patients. In addition, at the time data were collected some patients were still in remission and are thus right censored. The data were recorded in units of week. Here, there are only two censoring classes, $\mathcal{I}$ for patients who

came out of remission during the study and $\mathcal{R}$ for patients still in remission at the end of the study. For $Y_i \in \mathcal{I}$ we will take $c_{i,L} = y_i - 1/2$ and $c_{i,U} = y_i + 1/2$. For $Y_i \in \mathcal{R}$ the recorded value is $c_{i,L}$. The data are given in Table 3.1 in which patient index has been ordered with increasing recorded value. In a set of course notes for Statistics 415, Meeker (2003) fits both exponential and lognormal models to these data.

Patient	1	2	3	4	5	6	7	8	9	10
Recorded Value	1	1	2	2	2	6	6	6	7	8
$c_{i,L}$	0.5	0.5	1.5	1.5	1.5	5.5	5.5	5.5	6.5	7.5
$c_{i,U}$	1.5	1.5	2.5	2.5	2.5	6.5	6.5	6.5	7.5	8.5
Patient	11	12	13	14	15	16	17	18	19	20
Recorded Value	9	9	10	12	13	14	18	19	24	26
$c_{i,L}$	8.5	8.5	9.5	11.5	12.5	13.5	17.5	18.5	23.5	25.5
$c_{i,U}$	9.5	9.5	10.5	12.5	13.5	14.5	18.5	19.5	24.5	26.5
Patient	21	22	23	24	25	26	27	28	29	30
Recorded Value	29	31	42	45	50	57	60	71	85	91
$c_{i,L}$	28.5	31	41.5	45	50	56.5	59.5	71	85	90.5
$c_{i,U}$	29.5	∞	42.5	∞	∞	57.5	60.5	∞	∞	91.5

Table 3.1: Data on leukemia remission.

3.2 Truncation

A careful reading of the previous section will show that censoring is really a property of the process of observation, not a specified statistical model *per se*. In contrast, what we usually mean by truncation relates to the specification

of distributions in models. There are two major types of truncation, *re-normalization* and *Winsorization*.

3.2.1 Re-Normalization

It may occur that, for whatever reasons, we would like to model certain random variables $\{Y_i : \imath = 1, \ldots, n\}$ with given distributions having densities $f_i(y|\boldsymbol{\theta})$ with common support $y \in \Omega$. The densities are indexed by i here to allow for nonidentical situations such as those involving covariates. But it may also occur that the support of these densities extends beyond the range of possible values for our random variables. A simple example of this would be in a problem for which we might like to use normal distributions for the error terms of an additive error regression, but the responses must be positive and there are observations near the origin. In such situations one option is to truncate the support of the assigned densities, and then re-normalize to maintain a proper model.

Consider a set of densities $\{f_i(y|\boldsymbol{\theta}) : i = 1, \ldots, n\}$ with common support Ω. Consider also a set of random variables $\{Y_i : i = 1, \ldots, n\}$ such that the possible values of these variables are contained in some connected subset $\Omega_0 \subset \Omega$. Almost always, Ω is some interval on the real line (including unbounded intervals in one or both directions) and Ω_0 is an interval that has upper and/or lower endpoints of smaller absolute value than those of Ω. Then we can assign the random variables distributions having densities

$$g_i(y|\boldsymbol{\theta}) = \frac{f_i(y|\boldsymbol{\theta})}{\int_{\Omega_0} f_i(t|\boldsymbol{\theta})\, dt} \tag{3.4}$$

with common support $y \in \Omega_0$.

Example 3.2

Consider a set of random variables $\{Y_i : i = 1, \ldots, n\}$ such that possible values are strictly positive but otherwise a simple linear regression might be appropriate. How might we model such a situation? One thought would be to take the error terms in a simple linear regression to have normal distributions truncated below at zero. This model would be, for $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where the ϵ_i will be taken as independent and identically distributed with probability density function

$$f(\epsilon_i|\sigma^2) = (1/2)\,(2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\epsilon_i^2\right); \;\; \epsilon_i > 0. \tag{3.5}$$

This might be quite a difficult model to interpret, since we have what is essentially a simple linear regression but with only positive "error terms". That is, deviations from the regression line are not deviations from expected values because the expected values of the Y_i are no longer given by the regression line (and the variances are no longer given by σ^2). A better option might be to forego an attempt to write the model as an expression for the Y_i (as Y_i = something), and simply assign these variables normal distributions truncated below at zero and having parameters $\beta_0 + \beta_1 x_i$ and σ^2. That is, assuming covariates $x_i > 0$ have a positive influence on the expected values of the responses, the model consists of assigning Y_i the densities, for $i = 1, \ldots, n$, $0 < \beta_0 < \infty$, $0 < \beta_1 < \infty$, and $\sigma^2 > 0$,

$$f(y_i|\beta_0, \beta_1, \sigma^2) = \left[1 - \Phi\left(-\frac{\beta_0 + \beta_1 x_i}{\sigma}\right)\right]^{-1} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right], \tag{3.6}$$

where $\Phi(\cdot)$ is the distribution function (cumulative density) of a standard normal random variable. Now model (3.6) does not have the expected values

and variances of the Y_i equal to $\beta_0 + \beta_1 x_i$ and σ^2 either, but they are close, especially for large values of $(\beta_0 + \beta_1 x_i)/\sigma$.

3.2.2 Winsorization

It is not uncommon for statisticians to use distributions with unbounded support to model quantities that definitely have values they cannot be larger (or perhaps smaller) than. The weight of a collection of hummingbirds might be modeled with either normal or gamma distributions, even though we know a hummingbird cannot weight 3 metric tons. Actually, even modelling the weight of elephants would face the same difficulty. We may model chemical concentrations using gamma, inverse gamma, or lognormal distributions that have support on the entire positive line, even though we know there is some physical limit that applies to the problem. Our presumption in formulating such models is that tail probabilities become negligible rapidly and that, past a certain point, the distribution has essentially no probability mass.

There are some cases in which the scenario of the previous paragraph applies but we wish to formulate a model that is more reflective of physical reality. It may be the case, for example, that we know the particular maximum (or minimum) value that is possible. In other cases we may not have a known maximum (minimum), but we can pick a value that is reasonably larger (smaller) than any physically possible maximum (minimum). Winsorization may sometimes be used to develop an appropriate model in these cases. Rather than re-scaling a distribution based on a fixed truncation point (which is re-normalization) we simple "stack up" all the probability past a fixed point at that point. In the case of discrete distributions this amounts to nothing more than adding up all the probability past a certain value and

assigning it directly to that value. In contrast, Winsorization of continuous distributions then leads to a mixed continuous/discrete density, in which there is positive probability that the variable assumes a value equal to the Winsorization point.

Example 1.10 – Figs and Wasps

Ecologists are interested in systems that have evolved to reflect seed-eating pollinator mutualism. In such systems, pollinating insects or birds receive benefit from the plant in the form of food or substrate for laying eggs and the plants receive the benefit of pollination, allowing reproduction. One example is provided by figs (family *Moraceae*) and the wasps that pollinate them (family *Agaonidae*). Figs produce enclosed inflorescences (called syconia) that contain a large number of female florets (ovules – think of tiny enclosures in the shape of a cup) that produce chemicals that attract certain species of wasps. Foundress wasps enter these inflorescences and pollinate flowers while laying eggs into some of the ovules before dying within the inflorescence (Janzen, 1979). Wasp larvae develop for several weeks before males (who are wingless) emerge and compete for access to females. After mating, females collect pollen and exit the syconia through holes chewed out by male wasps. There are a fairly large number (hundreds) of ovules available in a fig syconia and yet this is still a finite resource available to the wasps. While the fig-wasp system can become quite complex (involving also a number of species of parasite wasps that gain benefit from the figs without providing any benefit to the plant), one of many questions that ecologists ask about this system is the amount of the available resource that is utilized by wasps. Data to address this question are gathered by collecting fig syconia, dissecting them

and counting the number of ovules that are occupied and unoccupied by wasps. One model that could be used in this problem would be a binomial (y occupied out of n total ovules). But the number of ovules is large and one could also contemplate fitting a model with Poisson random variables to such data, except that the support of Poisson probability mass functions is unbounded. As yet another alternative, then, one could formulate a model based on Winsorized Poisson probability mass functions of the form, for $\lambda > 0$ and positive integer R,

$$\begin{aligned} f(y|\lambda, R) &= \frac{1}{y!}\lambda^y \exp(-\lambda) I(y < R) \\ &+ \left[1 - \sum_{x=0}^{R-1} \frac{1}{x!}\lambda^x \exp(-\lambda)\right] I(y = R); \quad y = 0, 1, \ldots, R \end{aligned} \tag{3.7}$$

Clearly, the behavior of the probability mass function in (3.7) depends on the relative magnitudes of λ and R. If R is not substantially larger than λ it can easily occur that $Pr(Y = R) > Pr(Y = R - 1)$ which is typically not the behavior we would wish to have reflected. Kaiser and Cressie (1997) formulate a spatial model with Winsorized Poisson distributions but, along the way, demonstrate that as long as R is about 3λ to 4λ a Winsorized Poisson distribution behaves much like a regular Poisson distribution.

Chapter 4

Models Based on Stochastic Processes

We turn our attention now to models for, or that incorporate aspects of, *stochastic processes*. The most common models included in this category are time-series models, models of spatial processes, and models used in what is known as queuing theory. Each of these types of models are topics unto themselves (our department offers 3 courses on time series at different levels, 3 courses on spatial analysis at different levels, and queuing theory is covered in various places, notably in a service course offered to computer scientists). Thus, our objective is not to cover this topic in all of its glory but, rather, to communicate the basic concepts of stochastic processes and the manner in which they can be incorporated into statistical models.

4.1 Restrictions in Statistical Models

Consider the simple model, for $-\infty < \mu < \infty$, $\sigma > 0$ and $i = 1, \ldots, n$,

$$Y_i = \mu + \sigma\,\epsilon_i; \quad \epsilon_i \sim iid\, N(0,1).$$

This model offers several restrictions on the distribution of the variables Y_i. Certainly there is a restriction to normal distributions, but we have also specified that each random variable has the same expected value (μ) and the same variance (σ^2). Such restrictions serve a fundamental purpose, which is to give us multiple variables from the same *statistical population*. This is necessary for progress to be made, first of all in statistical abstraction but also if we have any hope of producing estimators with known properties. For example, contrast the above mode with the alternative

$$Y_i = \mu_i + \sigma_i\,\epsilon_i; \quad \epsilon_i \sim iid\, N(0,1).$$

What would we do with such a model without some restrictions or further modeling for the μ_i and/or σ_i? (not much would be a good answer).

We often place restrictions on distributions through the first two moments (mean and variance) by:

1. Specifying a constancy among a group of random variables, such as $\mu_i = \mu$ or $\sigma_i^2 = \sigma^2$.

2. Modeling unequal quantities as a function of covariates that depend on a small number of parameters, such as regression models for means or variance models such as those we covered in additive error regression in Statistics 520.

3. Modeling of means and/or variances as random variables that follow distributions with a small number of parameters, such as hierarchical or mixture models.

Just as in these cases with which we are already familiar, we will need to ensure that models for stochastic processes, or that incorporate aspects of stochastic processes, allow repeated observation of the same probabilistic behaviors. To foreshadow, it is through assumptions of stationarity that we obtain repeatability. Stationarity is a topic we will cover shortly, but first we need to discuss what is meant by a stochastic process in a bit more detail.

4.2 Stochastic Processes and Random Fields

The world we encounter on the scale of typical human experience (i.e., forget special relativity for the moment) is 4−dimensional in nature. Three of these are spatial dimensions and the fourth is temporal. A stochastic process is a collection of random variables indexed by one or more of these dimensions, in which restrictions on distributional characteristics (usually means and variances) are also functions of the indices. Such collections of random variables are often called *stochastic processes* for an index in one dimension, and *random fields* for indices in more than one dimension, although these terms may be used interchangeably.

We will present a few of the basic models for processes in time and space, but first list two examples that lie outside of what we will consider in order to indicate that we are not covering all possibilities in detail.

1. Let $t = 1, 2, \ldots$ index discrete points in time, and define a random variable for each of these points as $Y(t)$. Suppose that $Y(t) \in \Omega_Y =$

$\{0, 1, 2, 3\}$. Here, the set of possible values Ω_Y is often called the *state space* of the process $\boldsymbol{Y} = \{Y(t) : t = 1, 2, \ldots, \}$ since it is the set or space of the possible states that the process can assume. Now, for $j, k \in \Omega_Y$, define values $t_{j,k}$ as,

$$t_{j,k} = Pr[Y(t) = k \,|\, Y(t-1) = j].$$

The values $t_{j,k}$ may be collected in a matrix

$$T \equiv \begin{pmatrix} t_{0,0} & t_{0,1} & t_{0,2} & t_{0,3} \\ t_{1,0} & t_{1,1} & t_{1,2} & t_{1,3} \\ t_{2,0} & t_{2,1} & t_{2,2} & t_{2,3} \\ t_{3,0} & t_{3,1} & t_{3,2} & t_{3,3} \end{pmatrix},$$

such that $t_{j,k} \geq 0$ for all j, k, and $\sum_k t_{j,k} = 1$ for all j. The stochastic process $\boldsymbol{Y}$ is called a discrete Markov process (discrete because it is indexed at discrete points in time) that in this example also has a discrete state space Ω_Y. This *transition matrix* can be estimated from a finite sequence of observed values if suitable restrictions are placed on its component quantities, the $t_{j,k}$, generally as properties of the matrix T. Suitable restrictions include properties of T known as *irreducible*, *positive recurrent*, and *aperiodic*, which we will not go into further here.

2. Let $\{N(t) : t \geq 0\}$ represent the number of some type of events that have occurred between time 0 and time t. Suppose that,

   - $N(0) = 0$.
   - $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$ are independent for any disjoint intervals (s_1, t_1) and (s_2, t_2).

- For any $s, t \geq 0$ and any $x \in \{0, 1, \ldots\}$,

$$Pr[N(t+s) - N(s) = x] = \frac{1}{x!}(\lambda t)^x \exp(-\lambda t).$$

Then $\{N(t) : t \geq 0\}$ is a continuous-time Poisson process (which has discrete state space). In this instance, all necessary restrictions are built into the definition of the process.

4.3 Stationarity

For stochastic processes, the types of restrictions made on means and variances (including covariances) of the process random variables are often those necessary to produce *stationary* behavior in the model. There are actually several types of stationarity. To introduce these concepts of stationarity, we first set forth some general notation appropriate for random fields.

Let $\boldsymbol{s}$ denote a (non-random) variable that contains information on a "location" in a system of spatial/temporal indices. For example, in a two-dimensional spatial process $\boldsymbol{s}$ might be defined as $\boldsymbol{s} \equiv (u, v)$ for longitude u and latitude v, or some transformation of latitude and longitude that makes (u, v) amenable to Euclidean geometry (e.g., universal transverse-mercator coordinates, or UTMs). In a one-dimensional time series, we may have $\boldsymbol{s} \equiv t$, where t is a point in time. Notice that we have taken this location variable $\boldsymbol{s}$ to be continuous in its time/space domain, which leads to a continuous process $\boldsymbol{Y} \equiv \{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^d\}$, where $\mathcal{D}$ is the domain of the process. We assume from the outset that $var\{Y(\boldsymbol{s})\} < \infty$ for all $\boldsymbol{s} \in \mathcal{D}$.

1. First-Order Stationarity.

   The process $\boldsymbol{Y}$ is said to be *first-order* stationary if,

$$E\{Y(\boldsymbol{s})\} = \mu \qquad \forall \boldsymbol{s} \in \mathcal{D}$$

2. Intrinsic Stationarity.

   The process $\boldsymbol{Y}$ is said to be *intrinsically* stationary if it is first-order stationary and

   $$var\{Y(\boldsymbol{s}) - Y(\boldsymbol{s}+\boldsymbol{h})\} = V(\boldsymbol{h}) \quad \forall \boldsymbol{s}, \boldsymbol{h} \in \mathcal{D},$$

   for some function $V(\cdot)$. This concept of stationarity appears most often in spatial problems, but there is no reason it does not apply equally to temporal processes.

3. Second-Order Stationarity.

   The process $\boldsymbol{Y}$ is said to be *second-order* stationary if it is first-order stationary and

   $$cov\{Y(\boldsymbol{s}),\, Y(\boldsymbol{s}+\boldsymbol{h})\} = C(\boldsymbol{h}) \quad \forall \boldsymbol{s}, \boldsymbol{h} \in \mathcal{D},$$

   for some function $C(\cdot)$. While second-order and intrinsic stationarity are clearly related, they are not the same. In fact, it is possible to demonstrate that second-order stationarity implies intrinsic stationarity, but not the converse. Thus, second-order stationarity is a stronger condition than is intrinsic stationarity.

4. Strict Stationarity. Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m$ be any finite collection of locations in $\mathcal{D}$. Define

   $F_{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_m}(y_1, \ldots, y_m) =$

   $$Pr[Y(\boldsymbol{s}_1) \leq y_1, \ldots, Y(\boldsymbol{s}_m) \leq y_m].$$

   The process $\boldsymbol{Y}$ is said to be *strictly* stationary if, for all $\boldsymbol{h} \in \mathbb{R}^d$ and all $m \geq 1$,

   $$F_{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_m}(y_1, \ldots, y_m) = F_{\boldsymbol{s}_1+h,\ldots,\boldsymbol{s}_m+h}(y_1, \ldots, y_m)$$

Now, what do these various types of stationarity mean? Begin by considering a process in $\mathbb{R}^1$ such as a time series (could be a spatial transect, but time is adequate).

If a process in time is strictly stationary, then random variables for any set of times separated by a fixed distance have the same joint distribution. For example, $\{Y(1), Y(3), Y(10)\}$, $\{Y(6), Y(8), Y(15)\}$, and $\{Y(20), Y(22), Y(29)\}$ all have the same joint $3-$dimensional distribution. Similarly, $\{Y(4), Y(5)\}$ and $\{Y(150), Y(151)\}$ have the same $2-$dimensional joint distribution. This is, in fact, true for any number of random variables and any fixed difference in time. Clearly, this is a strong property. If a process in time is second-order stationary, then random variables for any set of times separated by a fixed distance have the same first two moments, but not necessarily the same distribution. Thus, strict stationarity implies second-order stationarity, but not the converse. Note, however, that if we specify normal distributions then second-order stationarity does implies strict stationarity since normal distributions are characterized by the first two moments.

If a process in time is intrinsically stationary, then random variables for any set of times separated by a fixed distance have variances of their differences that are the same. For example, $var\{Y(1) - Y(3)\} = var\{Y(6) - Y(8)\} = var\{Y(20) - Y(22)\}$. What is the relation between this and second-order stationarity? For any two random variables X and Y, $var(X - Y) = var(X) + var(Y) - 2cov(X, Y)$, or, $cov(X, Y) = (1/2)\{var(X) + var(Y) - var(X - Y)\}$. It is then entirely possible for $var\{Y(\boldsymbol{s}_i), -Y(\boldsymbol{s}_j)\}$ to be a function of $\boldsymbol{s}_i - \boldsymbol{s}_j$ alone (intrinsic stationarity) without $cov\{Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\}$ also being a function of $\boldsymbol{s}_i - \boldsymbol{s}_j$ alone (second-order stationarity). Intrinsic stationarity leads to second-order stationarity only with the additional restriction that $var\{Y(\boldsymbol{s}_i)\} = var\{Y(\boldsymbol{s}_j)\}$. Thus, intrinsic stationarity is a

weaker condition than is second-order stationarity. To generalize the above interpretations from $1-$dimensional real space to $d-$dimensional real space requires only replacing time differences with higher dimension displacement.

4.4 Two Fundamental Time Series Models

In this section we will consider stochastic processes for which we have a discrete set of indices indicating equally spaced points in time, $\boldsymbol{s}_i \equiv t$ for $t = 0, \pm 1, \pm 2, \ldots$. Two basic structures for temporal models, from which many more elaborate structures can be constructed (e.g., Newton, 1988; Shumway, 1988; Brockwell and Davis, 1991) are *moving average* and *autoregressive* models. In most texts on time series, these processes are presented in terms of a "backshift operator" $Bx_t = x_t - 1$. While this leads to compact notation and is useful in writing down results about these models, it does not necessarily promote understanding of the *structures* involved. Thus, we will avoid its use here, and present moving average and autoregressive models in a very simple form.

4.4.1 Moving Average Models

A basic moving average model formulation can be written as,

$$Y(t) = \sum_{k=0}^{q} \beta_k \, \epsilon(t-k), \tag{4.1}$$

in which we take $\beta_0 = 1$, and $\epsilon(t) \sim iid\, N(0, \sigma^2)$. Expression (4.1) would be said to represent a "qth order moving average model. To see the manner in which moving average models conceptualize a temporal process, consider a

$2nd$ order process written as

$$Y(t) = \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2}. \tag{4.2}$$

In (4.2) we see clearly that the process at time t is composed of a linear combination of independent error or innovation terms ϵ_t, ϵ_{t-1} and ϵ_{t-2}. Clearly, the process has mean 0 for all t and is thus automatically first-order stationary. Although we will not write down an explicit form here, it should also be clear that the variances of $Y(t)$ and covariances of $Y(t)$, $Y(t+s)$, $s \leq q$, will be the variance of the innovation terms (σ^2) times a polynomial in the coefficients, the βs. For time differences or lags of greater than q, the covariance will be 0. Thus, the process is second-order stationary.

A major concern with time series models of this type (this will be true also for autoregressive models) is whether parameters of the model (4.1) can be *identified* based on the covariance function of the process. The answer, in general, is no it cannot be. Consider an example taken from Newton (1988, p. 96), for a first order moving average process,

$$Y(t) = \epsilon_t + \beta\, \epsilon_{t-1},$$

which has first and second moments,

$$\begin{aligned} E\{Y(t)\} &= 0, \\ var\{Y(t)\} &= \sigma^2(1+\beta^2), \\ cov\{Y(t), Y(t+1)\} &= \beta\, \sigma^2, \\ cov\{Y(t), Y(t-1)\} &= \beta\, \sigma^2. \end{aligned}$$

Suppose that $var\{Y(t)\} = 5$ and $cov\{Y(t), Y(t+1)\} = 2$. Combinations of $\sigma^2 = 1$ with $\beta = 2$ or $\sigma^2 = 4$ with $\beta = 0.5$ both lead to these same

moments. In general, for a moving average process of order q, there are 2^q sets of parameters that lead to the same variance and covariances. There is only one set of parameters, however, that lead to what is called an *invertible* model. We will leave this topic until after we have introduced autoregressive models.

4.4.2 Autoregressive Models

The basic form of an autoregressive model is,

$$Y(t) = \sum_{k=1}^{p} \alpha_k Y(t-k) + \epsilon(t), \tag{4.3}$$

where, as before, $\epsilon(t) \sim iid\, N(0, \sigma^2)$. Expression (4.3) is said to represent a p^{th} order autoregressive process. Consider, analogous to our $2nd$ order moving average model (4.2) a $2nd$ order autoregressive model,

$$Y(t) = \alpha_1 Y(t-1) + \alpha_2 Y(t-2) + \epsilon(t). \tag{4.4}$$

Models (4.3) and (4.4) look like a linear regression of values of $Y(t)$ on previous values, which they are. Hence the name *autoregressive* which is more intuitive than the previous name of moving average for (4.1) or (4.2). The process represented by model (4.3) does not necessarily have constant mean, so it is usually assumed that the $Y(t)$ have constant mean zero. In practice, data are generally "de-trended" through regression (over time) or the process of "first-differencing" to remove at least linear trends in means of the $Y(t)$.

An autoregressive process is not necessarily second-order stationary unless some conditions are placed on the coefficients $\alpha_1, \ldots, \alpha_p$. This is also related to invertibility and will be covered shortly. The covariance of an autoregressive model can be determined using what are called the *Yule-Walker*

equations (e.g., Shumway 1988, p.135). To illustrate the ideas involved, consider a first order autoregressive model,

$$Y(t) = \alpha\, Y(t-1) + \epsilon(t),$$

with $\epsilon(t) \sim iid\, N(0, \sigma^2)$. Directly from this model we have the relations

$$\begin{aligned}
Y(t)Y(t) &= \alpha Y(t-1)Y(t) + \epsilon(t)Y(t) \\
Y(t)Y(t-1) &= \alpha Y(t-1)Y(t-1) + \epsilon(t)Y(t-1) \\
&\cdot \quad \cdot \\
&\cdot \quad \cdot \\
&\cdot \quad \cdot \\
Y(t)Y(t-k) &= Y(t-1)Y(t-k) + \epsilon(t)Y(t-k)
\end{aligned}$$

Since all $Y(t)$ are assumed to have expectation 0, taking expectations in these expressions yields,

$$\begin{aligned}
var\{Y(t)\} &= \alpha cov\{Y(t-1),\, Y(t)\} + \sigma^2 \\
cov\{Y(t),\, Y(t-1)\} &= \alpha var\{Y(t-1)\} + 0 \\
&\cdot \quad \cdot \\
&\cdot \quad \cdot \\
&\cdot \quad \cdot \\
cov\{Y(t),\, Y(t-k)\} &= cov\{Y(t-1),\, Y(t-k)\} + 0
\end{aligned} \tag{4.5}$$

The final term on the rhs of the first equality comes from the fact that

$$E\{Y(t)\epsilon(t)\} = E\left[\alpha Y(t-1)\epsilon(t) + \{\epsilon(t)\}^2\right] = \sigma^2,$$

since $Y(t-1)$ contains only random variables that are independent of $\epsilon(t)$. This also explains why the remaining right most terms are all 0.

Substituting the second line of (4.5) into the first, and assuming that we have an equal variance process gives an equality that allows derivation of the variance of the $Y(t)$,

$$\begin{aligned} var\{Y(t)\} &= \alpha^2 var\{Y(t-1)\} + \sigma^2 \\ var\{Y(t)\} &= \frac{\sigma^2}{1-\alpha^2} \end{aligned} \tag{4.6}$$

Substituting the first line of (4.5) into the second gives,

$$\begin{aligned} cov\{Y(t),\, Y(t-1)\} &= \alpha^2 cov\{Y(t),\, Y(t-1)\} + \alpha\sigma^2 \\ &= \frac{\alpha\,\sigma^2}{1-\alpha^2}. \end{aligned}$$

Continuing in this manner shows that

$$cov\{Y(t),\, Y(t-k)\} = \frac{\alpha^k\,\sigma^2}{1-\alpha^2}. \tag{4.7}$$

Equations (4.6) and (4.7) indicate that in order for an AR process to be stationary we must have $|\alpha| < 1$. Given that this is true, the correlations between values of the process separated by a distance k is, (here allowing k to take values $0, \pm 1, \pm 2, \ldots$),

$$corr\{Y(t),\, Y(t-k)\} = \alpha^{|k|}$$

An important point here is the distinction between a moving average model of order one and an autoregressive model of order one. In the moving average model, covariance (and hence correlation) between values separated by more than one lag are 0, while in the autoregressive model the covariance (and hence the correlation) decays more slowly over time as a power function of the coefficient α (which must be smaller than 1 in absolute value). This same

distinction extends to moving average and autoregressive models of higher orders.

4.4.3 Inversion

We have mentioned the issue called *inversion* previously. Inversion deals with conditions under which there exists a duality between moving average and autoregressive processes, our primary interest being situations in which both processes are stationary. Inversion concerns conditions under which the processes are the same. Box and Jenkins (1970, p. 50) point out that invertibility and stationarity are different properties. This is true, but it turns out that conditions needed to produce invertibility of stationary moving average processes are similar to those needed to produce stationarity in invertible autoregressive processes. We need to explain this remark, and do so in the following, although without the requisite derivations. For those see any of the standard time series textbooks referenced previously.

A key quantity in determination of both invertibility and stationarity is what is called the *characteristic polynomial* or *characteristic equation* which, for finite moving average (order q) and finite autoregressive (order p) processes, and a possibly complex argument z are,

$$\begin{aligned} h(z) &= 1 + \sum_{k=1}^{q} \beta_k \, z^k, \\ g(z) &= 1 - \sum_{k=1}^{p} \alpha_k \, z^k. \end{aligned} \tag{4.8}$$

Comments

1. In these notes, I have written both moving average (4.1) and autoregressive (4.3) models as sums rather than differences. It is also common

(e.g., Box and Jenkins 1970; Shumway 1988) to write moving average processes with coefficients being the negative of those in (4.1), in which case the characteristic polynomials of both Moving Average and Autoregressive models have the form of $g(z)$ given in (4.8); note also that Newton (1988) does neither of these, writing autoregressive models to isolate the error term $\epsilon(t)$ on the rhs of the model.

2. In consideration of a first order autoregressive process we arrived at the need to have $|\alpha| < 1$ in order for covariances to remain finite and, in fact, then also stationary. For a general finite autoregressive process of order p, this condition is translated into conditions on the roots (zeros) of the characteristic polynomial $g(z)$, since these will be determined by values z^0 that are functions of the coefficients $\{\alpha_1, \ldots, \alpha_p\}$. Similarly, conditions on the coefficients of a finite moving average process of order q can be specified in terms of conditions on the roots of the characteristic polynomial $h(z)$.

3. The conditions that produce desired behaviors in finite moving average and autoregressive processes turn out to be conditions on whether the roots of the characteristic polynomials lie *inside*, *on*, or *outside* of the unit circle (i.e., less than, equal, or greater than 1 in modulus).

We can now summarize, without proof, what can be a confusing array of results regarding moving average and autoregressive time series models.

1. Finite moving average processes are always second order stationary.

2. Finite autoregressive processes are second order stationary if all of the zeros of $g(z)$ in (4.8) are greater than 1 in modulus.

3. Finite moving average processes can be written as infinite autoregressive processes if the zeros of $h(z)$ in (4.8) are all greater than 1 in modulus.

4. Finite autoregressive processes can be written as (doubly) infinite moving average processes as long as none of the zeros of $g(z)$ in (4.8) is equal to one in modulus. *Note: a doubly infinite moving average processes is as in (4.1) but with the summation going from $-\infty$ to ∞, (see, e.g., Fuller 1996).* In addition, finite autoregressive processes can be written as (singly) infinite moving average processes if all of the zeros of $g(z)$ in (4.8) are greater than 1 in modulus.

Given invertible models (moving average and autoregressive) the question in an application is which representation is more parsimonious (adequate, with as few parameters as possible). Clearly, if a process is truly a moving average model of order one, using an autoregressive representation is not desirable as the moving average model would have 2 parameters, β and σ^2, while the autoregressive model would have an infinite number of parameters. The reverse would be true for a process that was in fact an autoregressive process of order one. This leads, in time series analysis, to the representation of processes as a combination of moving average and autoregressive models. These are called autoregressive-moving average (ARMA) models, and have the general form,

$$\sum_{j=0}^{p} \alpha_j Y(t-j) = \sum_{k=0}^{q} \beta_k \epsilon(t-k),$$

with $\alpha_0 = \beta_0 = 1$. This looks more familiar if we write it as,

$$Y(t) = \sum_{j=1}^{p} \alpha_j Y(t-j) + \sum_{k=1}^{q} \beta_k \epsilon(t-k) + \epsilon(t),$$

Note that the coefficients α_j in these two expressions are negatives of each other.

4.4.4 Dependence on the Past

One final point relative to moving average and autoregressive time series models is worthy of note (it is, perhaps, even more important from a modeling standpoint than the topic of inversion). This concerns the manner in which moving average and autoregressive models represent the dependence of the current $Y(t)$ on past values. We have already seen, through consideration of first order processes, that moving average models have pairwise correlation of zero for lags greater than 1 (in general this will be for lags greater than the order of the model q), while the pairwise correlation for autoregressive models dies off more slowly.

Now, consider the conditional distribution of $Y(t)$ given $Y(t-1) = y(t-1)$, in the case of first order autoregressive and first order moving average models. For the autoregressive model we have,

$$Y(t) = \alpha\, y(t-1) + \epsilon(t),$$

which clearly demonstrates that, given a value $Y(t-1) = y(t-1)$, the distribution of $Y(t)$ does not depend on any previous values of the process. On the other hand, for a first order moving average model,

$$\begin{aligned} Y(t) &= \beta\, \epsilon(t-1) + \epsilon(t) \\ &= \beta\, y(t-1) - \beta^2\, \epsilon(t-2) + \epsilon(t), \end{aligned}$$

so that, given $Y(t-1)$, $Y(t)$ is not independent of $Y(t-2)$, which is a function of $\epsilon(t-2)$ as $\epsilon(t-2) = Y(t-2) - \beta\epsilon(t-3)$.

How can this be? After all, we have shown that the covariance of $Y(t)$ and $Y(t-2)$ is 0 for a first order moving average process. How can they then be dependent? The answer lies in what a covariance matrix represents. A covariance matrix represents *pairwise dependence* in a *joint distribution* of variables. In the case of normality, this also represents dependence in the marginal distributions of any subset of component variables. But, even in the case of normality, the covariance matrix does not represent *conditional dependence*. For normal distributions conditional dependence is given by the *inverse* covariance matrix.

One way to think of this is that in a Gaussian (or joint normal) distribution, the elements of the covariance matrix represent dependence when all other variables than the pair involved are averaged over. This is marginal dependence. On the other hand, the inverse covariance matrix represents dependence when all other variables than the pair involved are conditioned on. In a first order moving average model, given $Y(t-1)$, $Y(t)$ is dependent on $Y(t-2)$ because they are both related to or dependent on $Y(t-1)$. But marginally, all of that shared dependence of $Y(t)$ and $Y(t-2)$ has been averaged out by integrating over possible values of $Y(t-1)$. In a first order autoregressive model, on the other hand, $Y(t)$ and $Y(t-2)$ are marginally dependent (under normality) but conditionally independent.

Another way to form some intuition about all of this is to consider the process of generation through time by which values of time series are produced. In a first order moving average model, it is innovation terms that are propagated through time to directly influence future variables. If we would condition on these innovation terms, then there would be no conditional dependence beyond lag one, that is, $Y(t)$ given $\epsilon(t-1)$ would be independent of all previous terms (either Ys or ϵs). But, because these innovations corre-

spond to latent (unobservable) variables, we condition on the actual value of the series at lag one $Y(t-1)$, which is composed partly of $\epsilon(t-1)$ but also partly of $\epsilon(t-2)$. Thus, conditioned on $Y(t-1)$ the current value $Y(t)$ is still dependent on $\epsilon(t-2)$ and, hence also $Y(t-2)$, and this extends to all previous terms $Y(t-3)$, $Y(t-4)$, etc. In a first order autoregressive model, it is the actual process variables (the Ys) that are directly propagated through time. Thus, all of the effect of previous ϵs is embodied in $Y(t-1)$ and, since this value directly affects $Y(t)$, conditioning on $Y(t-1)$ is the same as conditioning on all previous terms. That is, first order autoregressive models possess a first order Markov property in time (while p^{th} order autoregressive models are p^{th} order Markov in time).

4.4.5 Goals and Limitations of Traditional Time Series Models

We close our discussion of basic time series structures with a few thoughts on modeling issues related to these formulations. It should be noted, however, that the theory and application of time series forms a rich body of statistical knowledge and investigation, and there are many extensions of the basic ideas we have presented. Nevertheless, the following seem pertinent:

1. Everything we have considered has assumed zero mean processes (or constant mean processes in which the constant is easily removed by subtraction). In applications for which the mean is not constant, time series models are usually applied to series of residual quantities from which an estimated mean structure has been removed, or to a series of differenced values which results in what are called autoregressive-integrated-moving average models (ARIMA models). The point is that

time series modeling is focused on a signal plus noise conceptualization of scientific phenomena, in which the signal is removed and time series are used to describe remaining structure in the noise, or error, process.

2. Building on comment 1, the goals of time series analysis are generally those of data description, prediction, and forecasting without an overly great concern for conceptualization of the scientific problem under study, other than what might be incorporated in modeling of the mean structure.

3. A practical limitation of the models presented in this section is the assumption of equally-spaced time points $t = 0, \pm 1, \pm 2, \ldots$. While there are certainly a good number of applications in which this assumption is met – the references given previously contain a large number of examples and applications – there are also many problems that logically involve temporal dependence in which observations cannot be taken at equally spaced points in time, have a few missing values, or cannot be observed for certain periods of time which produces gaps in the record.

4. We have also, in our presentation of moving average and autoregressive models, assumed that innovation terms followed normal distributions. Following the same tradition as additive error models, time series models are most often presented without specific distributional assumptions, taking those terms to be simply random variables with expectation zero and constant variance. However, again in a manner similar to additive error models, there is an implicit appeal to normal distributions. We have simply made this explicit in the preceding presentation.

4.5 Random Fields

It is tempting, at this point, to discuss models for spatial processes as a distinct topic, but doing so would restrict the more general nature of models for *random fields*. In fact, everything covered in the previous subsection on time series models would fit under the heading of models for random fields. So time series models of the type discussed previously are simply one class of models for random fields. It is also true that problems other than those involving only temporal and/or spatial dependence can be considered within the context of models for random fields. Despite this, however, spatial problems present a natural situation which which to illustrate random fields in more than one dimension, and this subsection will have a decidedly spatial flavor and much discussion of the topics included will be found in the literature on spatial statistics (e.g., Haining 1990; Cressie 1993; Griffith and Layne 1999; Lawson 2001).

4.5.1 The Structure of Random Fields

Variables in a random field are indexed by a (typically non-random) *location* $\boldsymbol{s}$. For a collection of n random variables we will index them by their locations as $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$. It may be, for example, that in two dimensional geographic space, $\boldsymbol{s}_i = (u_i, v_i)$ where u_i is longitude and v_i is latitude. For random fields in general it is not necessary that locations $\boldsymbol{s}_i$ correspond to actual physical locations. Consider, for example, a longitudinal study in which n_j observations are taken over time for each of $j = 1, \ldots, k$ situations or subjects. We can place response random variables in a random field structure by defining $\boldsymbol{s}_i \equiv (j,\, t_j)$ where t_j is the time of an observation on a given subject j, for $j = 1, \ldots, k$. In a mathematical graph or network, we might assign

binary random variables to potential edges, and define $\boldsymbol{s}_i = (u_i, v_i)$ where u_i and v_i are the two nodes that would be joined by potential edge i. Nevertheless, a situation involving two dimensional space provides a convenient setting within which to develop an understanding of random fields, and one can think in terms of that setting for most of what follows.

We will, very briefly, present two random field structures called *continuous index* random fields and *discrete index* random fields. A third type of structure, *point processes*, will not be discussed here. The primary distinction among these types of random field models rests on the type of location indices $\boldsymbol{s}$ assumed for each. A general structure is given in the following:

Let $\mathcal{D} \subset \mathbb{R}^d$ be a subset of $\mathbb{R}^d$ that has positive volume. We define a general random field process as,

$$\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}, \tag{4.9}$$

where, at this point, we are making no assumptions about the quantities $Y(\boldsymbol{s})$ and $\boldsymbol{s}$ (i.e., random or non-random).

4.5.2 Continuous Index Random Fields

In a continuous index random field, we take, in the general process (4.9), $\mathcal{D}$ to be a fixed subset of $\mathbb{R}^d$, and allow $\boldsymbol{s}$ to vary continuously over $\mathcal{D}$. The response variables $Y(\boldsymbol{s})$ are taken as random. In general, $Y(\boldsymbol{s})$ may be multivariate, but we will consider only univariate processes. It is worthy of mention that, although the process itself is considered continuous, data will be available at only a discrete and finite number of locations in the form

$$\{y(\boldsymbol{s}_1), y(\boldsymbol{s}_2), \ldots, y(\boldsymbol{s}_n)\}.$$

In essence this means that, for a random field with no restrictions on expected values, variances, or distributional form, we have available only a partial realization of a unique stochastic process, that is, a sample of size less than one. This means then, from a modeling viewpoint, that restrictions on expected values, variances, or distributional forms are crucial in order to make progress in terms of estimation, inference and, even more fundamentally, statistical conceptualization or abstraction.

Consideration of dependence in a random field leads to a quantification of dependence that is an alternative to that of covariance – values of quantities that are closer together (in a random field distance) should be more similar than values of quantities that are farther apart. This concept of dependence is most easily understood if distance means geographic (i.e., spatial) distance, but a spatial setting is not necessary if a suitable metric can be defined for locations in a random field. The idea that things that are closer together should be more similar than things that are farther apart is directly represented in a quantity called the *variogram*, defined as a function which describes the variances of the differences of random variables at two locations, say $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$,

$$2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \equiv var\{Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)\}; \quad \text{all } \boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{D}. \tag{4.10}$$

In (4.10) the function $2\gamma(\cdot)$ is called the variogram. Just as functions must satisfy certain conditions to be density or mass functions, and matrices must satisfy certain conditions to be covariance matrices, so too must a function $2\gamma(\cdot)$ satisfy a certain condition to be a variogram. This condition is called *conditional negative definiteness*, and is defined as follows.

Definition

A variogram $2\gamma(\cdot)$ is conditionally negative definite if, for any finite number

of locations $\{\boldsymbol{s}_i : i = 1, \ldots, m\}$ and real numbers $\{a_i : i = 1, \ldots, m\}$ such that $\sum a_i = 0$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j 2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \leq 0. \tag{4.11}$$

Any function $2\gamma(\cdot)$ that satisfies (4.10) must have the property (4.11). Note, at this point, that we have made no assumptions regarding $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ in terms of mean (i.e., first order stationarity) or variance (i.e., constant variance). Note also, however, that by writing a variogram as a function of displacement between locations (the argument to $2\gamma(\cdot)$ in (4.10) is the displacement $\boldsymbol{s}_i - \boldsymbol{s}_j$) we have, in fact, assumed intrinsic stationarity if the process has constant mean.

An additional assumption is often made about the variogram, which may be checked by a data-driven model assessment, that the value of the variogram for two locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ depends only on the distance between the locations, $d_{i,j} \equiv ||\boldsymbol{s}_i - \boldsymbol{s}_j||$. For example, if $\boldsymbol{s}_i = (u_i, v_i)$ for a horizontal position u_i and vertical position v_i, Euclidean distance would be $d_{i,j} = \{(u_i - u_j)^2 + (v_i - v_j)^2\}^{1/2}$. In this case, we may write the variogram as a function of only a distance $h \in \mathbb{R}$ as,

$$2\gamma(h) = var\{Y(\boldsymbol{s} + \boldsymbol{w}) - Y(\boldsymbol{s})\}; \qquad \boldsymbol{s} \in \mathcal{D}, \tag{4.12}$$

for any $\boldsymbol{w}$ such that $||\boldsymbol{s} - \boldsymbol{w}|| = h$. Note that there may be many displacements $\boldsymbol{w}$ from $\boldsymbol{s}$ that give the same distance h. If a random field process has a variogram that satisfies (4.12) then we say the process is *isotropic*.

While we will not go into detail here, the goal in application of a continuous index random field model is usually prediction. Forecasting is possible, but becomes more tenuous than in time series because assumptions are being made that the form of a process remains similar for observations beyond the

scope of the data in more than one dimension. Forecasting based on data description but not understanding is dangerous even in one dimension but this danger is compounded when a lack of understanding is extended to more than one dimension of our physical world.

The general progression of developing a predictor for unobserved locations within the extent of the data is as follows.

1. The variogram $2\gamma(\cdot)$ is estimated from a given set of data (this is most often moment-based estimator) for a finite set of displacements (usually distances, under an assumption of isotropy).

2. A theoretical model that ensures conditional negative definiteness is fit to the estimated variogram values (much like a nonlinear regression).

3. A linear predictor to minimize prediction mean squared error is developed and turns out to be a function of observed data and values of the variogram, which are now taken from the model fitted in step 2.

4. Uncertainty in predictions may be derived by substituting the linear predictor into the prediction mean squared error which it minimizes, and then assuming normality.

In the area called *Geostatistics* such a prediction system is known as *kriging*. For a full development of kriging under various assumptions on the process (e.g., constant versus nonconstant mean, variances, etc.) see Cressie (1993).

As a final comment on continuous index random field models, note that nothing in the development presented has assumed constant variance for the process $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$. A variogram model, combined with a model for

variances yields a covariance model since,

$$\begin{aligned} cov\{Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)\} &= var\{Y(\boldsymbol{s}_i)\} + var\{Y(\boldsymbol{s}_j)\} \\ &- var\{Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)\}. \end{aligned}$$

For example, under an assumptions of constant variance σ^2, and an isotropic variogram $2\gamma(h)$, a covariance model for an isotropic, second order stationary process is,

$$2\,C(h) = 2\,C(0) - 2\,\gamma(h).$$

Thus, modeling the variogram plus the variance leads to a model for the covariance, but not necessarily the other way around. Nevertheless, it has become common to model continuous index spatial processes in terms of covariances rather than variograms.

4.5.3 Discrete Index Random Fields

To formulate a discrete index random field we take, in the general process of (4.9), $\mathcal{D}$ to be a fixed set of countable (usually finite) points, so that we have random field locations $\boldsymbol{s}_i \in \mathcal{D}$ for a discrete set of locations $\{\boldsymbol{s}_i : i = 1, \ldots\}$. As for continuous index processes, $Y(\boldsymbol{s}_i)$ are considered random variables.

In our (very) brief consideration of time series models, dependence was represented through the manner in which past values (autoregressive models) or past innovations (moving average models) where propagated through a process in a forward manner (there, through time). Models for continuous index random fields allowed an alternative formulation of dependence as a variogram model. Discrete index random fields present yet another vehicle by which dependence can be incorporated into a model, through what is called *neighborhood* structure in a *Markov random field*. Neighborhoods

specify which random variables are conditionally dependent or independent. It has become common to equate models for Markov random fields with what are called conditionally specified models, which we will discuss in this section. But it should be noted that models for Markov random fields are not a necessary consequence of discrete index random fields, nor are conditionally specified models a necessary consequence of Markov random fields. In fact, discrete index time series are discrete index random fields, some models for which (e.g., first order autoregressive models) possess a Markov property. Thus, one could consider a first order autoregressive model a model for a Markov random field, although it would not typically be considered a conditionally specified model.

First, we define what is meant by a neighborhood structure for a discrete index random field. The neighborhood of a location $\boldsymbol{s}_i$ is that set of locations N_i such that the full conditional distribution of the random variable $Y(\boldsymbol{s}_i)$ functionally includes the variables at locations contained in N_i. In practice, neighborhoods are usually determined by other than statistical considerations and are often, in fact, rather arbitrary in nature. Determination of appropriate neighborhoods is a difficult problem in the application of the types of models to be considered. Examples of neighborhoods for spatial problems with $\boldsymbol{s}_i \equiv (u_i, v_i)$ include:

1. If $\boldsymbol{s}_i$ denotes the center of a county or other political subdivision, we might define N_i to include those counties that share a common border with the county of which $\boldsymbol{s}_i$ is the centroid.

2. If the values $u_i : i = 1, \ldots, C$ and $v_i : i = 1, \ldots, R$ denote equally spaced vertices on a regular grid, we might specify that $N_i \equiv \{(u_j, v_j) : (u_j = u_i, v_j = v_i \pm \delta)$ or $(u_j = u_i \pm \delta, v_j = v_i)\}$, where δ is the grid

spacing. This is known as a *four nearest neighbors* structure.

3. For locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ that are either uniformly or non-uniformly distributed in space, we might define $N_i \equiv \{\boldsymbol{s}_j : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| \leq \kappa\}$ for some predetermined distance κ.

Simultaneous Autoregressive Model

If specified neighborhoods are used to define dependence parameters $\{b_{i,j} : i, j = 1, \ldots, n\}$, the following model has sometimes been called a *simultaneous autoregressive model* (SAR):

$$Y(\boldsymbol{s}_i) = \mu_i + \sum_{j=1}^{n} b_{i,j}\{Y(\boldsymbol{s}_j) - \mu_j\} + \epsilon_i, \tag{4.13}$$

where (usually), $\epsilon_i \sim iid\, N(0,\, \sigma^2)$, and $b_{i,i} = 0$, for all i. Model (4.13) is called an autoregressive model because of the structure of $Y(\boldsymbol{s}_i)$ regressed on values of $Y(\boldsymbol{s}_j)$, but it does **not** share properties with a time series autoregressive model. For one thing, as shown in Cressie (1993, p. 406), the error ϵ_i is not independent of the $\{Y(\boldsymbol{s}_j) : j \neq i\}$.

Markov Random Fields

We are familiar with the standard Markov assumption in time; given the entire past, the present depends on only the most immediate past. What does this imply for a joint distribution of variables on a one-dimensional random field (time or otherwise)? Let $\{Y_1, Y_2, \ldots, Y_n\}$ denote ordered random variables such as through time or on a spatial transect. Let $p(\cdot)$ denote a generic probability density or mass function corresponding to its arguments. That is, $p(y)$ is the marginal density of Y, $p(y_1, y_2)$ is the joint density of Y_1 and

Y_2, $p(y_1|y_2)$ is the conditional density of Y_1 given Y_2, and so forth. Then, as is always true,

$$p(y_1, \ldots, y_n) = p(y_1)p(y_2|y_1) \ldots, p(y_n|y_1, \ldots, y_{n-1})$$

which, by the usual Markov property, becomes

$$p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2|y_1) \ldots, p(y_n|y_{n-1}).$$

Note also that the Markov property implies that

$$p(y_i|\{y_j : j = 1, \ldots, i-2\}) = p(y_i|y_{i-2}).$$

What is implied about the full conditional distribution of a random variable Y_i is then,

$$
\begin{aligned}
& p(y_i|\{y_j : j \neq i\}) \\
&= \frac{p(y_1, \ldots, y_n)}{p(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)} \\
&= \frac{p(y_1)p(y_2|y_1)\ldots p(y_{i-1}|y_{i-2})p(y_i|y_{i-1})}{p(y_1)p(y_2|y_1)\ldots p(y_{i-1}|y_{i-2})} \\
&\times \frac{p(y_{i+1}|y_i)p(y_{i+2}|y_{i+1})\ldots p(y_n|y_{n-1})}{p(y_{i+1}|y_{i-1})p(y_{i+2}|y_{i+1})\ldots p(y_n|y_{n-1})} \\
&= \frac{p(y_i|y_{i-1})p(y_{i+1}|y_i)}{p(y_{i+1}|y_{i-1})} \\
&= \frac{p(y_i|y_{i-1})p(y_{i+1}|y_i, y_{i-1})}{p(y_{i+1}|y_{i-1})} \\
&= \frac{p(y_i, y_{i-1})p(y_{i-1}, y_i, y_{i+1})p(y_{i-1})}{p(y_{i-1})p(y_{i-1}, y_i)p(y_{i-1}, y_{i+1})} \\
&= \frac{p(y_{i-1}, y_i, y_{i+1})}{p(y_{i-1}, y_{i+1})} \\
&= p(y_i|y_{i-1}, y_{i+1})
\end{aligned}
$$

Thus, the typical Markov property in one dimension (e.g., time) implies that the conditional distribution of Y_i given all other variables depends on only the adjacent variables Y_{i-1} and Y_{i+1}.

It is the structure of such full conditional distributions that are the concern of Markov random fields. A collection of random variables $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ are said to constitute a Markov random field if, for each $i = 1, \ldots, n$, the conditional density or mass functions satisfy,

$$p(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}) = p(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in N_i\}), \tag{4.14}$$

where $\{N_i : i = 1, \ldots, n\}$ are neighborhoods.

4.6 Conditional Model Specification

While the definition of Markov random fields just given indicates that neighborhoods are determined by conditional distributions, in model formulation we typically want to reverse this progression by starting with defined neighborhoods and then using these to write forms for conditional density or mass functions. This method of formulating models is called *conditional model specification* and owes a great deal to early work by Besag (1974). In this method of modeling, we specify forms for the n full conditionals making use of neighborhoods as in (4.14) to reduce the complexity of these distributions. The key for modeling is to ensure that a joint distribution exists that has the corresponding set of specified conditionals. The key for statistical analysis is to identify this joint and make use of it in statistical estimation and inference procedures. This becomes a long and involved topic, beyond the scope of our class. See Besag (1974) for models based on one-parameter exponential families. Kaiser and Cressie (2000) relax some of the assumptions of Besag, and extend this modeling approach to multiple parameter exponential family conditionals. Kaiser (2001) gives an introductory overview of this approach to modeling. Arnold, Castillo and Sarabia (1992) provide some characterization results, especially for bivariate settings. Probably the most common conditional model in applications is formed from conditional normal distributions and is often called the *conditional autoregressive model* (CAR).

4.6.1 Conditional Autoregressive Model

Let $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ be a set of n random variables with locations $\boldsymbol{s}_i$; $i = 1, \ldots, n$, and let the full conditional densities of these random variables be given by,

$$f_i(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}) =$$

$$\frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left[\frac{-1}{2\tau_i^2}\left\{y(\boldsymbol{s}_i) - \mu(\{y(\boldsymbol{s}_j) : j \neq i\})\right\}^2\right], \tag{4.15}$$

where

$$\begin{aligned} \mu(\{y(\boldsymbol{s}_j) : j \neq i\} &\equiv E\left[Y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}\right] \\ \tau_i^2 &\equiv var\left[Y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}\right] \end{aligned}$$

Now, further model

$$\mu(\{y(\boldsymbol{s}_j) : j \neq i\}) = \alpha_i + \sum_{j=1}^{n} c_{i,j}\left\{y(\boldsymbol{s}_j) - \alpha_j\right\}, \tag{4.16}$$

subject to the conditions that $c_{i,j}\tau_j^2 = c_{j,i}\tau_i^2$, $c_{i,i} = 0$; $i, j = 1, \ldots, n$, and $c_{i,j} = 0$ unless $\boldsymbol{s}_j \in N_i$, the neighborhood of $\boldsymbol{s}_i$. It is common to take $\tau_i = \tau$ in this model so that the condition on the $c_{i,j}$ reduces to $c_{i,j} = c_{j,i}$.

Let C denote the $n \times n$ matrix with ij^th element $c_{i,j}$ and M the $n \times n$ matrix with diagonal elements τ_i^2; for constant conditional variance $M = \tau^2 I_n$ where I_n is the $n \times n$ identity matrix. As shown by Besag (1974) and Cressie (1993), the joint distribution of $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$ is then

$$\boldsymbol{Y} \sim Gau(\boldsymbol{\alpha}, (I_n - C)^{-1}M), \tag{4.17}$$

provided that the $n \times n$ matrix $(I_n - C)$ is invertible and the $n \times n$ matrix $(I_n - C)^{-1}M$ is positive definite. In (4.17) $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_n)^T$ and the elements

of $\boldsymbol{\alpha}$ may take on any real value. In order to use this model in practice, the $c_{i,j}$ must be subject to additional modeling to reduce the number of free parameters, such as $c_{i,j} = h(\boldsymbol{\eta})$ for a low-dimensional parameter $\boldsymbol{\eta}$. Then the need for $(I_n - C(\boldsymbol{\eta}))-1M$ to be positive definite results in restrictions on the parameter space of $\boldsymbol{\eta}$, that usually are affected by the way neighborhoods are defined.

A strength of conditional model specification is flexibility in modeling dependence. While dealing with directional dependence in a model for a continuous index random field becomes quite complicated, doing so for a Markov random field model is quite simple. We illustrate here for a conditionally specified model with Gaussian conditionals as in (4.15). Suppose we have defined locations as $\boldsymbol{s}_i = (u_i, v_i)$ where u_i denotes a horizontal coordinate and v_i a vertical coordinate on a regular grid. In addition, suppose that a four-nearest neighborhood structure is to be used so that, for $i = 1, \ldots, n$

$$N_i = \{\boldsymbol{s}_j : (u_j = u_i, v_j = v_i \pm 1) \text{ or } (u_j = u_i \pm 1, v_j = v_i)\} \tag{4.18}$$

A model with isotropic (unidirectional) dependence might take conditional densities as in (4.15) with conditional expectations as in (4.16) where

$$c_{i,j} = \begin{cases} \eta & j \in N_i \\ 0 & \text{otherwise} \end{cases}$$

Note that by definition $\boldsymbol{s}_i \notin N_i$, that is, locations are not neighbors of themselves.

Now, to build different north-south and east-west dependencies into this

model, we simply partition the neighborhoods as

$$\begin{aligned} N_{i,u} &= \{\boldsymbol{s}_j : u_j = u_i \pm 1, v_j = v_i\} \\ N_{i,v} &= \{\boldsymbol{s}_j : u_j = u_i, v_j = v_i \pm 1\} \\ N_i &= N_{i,u} \cup N_{i,v} \end{aligned} \tag{4.19}$$

Dependence parameters are then defined as

$$c_{i,j} = \begin{cases} \eta_u & j \in N_{i,u} \\ \eta_v & j \in N_{i,v} \\ 0 & \text{otherwise} \end{cases}$$

The modeling of dependence parameters $c_{i,j}$ can also incorporate time, functions of spatial locations, and auxiliary variables such as wind direction (e.g., Kaiser *et al.* 2002).

4.6.2 Models with Exponential Family Conditional Distributions

Consideration of conditionally specified models has largely been restricted to full conditional distributions in exponential families. Even among these, the conditional autoregressive model represents a very special situation in that the joint distribution is available in closed form. Typically, the joint distribution that corresponds to a set of specified full conditionals can only be shown to exist, and then identified up to an unknown constant of proportionality (constant with respect to the joint as a function of values for the random variables). Unfortunately, that constant is a function of the parameter values, which then makes estimation and inference difficult. Without presenting details, which can be found in Besag (1974) and Kaiser and Cressie (2000),

1. Certain conditions must be met by the full conditionals specified to ensure that a compatible joint exists, that is, a joint that has the conditionals that have been specified. Although there are others, a fundamental need is for the support of joint and marginal distributions to satisfy either Besag's (1974) *positivity condition* or the weaker *Markov random field condition* of Kaiser and Cressie (2000). The positivity condition, which is often satisfied, states that if $\Omega_i; \;\; i = 1, \ldots, n$ denote the supports of the marginal distributions of $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$, then the support of the joint distribution should be $\Omega = \Omega_1 \times \Omega_2 \times \ldots, \times \Omega_n$.

2. Given that the appropriate conditions are satisfied, one can construct, from those conditionals, a set of functions H_i, $H_{i,j}$, $H_{i,j,k}, ldots, H_{1,2,\ldots,n}$ and what is called the *negpotential function*,
$Q(\boldsymbol{y}|\boldsymbol{\theta}) =$

$$\sum_{1 \leq i \leq n} H_i[y(\boldsymbol{s}_i)|\boldsymbol{\theta}] + \sum_{1 \leq i < j \leq n} \sum H_{i,j}[y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)|\boldsymbol{\theta}] + \ldots + H_{1,2,\ldots,n}[y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)|\boldsymbol{\theta}]. \tag{4.20}$$

Under an assumption introduced by Besag (1974) called *pairwise-only dependence*, only the first two sums on the right hand side of (4.20) are used. There have been some investigations into the restrictions that result from this assumption (e.g., Tjelmeland and Besag 1998; Lee, Kaiser and Cressie 2001) but this remains a common assumption in applications because it greatly simplifies analysis.

3. The importance of (4.20) is that the joint distribution $f(\boldsymbol{y}|\boldsymbol{\theta})$ may then be recovered from the conditionals as,

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{\exp[Q(\boldsymbol{y}|\boldsymbol{\theta})]}{\int \exp[Q(\boldsymbol{y}|\boldsymbol{\theta})] \, d\mu(\boldsymbol{y})}, \tag{4.21}$$

where μ is either counting or Lebesque measure. What renders (4.21) difficult to work with is that the integral in the denominator is over the joint support of $f(\boldsymbol{y}|\boldsymbol{\theta})$, which is typically of high dimension. For example, in a model with n binary conditional distributions, the denominator of (4.21) is a sum over 2^n terms, and in even smallish spatial problems n will be in the hundreds.

4.6.3 Binary Conditionals Model

We present one additional example of a conditionally specified model based on binary conditional distributions. Assume that we have available locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, random variables $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$, and neighborhoods $N_1, \ldots, N_n$. Let $\boldsymbol{y}(N_i) = \{y(\boldsymbol{s}_j) : \ \boldsymbol{s}_j \in N_i\}$, and specify the full conditional probability distributions, for $0 < \theta_i < 1$ and $i = 1, \ldots, n$,

$$f(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : \ j \neq i\}) = f(y(\boldsymbol{s}_i)|\boldsymbol{y}(N_i)) = \theta_i^{y(\boldsymbol{s}_i)} \ (1-\theta_i)^{1-y(\boldsymbol{s}_i)}; \ \ y(\boldsymbol{s}_i) = 0, 1. \tag{4.22}$$

The first equality in (4.22) is due to the Markov assumption in Markov random fields, stating that the full conditional distribution is the same as the distribution conditioned on only neighbors.

The model is then fleshed out by taking, for $0 < \kappa_i < 1$ and $-\infty < \eta_{i,j} < \infty$, $i, j = 1 \ldots, n$,

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{j=1}^{n} \eta_{i,j} \left\{y(\boldsymbol{s}_j) - \kappa_j\right\}, \tag{4.23}$$

subject to the restrictions that $\eta_{i,i} = 0$ and $\eta_{i,j} = 0$ unless $\boldsymbol{s}_j \in N_i$, $i, j = 1, \ldots, n$.

The negpotential function (4.20) for this model is

$$Q(\boldsymbol{y}|\boldsymbol{\kappa},\boldsymbol{\eta}) = \sum_{1\le i\le j} y(\boldsymbol{s}_i)\left[\log\left(\frac{\kappa_i}{1-\kappa_i}\right) - \sum_{j\neq i}\eta_{i,j}\kappa_j\right] + \sum\sum_{1\le i<j\le n}\eta_{i,j}y(\boldsymbol{s}_i)y(\boldsymbol{s}_j). \tag{4.24}$$

Note that $\exp\{Q(\boldsymbol{y}|\boldsymbol{\kappa},\boldsymbol{\eta})\} \ge 0$ and, assuming positivity, $\exp\{Q(\boldsymbol{0}|\boldsymbol{\kappa},\boldsymbol{\eta})\} = 1$ so that the denominator of (4.21) is finite and positive.

Given the positivity condition (stronger than absolutely necessary), pairwise-only dependence, and the conditions of Theorem 3 in Kaiser and Cressie (2000), a joint distribution that has the conditionals (4.22) exists and is given by

$$f(\boldsymbol{y}|\boldsymbol{\kappa},\boldsymbol{\eta}) = \frac{\exp\{Q(\boldsymbol{y}|\boldsymbol{\kappa},\boldsymbol{\eta})\}}{\sum_{\boldsymbol{t}\in\Omega}\exp\{Q(\boldsymbol{t}|\boldsymbol{\kappa},\boldsymbol{\eta})\}}. \tag{4.25}$$

As previously mentioned, the joint support Ω contains 2^n elements and computation of (4.25) is computationally prohibitive. Estimation and inference, either non-Bayesian or Bayesian, will rely on procedures other than direct manipulation of (4.25).

To be useful, the number of free parameters must be reduced from the set containing $\kappa_i;\ \ i = 1,\ldots,n$ and $\eta_{i,j};\ \ i,j = 1,\ldots,n$, even with the restrictions imposed in (4.23). To accomplish this requires additional modeling of what are called the *large scale* and *small scale* model components. The large scale model component consists of a model for the κ_i, which are close to (leaving the precise meaning of this vague) the marginal expected values of the $Y(\boldsymbol{s}_i);\ \ i = 1,\ldots,n$ (Caragea and Kaiser 2009). Large scale model

components that suggest themselves immediately include

$$\begin{aligned} \kappa_i &= \kappa, \\ \kappa_i &= h(\boldsymbol{x}_i, \boldsymbol{\beta}), \\ g(\kappa_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta}. \end{aligned} \tag{4.26}$$

In (4.26) $\boldsymbol{x}_i$ represents a covariate value at the location $\boldsymbol{s}_i$; $i = 1, \ldots, n$, and $h(\cdot)$ and $g(\cdot)$ are known functions. The second line of (4.26) is essentially a nonlinear regression model, while the third line gives a more restricted form similar to generalized linear models. Note that the domain of $h(\cdot)$ is the real line and the range is $(0, 1)$ while the reverse is true for $g(\cdot)$. The small scale model component consists of modeling the dependence parameters, the $\eta_{i,j}$; $i, j = 1, \ldots, n$. Models for the small scale component are very flexible in the same way as for the parameters $c_{i,j}$ in the conditional autoregressive model described previously. In particular, constant $\eta_{i,j}$, different values for different partitions of the neighborhood, or modeling in terms of covariates are all possible.

Chapter 5

Models for Networks

This section presents an introduction to probability models for use with networks, or what are frequently called random graph models. Applications of networks or mathematical graphs appear in a wide variety of disciplines, such as sociology, economics, political science, biology, animal behavior, and others (e.g., Lusseau 2003, Sporns *et al.* 2004, Hoff and Ward 2005, Simpson, Hayasaka and Laurrienti 2011). Because of this, research into graphs has been conducted in many fields, including mathematics, computer science, and statistics, as well as applied disciplines. Our focus here is on the formulation and use of stochastic models, so there is a large portion of the overall topic of graphs and the use of graphs in representation and analysis of problems that we will not cover.

5.1 Graphs and Approaches to the Analysis of Graphs

Before we approach the topic of random graph models, we briefly review what is meant by a mathematical graph, and how the use and analysis of graphs has been approached by mathematicians, computer scientists, and statisticians. A mathematical graph is a structure that consists of a set of nodes (or vertices) and a set of edges (or joins) that connect some or all possible pairs of nodes. In applications, nodes are typically associated with people, animals, places or other objects. In social networks, nodes are often individual people. In biology, nodes may be connected with genes, cells, portions of the brain, animals, or components of ecosystems. In communications science, nodes may be cell phone towers or transformers or transfer stations on a power grid. Edges typically represent some type of relation or relationship between nodes. If nodes represent physicists, edges may represent joint authorship of papers. If nodes represent brain cells, edges may represent activation from a single stimulus. If nodes represent web domains, edges may represent direct links. Edges may be either undirected or directed. Undirected edges represent bilateral relations, such as friendship, while directed edges may represent causation such as parent to offspring, or hierarchical organization such as links from web site A to web site B. Graphs typically contain either all undirected or all directed edges.

Mathematicians are frequently interested in what is called graphical enumeration, the problem of counting how many graphs are possible that satisfy certain conditions. For example, how many undirected graphs can be constructed from n distinct (or labeled) nodes; turns out this is $2^{n(n-1)/2}$,

where $n(n-1)/2$ is the number of potential edges among n distinct nodes. Other problems are categorized as subgraph identification, such as finding the largest subgraph of a given graph such that all nodes in the subgraph are joined by edges to all other nodes in the subgraph (called a complete subgraph).

Computer scientists often use graphs to represent data structures, which can lead to efficient ways to store and retrieve information, or to represent networks of communication, or computational devices. They are interested in the development of algorithmic models of graphs, that is, in the development of algorithms that can produce graphs with certain properties. For example, the number of edges connected to a given node is called the degree of that node. The empirical distribution of degrees for a graph with n nodes is called the degree distribution. It turns out that different types of degree distributions correspond to graphs with different structures, such as many small complete subgraphs (or clusters) or the occurrence of a few super-nodes that have a much higher degree than the other nodes in a graph. These, and other, structures in a graph can be related to important real-world phenomena, such as how robust a communication network is to failure of nodes. While the algorithmic models developed may contain some random decisions, they typically have the form of games in which nodes and edges are added or subtracted from a graph according to a set of rules (although these rules can become enormously complex).

Statisticians are typically interested in models for graphs that result in a probabilistic mechanism for generating graphs. These models are usually formulated by defining a set of binary random variables that correspond to the occurrence of potential edges for a graph with a fixed number of nodes. If there are n nodes there are $k = n(n-1)/2$ possible edges. A model

for the joint distribution of these random variables then gives a likelihood for any possible graph that could be generated by the model. The analysis of an observed graph then proceeds in the same manner as the analysis of nearly any set of data. The observed graph with n nodes contains a subset of the k possible edges. The random variables that correspond to these edges are taken as having an observed value of 1, and the variables that correspond to the other potential edges are taken as having an observed value of 0. Using these data and the likelihood afforded by the model, parameters are estimated and inferences made using either non-Bayesian or Bayesian approaches. In more complicated models for random graphs, nodes may be assumed to result from a probabilistic mechanism as well as edges. This may be the case, for example, if nodes are not fixed and correspond to entities with spatial position. Nodes might then be assumed to arise from a spatial point process. To date, these more complex possibilities have received relatively little attention in the literature.

5.2 Graph Properties and Graph Topology

A great deal of attention has focused on identifying and defining various properties or features of graphs. Some concepts are well defined. For example, a graph is called connected if there is a path of edges that one can traverse to get from any node to any other node, or is said to be bipartite if there is a partition of the nodes into two sets such that edges only join nodes from one set with nodes from the other. Other concepts are less definitive. The concept of what are called scale free graphs is that parts of the network (subgraphs) have a similar structure to the whole graph. This is loosely quantified by defining scale free graphs as those having power law degree

distributions, which lead to a relatively small number of nodes with high degrees and a larger number of nodes with smaller degrees. But an exact quantification of this concept is elusive, because how close to a power law degree distribution is close enough to claim a scale free property? Even more vague is the idea of a small world property for graphs. The concept here is that the number of steps needed to get from most nodes to most other nodes is small. Graphs that would be said to have a small world property tend to contain many connected subgraphs and also many hubs, which are nodes with relatively high degrees. There are any number of indices that have been proposed as possibly quantifying the concepts of small world and scale free properties, but none have gained anything near to universal acceptance.

There are various configurations of graph topology that get used to quantify features of realized graphs, usually by counting the number of occurrences in observed or realized graphs from a model. These include the number of edges, the number of triangles, and the number of $k-$stars (nodes that have edges with k other nodes, none of which have edges between them). These ideas have spawned elaborate extensions, including what are called "alternating $k-$stars", "alternating $k-$triangles", and "alternating $k-$two paths" (Snijders, Pattison, Robins and Handcock 2006). Substantive interpretation of these features of graph topology is somewhat elusive, aside from the number of triangles, which is connected with the notion of transitivity. Transitivity is the idea that friends tend to be friends. In other words, the idea is that two friends of yours are more likely to be friends of each other than are two randomly chosen people. Thus, in a graph in which edges represent friendships among people, there should be more triangles than two-stars. What is called the clustering index or clustering coefficient is, in fact, some ratio of the number of triangles to potential triangles. There are a number of

formulas for this index that do not all agree with each other (Bollobas and Riordan 2003).

5.3 Random Graph Models

Models for random graphs may be considered to consist largely of four classes, Erdos-Renyi models, block and covariate models, exponential random graph models, and local structure graph models. We briefly describe each of these in turn, but first define the situation to be considered. We will assume there exists an observed graph containing n nodes and some number of edges greater than 0 and less than $k = n(n-1)/2$. Index possible edges among the n nodes as $i = 1, \ldots, k$. Let $\boldsymbol{s}_i$ and $\boldsymbol{u}_i$ denote the two nodes that would be joined by edge i; $(\boldsymbol{s}_i, \boldsymbol{u}_i) \in \mathcal{P}$ where $\mathcal{P}$ denotes the set of unique pairs of $\{1, \ldots, n\}$. Define random variables, for $i = 1, \ldots, k$,

$$Y_i = \begin{cases} 1 & \text{if edge } i \text{ is realized} \\ 0 & \text{otherwise.} \end{cases}$$

5.3.1 Erdos-Renyi Graphs

The simplest random graph model for a set of n fixed nodes takes the random variables $Y_1, \ldots, Y_k$ to be independent and identically distributed with common probability mass function, for some $0 < p < 1$,

$$Pr[Y_i = y_i] = f(y_i|p) = p^{y_i}\,(1-p)^{1-y_i}; \quad y_i = 0, 1. \tag{5.1}$$

This is called the Erdos-Renyi graph model, named after the two mathematicians that suggested it in the 1950s. In fact, what is now called the Erdos-Renyi model was long (and still is by some) considered *the* model for random graphs. That is, the term random graph referred to graphs generated

by this model. In this context, any number of questions about the expected number of graph structures can be approached through what are essentially combinatorial methods, as the number of possible realizations of an Erdos-Renyi model is 2^k, and the number of those that meet certain criteria can often be counted as well.

5.3.2 Block and Covariate Models

In block models, attributes of the nodes are used to create groups. The simplest is to consider nodes of different colors, red, blue, white, and so forth. Blocks of edges are created based on the groups of nodes they connect; red-red, red-blue, red-white, blue-blue, etc. Random variables within blocks are assumed to be independent and identically distributed with a block-specific probability of realization, and with independence among blocks as well. This is nothing more than partitioning the possible pairs of nodes into independent blocks and assigning an Erdos-Renyi model to each block.

In what are being called covariate models here, assume there are auxiliary variables associated with each Y_i (each possible edge) in the form of covariates $\boldsymbol{x}_i^T = (x_{i,1}, \ldots, x_{i,p})$. We wish to model the probabilities of edge realization as functions of these covariates. Note that block models can be considered to be covariate models in which covariates identify block membership. If there are B blocks, one way to accomplish this is to let $\boldsymbol{x}_i$ be a $B-$vector of values with one element equal to 1 and $B-1$ elements equal to 0.

Block and covariate models then become models for independent binary responses, such as basic generalized linear models with binary random components. In particular, given choice of a suitable link function $g(\cdot)$, marginal

probability mass functions (5.1), and $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$,

$$g(p_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}; \quad i = 1, \ldots, k.$$

5.3.3 Exponential Random Graph Models

To formulate an exponential random graph model, we no longer assume that the random variables $Y_i;\ i = 1, \ldots, k$ are independent. Let $\boldsymbol{s}(\boldsymbol{y}) = (s_1(\boldsymbol{y}), \ldots, s_q(\boldsymbol{y}))^T$ denote a $q-$vector of statistics that depend on an observed (or possible) graph. Counts of features in graph topology such as the number of edges, the number of 2−stars, and/or the number of triangles are often taken as the components of $\boldsymbol{s}(\boldsymbol{y})$. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$ be a vector of q parameters. The joint probability mass function of $Y_1, \ldots, Y_k$ is then specified as

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{\exp\left[\boldsymbol{s}(\boldsymbol{y})^T \boldsymbol{\theta}\right]}{\sum_{\boldsymbol{t} \in \Omega} \exp[\boldsymbol{s}(\boldsymbol{t})^T \boldsymbol{\theta}]}, \tag{5.2}$$

where Ω is the $k-$fold Cartesian product of $\{0, 1\}$. The distribution (5.2) is said to be in Gibbsian form, because it has a form similar to a distribution of particle states in statistical mechanics called a Gibbs or Boltzmann distribution. No other mystical connections. Notice that the normalizing constant in the denominator of (5.2) is a sum over 2^k terms, and is usually computationally infeasible to evaluate. Thus, estimation of the parameters $\boldsymbol{\theta}$ on the basis of an observed graph must rely on something other than direct evaluation of the likelihood.

Exponential random graph models are often lauded as being able to accommodate nearly any pattern of dependencies among the random variables that represent potential edges. This is true, but determining what that pattern is for a given model is difficult, if not impossible. That is, these models

accommodate, but do not identify or control, the dependence structure that may exist among the random variables $Y_1, \ldots, Y_k$.

5.3.4 Local Structure Graph Models

A more recent development has been what are called Local Structure Graph models (Casleton, Nordman and Kaiser 2016). In some ways these are not so much a different type of graph model as they are a principled way to construct an exponential random graph model. These are called local structure graph models because they model the probability distribution of the edge random variables as a binary Markov random field. That is, in these models, we take the random variables $Y_1, \ldots, Y_k$ to be associated with k locations in a Markov random field. Given a definition of neighborhoods for each of these locations and values for the random variables at neighboring locations, $\boldsymbol{y}(N_i)$; $i = 1, \ldots, k$, the full conditional probability mass functions for $Y_1, \ldots, Y_k$ are specified to be

$$f(y_i|\boldsymbol{y}(N_i), \theta_i) = \theta_i^{y_i} (1 - \theta_i)^{1-y_i}; \quad y_i = 0, 1, \tag{5.3}$$

where

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{j \in N_i} \eta_{i,j}(y_j - \kappa_j), \tag{5.4}$$

which is the model of expressions (4.22) and (4.23) for random variables written as Y_i rather than $Y(\boldsymbol{s}_i)$. The same conditions that are required for a binary conditionals Markov random field model to correspond to a joint distribution that may be identified through the use of a negpotential function apply to local structure graph models. Under these conditions, the joint probability mass function is then given by (4.25) which will be seen to also be a Gibbsian form of distribution. As for exponential random graph

models, it will not be possible to compute the value of the joint probability mass function for a local structure graph model. This is again because of the denominator in (4.25). As a result, fitting a local structure graph model to data contained in an observed graph will require methods that do not rely on direct evaluation of likelihoods.

The value of local structure graph models is that they offer a mechanism through which to understand and control dependencies in edge formation in line with whatever substantive processes underlie an observed graph. Dependencies in a local structure graph model are governed by the definition of neighborhoods and the parameters $\eta_{i,j}$ in (5.4). Thus, although the joint distribution has a form similar to exponential random graph models, we consider local structure graph models to be a different class of models. For exponential random graph models, modeling takes place at the level of the joint distribution, and the conditionals are whatever may result. For local structure graph models, modeling takes place locally, at the level of the conditional distributions, and the joint is whatever may result (subject to conditions that ensure it exists). Since conditional dependencies occur at the level of the conditional distributions, the approach of local structure graph models offers greater opportunity to model these dependencies in a meaningful way within the context of a given problem.

To avoid confusion, there is one additional issue with local structure graph models that should be discussed at this point. The reader may be aware that any Markov random field defines a mathematical graph. The locations of the random field are nodes in the graph, and any two locations that are neighbors of each other are joined by an edge. In a local structure graph model, edges of the original graph are locations in a Markov random field and are thus nodes in the Markov random field *graph of edges*. Edges in the graph of edges

do not appear in the original graph. The original graph (or data graph) is assumed to be a realization of a random graph model. The graph of edges is a fixed graph that is defined by possible edges in the original graph and the definition of neighborhoods in the graph of edges.

Chapter 6

Complex Hierarchical Models

In Statistics 520 we introduced hierarchical models in which parameters of a data model were modeled as random variables in their own right. That is, we had the basic structure:

Data Model: $f(\boldsymbol{y}|\boldsymbol{\theta})$
Parameter Model: $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$
Prior: $\pi(\boldsymbol{\lambda})$.

In what we have seen to this point, the components of $\boldsymbol{\theta}$ were always taken to be independent. We now wish to make use of the same ideas, but in terms of structures that place greater complexity (and have greater flexibility) on the Parameter Model component of this hierarchical structure. It is natural to consider hierarchical models in the context of Bayesian analysis, and we will do so here for the most part. It is important to keep in mind, however, that we could formulate complex hierarchical models without the final stage prior. So, while hierarchical models lend themselves to Bayesian analysis, they are not necessarily Bayesian Models.

Throughout this chapter we will emphasize a point made in the first chapter of the Stat 520 notes, that a key to construction of a meaningful model is to know how the model is representing the scientific mechanism of interest. This becomes critical in the use of hierarchical models because it is possible to construct many mathematical and probabilistic structures. Without careful consideration of how those structures are representing underlying scientific mechanisms it is easy to arrive at mathematically elaborate, yet substantively shallow, stochastic conceptualizations of problems. Since many hierarchical models with complex structure involve variables that fail to be independent, connecting models with physical and biological problems often centers on how dependencies are incorporated in models.

6.1 Extending a Beta-Binomial Model

This section discusses a number of models within the general category of beta-binomial responses that illustrate two of the points made in the introductory comments to this chapter, that analysis of hierarchical models is not constrained to Bayesian methods, and that a variety of models can be constructed by incorporating dependencies in different ways. In Stat 520 we saw a Beta-Binomial model several times. The model was formulated as follows. Let $Y_1, \ldots, Y_n$ denote independent binomial random variables with binomial sample sizes $m_1, \ldots, m_n$ considered fixed and known (non-random) and parameters θ_i. The probability mass functions for the Y_i are, for $0 < \theta_i < 1$ and $i = 1, \ldots, n$,

$$f_i(y_i|\theta_i) = \frac{m_i!}{y_i!(m_i - y_i)!}\,\theta_i^{y_i}\,(1-\theta_i)^{m_i - y_i}; \quad y_i = 0, 1, \ldots, m_i. \tag{6.1}$$

Now assume that the θ_i are independent and identically distributed following

a common beta distribution with probability density function, for $0 < \alpha$, $0 < \beta$, and $i = 1, \ldots, n$,

$$g(\theta_i|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta_i^{\alpha-1}\,(1-\theta_i)^{\beta-1}; \quad 0 < \theta_i < 1. \tag{6.2}$$

The joint data model is then

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\theta_i), \tag{6.3}$$

and the marginal model is

$$h(\boldsymbol{y}|\alpha, \beta) = \prod_{i=1}^{n} h_i(y_i|\alpha, \beta), \tag{6.4}$$

where

$$h_i(y_i|\alpha, \beta) = \int_0^1 f_i(y_i|\theta_i)\, g(\theta_i|\alpha, \beta)\; d\theta_i. \tag{6.5}$$

6.1.1 Random Binomial Sample Sizes

One application of a beta-binomial model we covered in Stat 520 was the analysis of toxicity tests conducted with fish of the genus *Gambusia* in the Central Valley of California. The study design was that gravid female fish were captured in two areas and held in the laboratory until birth (*Gambusia* give live birth). The responses were the number of live young among "litters" of fish of varying sizes. It was pointed out then that the size of the litters were actually not fixed, and could be considered random variables along with the responses of numbers of live young. The data from this example are presented in Table 6.1.

We can extend our beta-binomial model by taking the total number of young born to female $i = 1, \ldots, n$ to be connected with random variables $M_1, \ldots, M_n$, and assigning these distributions of their own. Because adult

Volta NWR		San Luis Drain	
No. Live	Total	No. Live	Total
28	28	36	40
31	31	33	34
9	11	27	28
68	68	4	18
32	32	13	18
37	37	22	26
19	19	20	24
17	17	20	22
26	26	38	41
52	52	21	21
30	30	20	25
46	46	26	27
0	9	7	16
47	51	18	18
22	22	23	25
18	19		
62	64		
4	5		

Table 6.1: Number of live and total young born to *Gambusia* collected from the Volta National Wildlife Refuge and the San Luis Drain in Central California.

fish were housed individually in this study, it is reasonable to assume that the M_i are independent. The total number of young are integer-valued variables,

and one might consider assuming that the M_i has a Poisson distribution. A question is whether the collection of M_i should be considered identically distributed or not. Two model structures seem immediately possible. One takes $M_1, \ldots, M_n$ to be independent and identically distributed having common probability mass function

$$f_M(m_i|\lambda) = \frac{1}{m_i!}\,\lambda^{m_i}\,\exp(-\lambda); \quad m_i = 0, 1, \ldots \tag{6.6}$$

Another possibility would be to assume the M_i are independent with probability mass funcitons, for $i = 1, \ldots, n$,

$$f_M(m_i|\lambda_i) = \frac{1}{m_i!}\,\lambda_i^{m_i}\,\exp(-\lambda_i); \quad m_i = 0, 1, \ldots \tag{6.7}$$

and then take $\lambda_1, \ldots, \lambda_n$ to be independent and identically distributed with common gamma probability density function,

$$g(\lambda_i|\eta, \psi) = \frac{\eta^{\psi}}{\Gamma(\psi)}\,\lambda_i^{\psi-1}\,\exp(-\eta\lambda_i); \quad \lambda_i > 0. \tag{6.8}$$

One question, then, is whether we want to model the litter sizes as identically distributed Poisson variables, or assign them a mixture model in a manner similar to the binomial response variables. At least a part of the motivation for using a beta-binomial mixture model with these data was the variability in observed proportions in the data. Consider the data in Table 6.1 from the San Luis Drain. A test for homogeneity of binomial proportions (e.g., Snedecor 1967, p. 420) yields a $p-$vlaue of 1.68×10^{-13}. We could conduct a crude test to see if a common Poisson distribution seems reasonable to model the litter sizes by dividing the range of the data into bins, computing the estimated probabilities of those bins from a fitted Poisson model, and then conducting a χ^2 test. The maximum likelihood estimator of the Poisson parameter λ from a model having independent marginal distributions (6.6)

is $\bar{m} = (1/n)\sum m_i$. A simple binning strategy for the data from the San Luis Drain in Table 6.1 would be $10-19$, $20-29$, $30-39$ and $40-49$. We should also include bins of < 10 and > 49 because the fitted Poisson model will have some probability in those regions. Denote bins for values < 10 as B_1, up to values of > 49 as B_6. To make the procedure amenable to the use of other than 6 bins, let the bins be $B_1, \ldots, B_k$. Denote the probabilities that a Poisson random variable with parameter $\bar{m}$ falls into bin B_j as P_j; $\;j = 1, \ldots, k$. The χ^2 test statistic is then

$$T = \sum_{j=1}^{k} \frac{1}{P_j} \left[(1/n) \sum_{i=1}^{n} I(m_i \in B_j) - P_j \right]^2 . \tag{6.9}$$

Conducting this procedure for the data corresponding to the San Luis Drain in Table 6.1 results in a test statistics of $T = 3.687$ and an associated $p-$value (with 5 degrees of freedom) of $p = 0.5953$, which provides no evidence that the distribution of litter sizes differs from what might be expected from a common Poisson distribution.

We would then model the observable random variables $Y_1, \ldots, Y_n$ as independent with probability mass functions (6.1), the observable random variables $M_1, \ldots, M_n$ as independent with common probability mass function (6.6), and the binomial parameters $\theta_1, \ldots, \theta_n$ as independent with common probability density (6.2). Absent any scientific indication that the litter sizes should be related to the probability of live birth, we might assume independence between the Y_i, M_i, and θ_i. The parameters to be estimated in this model are then α, β, and λ. If there were a plausible mechanism that would cause the proportions of live young to be related to the litter sizes, this would most likely focus on the physiological processes of implantation and embryonic development. Under this situation we would want to formulate a bivariate model for the (Y_i, M_i) pairs. We will defer discussion of this

possibility until later in this section.

6.1.2 Dependence Between Data Model Paramters 1

In some problems for which we might consider application of a beta-binomial model, it is reasonable to consider the number of binomial trials as random, and to suppose the number of trials might be related to the realized probability of success. Consider, for example, modeling the number of three point shots taken (and made) by a college basketball player. Players have an overall probability of success at three point shots, but this probability varies substantially among games. It would not be unreasonable to model success at three point shooting across games for a given player with a beta-binomial model. At the same time, it is also reasonable to suppose that the number of three point shots taken in a game may well depend on the probability of success that is being realized in that game, either as a matter of confidence of the player or instructions from the coach. In this situation we might like to develop a beta-binomial model that contains dependence between the binomial probabilities and the number of binomial trials. Zhu, Eickhoff and Kaiser (2003) developed such a model, although their application involved the number of capture attempts made by foraging green herons. In our current notation, the model of Zhu *et al.* took response variables $Y_1, \ldots, Y_n$ to be conditionally independent with probability mass functions (6.1), binomial parameters $\theta_1, \ldots, \theta_n$ to be independent and identically distributed with probability density functions (6.2), and number of binomial trials $M_1, \ldots, M_n$ to be conditionally independent having Poisson distributions truncated at zero. Truncation at zero was appropriate in the application under consideration by Zhu *et al.* but might not be in all applications. Specifically, the data

model consists of two parts, both for $i = 1, \ldots, n$,

$$f(y_i, |m_i, \theta_i) = \frac{m_i!}{y_i!\,(m_i - y_i)!}\,\theta_i^{y_i}\,(1-\theta_i)^{m_i - y_i}; \quad y_i = 0, 1, \ldots, m_i, \tag{6.10}$$

where $0 < \theta_i < 1$, and

$$f_M(m_i|\lambda_i) = \frac{1}{m_i!}\,\lambda_i^{m_i}[\exp(\lambda_i) - 1]^{-1}; \quad m_i = 1, 2, \ldots \tag{6.11}$$

where $0 < \lambda_i$. To reflect a "decision" process that relates the number of attempts to the realized binomial parameters, these authors took, for $i = 1, \ldots, n$,

$$\lambda_i = \exp\left\{\eta_0 + \eta_1 \log\left(\frac{\theta_i}{1-\theta_i}\right)\right\}, \tag{6.12}$$

where $-\infty < \eta_0 < \infty$ and $-\infty < \eta_1 < \infty$. Notice here that dependence is incorporated in the model through a specified relation between parameters of the two parts of the data model, $f(y_i|\theta_i, m_i)$ and $f_M(m_i|\lambda_i)$. Only the θ_i were explicitly modeled, as independent and identically distributed random variables with common density (6.2) as in a typical beta-binomial model. Thus, while the data model parameters for both portions of the model are random, the λ_i are functionally related to the θ_i, $i = 1, \ldots, n$. The consequences of this are that the distributions of the M_i are specified conditionally on the θ_i (with parameters η_0 and η_1), and the left hand side of (6.11) can be written as $f_M(m_i|\theta_i, \eta_0, \eta_1)$. The complete marginal likelihood for the model is then given as

$$L(\alpha, \beta, \eta_0, \eta_1) = \prod_{i=1}^{n} \int_0^1 f(y_i, |m_i, \theta_i)\, f_M(m_i|\theta_i, \eta_0, \eta_1)\, g(\theta_i|\alpha, \beta)\; d\theta_i. \tag{6.13}$$

In Zhu *et al.*(2003) the parameters of this model were estimated using a version of the EM algorithm. If we wished to enact a Bayesian analysis we would need to assign a prior to the parameter vector $(\alpha, \beta, \eta_0, \eta_1)^T$.

6.1.3 Dependence Between Data Model Parameters 2

Another example we discussed in Stat 520 was a situation involving response random variables that were connected with the number of misdemeanor interactions (per month or quarter) between police and adolescent males in small towns. We used this example to suggest the possibility of a Poisson-gamma mixture as a possible model, and were hypothetically interested in whether there was a difference between towns that had civic recreational facilities, such as a city gym, and those that did not. Now consider a hypothetical extension of this example, in which another observed variable is the number of those previously identified interactions that result in formal criminal charges. This is an example of a situation in which we might consider the use of a combined beta-binomial and Poisson-gamma model, similar to what was discussed previously in the example with *Gambusia spp.* There, we did wonder just a bit about whether litter size and proportion of live young might be related, which could be due to a physiological mechanism that operates within an individual adult female. But here, our focus is on processes that are operational at a higher level than the data model – the tendencies for certain events to occur in small towns with and without after-school recreational facilities. We might also wonder, in this example, whether there is some type of relation between number of incidents and the proportion of those that result in criminal charges. Certainly, both the police forces and the cit attorney's offices (responsible for filing criminal charges) have certain "personalities" that can vary from town to town. But, there is no guarantee that they do so in a consistent manner. That is, possible dependence between the two responses of number of police-adolescent interactions and number of criminal charges is less direct (in terms of mechanism) than the possible de-

pendencies in either of the two previous examples. We might, then, consider a model for one group (i.e., towns with or without recreational facilities) in which the response variables of number of police-adolescent interactions $M_1, \ldots, M_n$, have conditionally independent distributions, for $0 < \lambda_i$,

$$f_M(m_i|\lambda_i) = \frac{1}{m_i!}\,\lambda_i^{m_i}\,\exp(-\lambda_i); \quad m_i = 0, 1, \ldots \tag{6.14}$$

Assume that the response variables $Y_1, \ldots, Y_n$ are also conditionally independent and have probability mass functions, for $0 < \theta_i < 1$,

$$f(y_i|m_i, \theta_i) = \frac{m_i!}{y_i!\,(m_i - y_i)!}\,\theta_i^{y_i}\,(1-\theta_i)^{m_i - y_i}; \quad y_i = 0, 1, \ldots, m_i. \tag{6.15}$$

Here, we want to assign the data model parameter pairs (λ_i, θ_i) a joint distribution that allows dependence between the two. This could be accomplished in several ways:

1. Define transformations of λ_i and θ_i that each have range equal to the entire real line, such as $\psi_i = \log(\lambda_i)$ and $\phi_i = \log(\theta_i/(1-\theta_i)$. Then model the joint distributions of independent pairs (ψ_i, ϕ_i) with a common bivariate normal distribution (within one comparison group, either with or without recreational facilities).

2. Model the (λ_i, θ_i) as a marginal and conditional pair, $g_1(\theta_i|\lambda_i)$ and $g_2(\lambda_i)$. Take $\lambda_1, \ldots, \lambda_n$ as independent with common distribution

$$g_2(\lambda_i|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\lambda_i^{\alpha-1}\,\exp(-\beta\,\lambda_i); \quad 0 < \lambda_i. \tag{6.16}$$

   Then let $\theta_1, \ldots, \theta_n$ be independent with beta distributions that have been given a mean value parameterization (μ_i, ϕ),

$$g_1(\theta_i|\mu_i, \phi) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu_i)\,\Gamma(\phi(1-\mu_i))}\,\theta_i^{\phi\mu_i - 1}\,(1-\theta_i)^{\phi(1-\mu_i)-1}; \quad 0 < \theta_i < 1. \tag{6.17}$$

In (6.17), the parameters μ and ϕ are related to the typical α, β as $\mu = \alpha/(\alpha+\beta)$ and $\phi = \alpha + \beta$. Finally, relate the μ_i to the λ_i as

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1\,\lambda_i; \quad i = 1, \ldots, n,$$

and note that this is really nothing more than a beta regression model in which we allow the expected values μ_i to vary but hold the dispersion parameter ϕ constant across observations. The left hand side of (6.17) may now be written as $g_1(\theta_i|\lambda_i, \beta_0, \beta_1, \phi)$. The marginal likelihood for observable variables is then

$$L(\beta_0, \beta_1, \phi, \alpha, \beta) =$$

$$\prod_{i=1}^{n} \int_0^\infty \int_0^1 f(y_i|m_i, \theta_i)\, f_M(m_i|\lambda_i)\, g_1(\theta_i|\lambda_i, \beta_0, \beta_1, \phi)\, g_2(\lambda_i|\alpha, \beta)\; d\theta_i\; d\lambda_i. \tag{6.18}$$

Computation of the integrals in (6.18) would likely prove extremely difficult. A Bayesian approach would focus on the joint posterior

$$p(\beta_0, \beta_1, \phi, \alpha, \beta, \{\theta_i : i = 1, \ldots, n\}, \{\lambda_i : i = 1, \ldots, n\}|\boldsymbol{y}, \boldsymbol{m}), \tag{6.19}$$

and integration over the values of θ_i and λ_i would be accomplished through simulation of Markov chains.

3. A third approach toward modeling the (θ_i, λ_i) as independent pairs, with each pair having a bivariate distribution, would be to assign conditional distributions, $p_1(\theta_i|\lambda_i, \psi_1)$ and $p_2(\lambda_i|\theta_i, \psi_2)$. This is then a Markov random field model in which the neighbor of θ_i is λ_i and the neighbor of λ_i is θ_i, for $i = 1, \ldots, n$. Care would need to be exercised in this attempt to ensure that a compatible joint distribution exists. Estimation would require methods that we have not yet covered, under either non-Bayesian or Bayesian approaches.

6.1.4 Dependence Between Observable Variables

The previous two subsections have considered ways that the values of data model parameters θ_i and λ_i might be related to each other, functionally or stochastically. We return now to the possibility mentioned at the close of Section 6.1.1, that the observable variables Y_i and M_i might be directly related, due to the influence of a shared scientific mechanism. This is probably the most difficult type of dependence to model in hierarchical models because it requires the use of a joint data model rather than relying on conditional independence to formulate the data model. The choices are few at this stage of our understanding. One could transform the observable variables in a way that would allow the use of a multivariate (in this case, bivariate) normal distribution in the model. One could attempt to formulate a Markov random field model consisting of conditional distributions of different forms (here, binomial and Poisson). Or, one could attempt to use the ideas of copulas to develop a joint data model (e.g., Nelsen 2006). If time allows we may cover an introduction to the use of copulas in model formulation later in the course. The idea can be relatively easily understood relative to simulation. Very briefly, suppose we wish to simulate values of random variables X and Y, with marginal distribution functions $F(x)$ and $G(y)$ but such that X and Y are correlated. We could generate values (w, z) from a bivariate normal with specified correlation, then use the probability integral transform to create $u = \Phi(w)$ and $v = \Phi(z)$, which will both have marginal uniform distributions, but still be correlated. Finally, we would take $x = F^{-1}(u)$ and $y = G^{-1}(v)$ to arrive at two correlated values with marginal distributions F and G. There are, of course, some additional particulars to be considered.

6.1.5 Take Away Points

Although this section dealt with models in which we wanted dependence to be modeled between pairs of random variables with independence assumed among pairs, the basic ideas presented can also be used with larger numbers of potentially dependent variables, even if those variables represent the same observable quantity on different sampling units, such as in a spatial or spatio-temporal problem. Combining the examples presented in this section with some things we already knew, some important summary points are:

1. Hierarchical models offer great flexibility in accounting for variability in responses of one or more types, as well as in representing dependencies within a problem. Because of this, it is important to design a hierarchical model so that the levels of the hierarchy match, to the best of our knowledge, various scientific mechanisms involved in producing the observed data.

2. Dependencies may be effectively introduced through a combination of conditionally independent data distributions and parameters that either

   (a) have common values for a number of observable random variables (e.g., a classic random effects model).

   (b) are functionally related, having probability distributions that are induced through a combination of assignment and transformation (e.g., the model of Section 6.1.2).

   (c) modeled with a joint distribution (e.g., the models of Section 6.1.3).

3. There are situations in which the most scientifically plausible way to incorporate dependence is through the data model directly. These problems present challenging situations in terms of developing both substantively pleasing and mathematically useful models.

6.2 Hierarchical Models and Scientific Processes

We are now prepared to examine situations for which hierarchical modeling is probably the most profitable and natural approach available, and covers a large portion of problems in which observations are collected at points in time and space. The development presented here owes a great deal to Berliner 1996; Berliner, Millif and Wikle 2003; Cressie *et al.* 2009; Cressie and Wikel 2011. The fundamental idea is straightforward, to separate the process of observation and the scientific process of interest into levels of a hierarchy.

6.2.1 Basic Hierarchical Structures

Let square brackets denote distribution, so that $[X]$ is the distribution of a random variable X, $[Y|X]$ is the distribution of Y given X, and so forth. Consider modeling observable random variables $\boldsymbol{Z}$ that are the result of some physical or biological process we will represent as $\boldsymbol{Y}$. Given the process $(\boldsymbol{Y})$ we obtain a realization through a measurement or observation process, which may be modeled as a distribution that depends on some additional parameters $[\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}]$. The process itself is modeled in terms of random variables $\boldsymbol{Y}$, assigned a distribution that also may depend on some additional param-

eters $[\boldsymbol{Y}|\boldsymbol{\lambda}]$. Finally, in a Bayesian context, we will assign the collection of parameters a prior distribution $[\boldsymbol{\theta}, \boldsymbol{\lambda}]$. Note here that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ may, or may not, consist of mutually exclusive components, and we use the notation $(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{\theta} \cup \boldsymbol{\lambda}$. In this framework $\boldsymbol{Z}$ corresponds to observable quantities, $\boldsymbol{Y}$, $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ to unobservable quantities. The fundamental posterior for making inference about any unobservable quantity is then

$$[\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{Z}] \propto [\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}]\,[\boldsymbol{Y}, |\boldsymbol{\lambda}]\,[\boldsymbol{\theta}, \boldsymbol{\lambda}]. \tag{6.20}$$

Expression (6.20) follows from $[\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}] = [\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\lambda}]$ and $[\boldsymbol{Y}|\boldsymbol{\lambda}] = [\boldsymbol{Y}|\boldsymbol{\lambda}, \boldsymbol{\theta}]$.

Multiple Data Sources

The basic hierarchical structure defined previously is

- Data Model: $[\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}]$
- Process Model: $[\boldsymbol{Y}|\boldsymbol{\lambda}]$
- Prior: $[\boldsymbol{\theta}, \boldsymbol{\lambda}]$

Now, suppose we have a problem in which several data sources, say $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_k$, all depend on the process $\boldsymbol{Y}$. For example, bathymetry is important in how a marine oil spill will expand, and thus in determining where resources (people, booms, etc.) should be deployed to best control environmental damage. In a given region, we may have some measurements of ocean depth from sonar readings taken by marine vessels. These observations are likely to be reasonably high precision, but rather sparse in coverage. We may also have less precise, but more complete spatial coverage, from an aerial LIDAR (Light Detection and Ranging) data source. Both of these depend on the process of interest, ocean depth, but neither are measured without error.

Assuming the data sources are conditionally (given the process $\boldsymbol{Y}$) independent, we can develop a data model as

$$[\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_k|\boldsymbol{Y}, \boldsymbol{\theta}] = [\boldsymbol{Z}_1|\boldsymbol{Y}, \boldsymbol{\theta}_1]\,[\boldsymbol{Z}_2|\boldsymbol{Y}, \boldsymbol{\theta}_2]\ldots[\boldsymbol{Z}_k|\boldsymbol{Y}, \boldsymbol{\theta}_k], \tag{6.21}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)^T$.

Evolution of Processes in Space and Time

It is quite common that the process of interest $\boldsymbol{Y}$ involves a temporal and/or spatial component. If the process is one that evolves through time, such as meteorological quantities (e.g., temperature and humidity) at a given point in space, this can be represented as $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T)$. Often, a model will involve only one type of random variable (e.g., temperature **or** humidity) so that $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$. In these situations, modeling of the entire process is often accomplished at time T as

$$[\boldsymbol{Y}|\boldsymbol{\lambda}] = [Y_T|Y_{T-1}, \ldots, Y_1, \boldsymbol{\lambda}]\,[Y_{T-1}|Y_{T-2}, \ldots, Y_1, \boldsymbol{\lambda}]\ldots, [Y_1|\boldsymbol{\lambda}]. \tag{6.22}$$

A classic example of this structure occurs when we also make the traditional Markov assumption in time, namely that, for $t \in \{2, 3, \ldots\}$

$$[Y_t|Y_{t-1}, \ldots, Y_1] = [Y_t|Y_{t-1}],$$

in which case (6.22) at time T becomes

$$[\boldsymbol{Y}|\boldsymbol{\lambda}] = [Y_1|\boldsymbol{\lambda}] \prod_{t=2}^{T} [Y_t|Y_{t-1}, \boldsymbol{\lambda}]. \tag{6.23}$$

A common structure for such a process is that of an autoregressive time series, applied to either the process variables $Y_1, \ldots, Y_T$, or additive error terms. These two possibilities give, for $t = 2, \ldots,$

$$Y_t = \alpha Y_{t-1} + \epsilon_t, \tag{6.24}$$

$$Y_t = \mu_t + w_t; \quad w_t = \gamma w_{t-1} + \epsilon_t, \tag{6.25}$$

where $\epsilon_t \sim iidN(0, \sigma^2)$ and suitable restrictions are placed on α or γ to ensure that the processes are second order stationary. The structure of (6.25) is useful in situations for which the process $\boldsymbol{Y}$ is believed to depend on external covariates, which are then incorporated through additional modeling of the μ_t, while the structure of (6.24) is commonly used if the process $\boldsymbol{Y}$ represents the expected values of the observable random variables $\boldsymbol{Z}$.

If the process $\boldsymbol{Y}$ involves a spatial component we may make a Markov assumption in terms of defined neighborhoods. Again, it is frequently the case that we restrict attention to random variables connected with a single type of quantity, and we may represent these random variables at spatial locations $\boldsymbol{s}_i$ as $Y(\boldsymbol{s}_i)$; $i = 1, \ldots, n$. If neighborhoods of these spatial locations are defined as N_i; $i = 1, \ldots, n$, a Markov assumption may have the form

$$[Y(\boldsymbol{s}_i)|\boldsymbol{\lambda}, \{Y(\boldsymbol{s}_j) : j \neq i\}] = [Y(\boldsymbol{s}_i)|\boldsymbol{\lambda}, \{Y(\boldsymbol{s}_j) : j \in N_i\}]. \tag{6.26}$$

Formulation of a model for the process in terms of the conditional distributions (6.26) would then need to proceed in a manner that ensures the existence of a compatible joint distribution for $\boldsymbol{Y}$. A frequently encountered technique is to model a spatial process as a conditional autoregressive structure with full conditional distributions,

$$f(y(\boldsymbol{s}_i)|\boldsymbol{y}(N_i), \mu(\boldsymbol{s}_i), \tau_i) =$$
$$\frac{1}{[2\pi\tau_i^2]^{1/2}} \exp\left[-\frac{1}{2\tau_i^2}\{y(\boldsymbol{s}_i) - \mu(\boldsymbol{s}_i)\}^2\right]; \quad -\infty < y(\boldsymbol{s}_i) < \infty, \tag{6.27}$$

where $\boldsymbol{y}(N_i) = \{y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in N_i\}$, and

$$\mu(\boldsymbol{s}_i) = \alpha_i + \sum_{j:\boldsymbol{s}_j \in N_i} \eta_{i,j}\{y(\boldsymbol{s}_j) - \alpha_j\}. \tag{6.28}$$

This CAR model requires suitable restrictions on the parameters to ensure that a compatible joint exists, in which case it will be a joint Gaussian

distribution as discussed in Chapter 4. Under conditions for existence of a joint, the marginal expected values are,

$$E(\boldsymbol{Y}) = \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T.$$

The components of $\boldsymbol{\alpha}$ may be modeled as functions of spatial covariates, including trends (e.g., over latitudes). These marginal expectations describe what is called the large-scale model structure. Conditional covariances are determined by the parameters $\eta_{i,j}$ and τ_i^2; $i, j = 1, \ldots, n$, and these describe what is called the small-scale model structure.

Processes that include both spatial and temporal structures are challenging to model, and a variety of particular models have been proposed for this purpose. Two quite general approaches that are often adequate are to view the overall process as (1) a set of temporal processes that vary over space, or (2) a set of spatial processes that evolve over time. The choice is sometimes governed by the amounts of data available in the dimensions of space and time. For example, Haslett and Raftery (1998) modeled wind speeds in Ireland using largely time series structures in their models. The data available to these authors consisted of $6,226$ daily readings at 12 spatial locations. In contrast, Kaiser, Daniels, Furakawa and Dixon (2002) modeled particulate matter air pollution as a conditionally specified Gaussian spatial process in which the dependence parameters were allowed to vary over time.

6.3 Approaches to Analysis

In this section we provide an overview of ways that estimation and inference may be approached for hierarchical models.

6.3.1 Non-Bayesian Analysis

There is nothing inherently Bayesian about hierarchical models. Cressie and Wikle (2011, p. 23) term the basic structure of $[\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}]$ and $[\boldsymbol{Y}|\boldsymbol{\lambda}]$ *empirical hierarchical* models. In the general notation developed in this chapter, non-Bayesian estimation would proceed by deriving an objective function based on the marginal likelihood,

$$[\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{\lambda}] = \int [\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}]\,[\boldsymbol{Y}|\boldsymbol{\lambda}]\,d\boldsymbol{Y}. \tag{6.29}$$

For full maximum likelihood estimation, $[\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{\lambda}]$ would be used directly. It may occur, however, that the integral in (6.29) is intractable or the result is computationally difficult to deal with, often due to dimension of integral. There are a number of options that may then be considered.

EM Algorithm

The EM Algorithm provides a way to locate marginal maximum likelihood estimates in the face of missing information or incomplete data. In (6.29) we might consider the process $\boldsymbol{Y}$ to be missing information. We would then need to work with the complete data likelihood,

$$[\boldsymbol{Z},\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\lambda}] = [\boldsymbol{Z}|\boldsymbol{\theta}]\,[\boldsymbol{Y}|\boldsymbol{\lambda}]$$

and, in particular, the expected value of the logarithm of this quantity taken with respect to the distribution $[\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\lambda}]$ with fixed parameters $\boldsymbol{\theta}^{(p)}$ and $\boldsymbol{\lambda}^{(p)}$,

$$Q = \int \log\left([\boldsymbol{Z},\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{\lambda}]\right)\,[\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta}^{(p)},\boldsymbol{\lambda}^{(p)}]\,d\boldsymbol{Y}. \tag{6.30}$$

In the EM Algorithm, a sequence of Q functions become the objective functions for a sequence of maximizations. Although the integral in (6.30) hardly

looks simpler than that in (6.29), it turns out that it sometimes is simpler, or the function Q is more easily maximized than is the logarithm of $[\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{\lambda}]$ in (6.29). With either use of (6.29) for direct maximization, or the use of (6.30) for use in the EM Algorithm, we can often justify interchanging differentiation and integration, and derivatives may be evaluated through numerical integration.

Monte Carlo Maximum Likelihood

In Monte Carlo Maximum Likelihood we attempt to evaluate the integral (6.29), but do so numerically using importance sampling. Consider a situation in which both $[\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}]$ and $[\boldsymbol{Y}|\boldsymbol{\lambda}]$ have density or mass functions that can be written in Gibbsian form, which means that for some functions Q_z and Q_y,

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}) &= \frac{\exp\{Q_z(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})\}}{k_z(\boldsymbol{y},\boldsymbol{\theta})}, \\ p(\boldsymbol{y}|\boldsymbol{\lambda}) &= \frac{\exp\{Q_y(\boldsymbol{y}|\boldsymbol{\lambda})\}}{k_y(\boldsymbol{\lambda})}, \end{aligned} \tag{6.31}$$

where $k_z(\boldsymbol{y},\boldsymbol{\theta}) = \int exp\{Q_z(\boldsymbol{t}|\boldsymbol{y},\boldsymbol{\theta})\}\, d\boldsymbol{t}$ and $k_y(\boldsymbol{\lambda}) = \int \exp\{Q_y(\boldsymbol{t}|\boldsymbol{\lambda})\}\, d\boldsymbol{t}$. The marginal log likelihood may then be written as

$$\ell(\boldsymbol{\theta},\boldsymbol{\lambda}) = \log\left\{\int \exp\left[Q_z(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}) + Q_y(\boldsymbol{y}|\boldsymbol{\lambda}) - k_z(\boldsymbol{y},\boldsymbol{\theta})\right]\, d\boldsymbol{y}\right\} - \log\{k_y(\boldsymbol{\lambda})\}. \tag{6.32}$$

Now, the distribution of $\boldsymbol{Y}$ may involve complex dependencies, and we may not even be able to determine $k_y(\boldsymbol{\lambda})$ in closed form; we will assume here that we can determine $k_z(\boldsymbol{y},\boldsymbol{\theta})$ in closed form. But suppose that there are two other distributions with supports that dominate that of $p(\boldsymbol{y}|\boldsymbol{\lambda})$, $m_1(\boldsymbol{y}|\psi_1)$ and $m_2(\boldsymbol{y}|\psi_2)$, say, and that we produce samples $\boldsymbol{y}_r^{(1)}$ and $\boldsymbol{y}_r^{(2)}$; $\quad r = 1,\ldots,M$ from these two distributions. Then we can form a Monte Carlo approximation

to (6.32) as

$$\begin{aligned}\ell_M(\boldsymbol{\theta},\boldsymbol{\lambda}) &= \log\left\{\frac{1}{M}\sum_{r=1}^{M}\frac{1}{m_1(\boldsymbol{y}_r^{(1)}|\psi_1)}\exp\left[Q_z(\boldsymbol{z}|\boldsymbol{y}_r^{(1)},\boldsymbol{\theta})+Q_y(\boldsymbol{y}_r^{(1)}|\boldsymbol{\lambda})-\log\{k_z(\boldsymbol{y}_r^{(1)},\boldsymbol{\theta})\}\right]\right\} \\ &\quad -\log\left\{\frac{1}{M}\sum_{r=1}^{M}\frac{1}{m(\boldsymbol{y}_r^{(2)}|\psi_2)}\exp\left[Q_y(\boldsymbol{y}_r^{(2)}|\boldsymbol{\lambda})\right]\right\} \qquad (6.33)\end{aligned}$$

The Monte Carlo log likelihood function (6.33) may then be maximized in $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ using appropriate methods, which may involve Newton-type algorithms, direct search, and/or profiling. Importance sampling plays two roles in this procedure. First, we would like for the sampling distributions m_1 and m_2 to be easier to sample from than the distribution $p(\boldsymbol{y}|\boldsymbol{\lambda})$ which is what we would otherwise sample from for a Monte Carlo approximation to the log likelihood. Secondly, and perhaps more importantly, we want to use one sample from these distributions to evaluate (approximate) the log likelihood repeatedly, over iterations of a maximization procedure. An example of the use of Monte Carlo Maximum Likelihood with a spatial hierarchical model may be found in Kaiser, Cressie and Lee (2002). There, the data model $[\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}]$ consisted of conditionally independent binomial distributions with parameters (in our current notation) $\boldsymbol{Y}$, and there was no $\boldsymbol{\theta}$. The process model $[\boldsymbol{Y}|\boldsymbol{\lambda}]$ was formulated as a Markov random field model having conditional beta distributions and neighborhoods defined by distance between locations. Asymptotic results were used to compute Wald-like theory intervals and confidence regions for the parameters.

Composite Likelihood

Even if the full integral in (6.29) can be evaluated, the dimension of the resulting marginal distribution of $\boldsymbol{Z}$ may pose computational difficulties. For example, in some spatial problems, the dimension of $[\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{\lambda}]$ may be in the

tens or even hundreds of thousands. In these cases, it may be possible to use lower dimensional distributions to construct likelihood "pieces", such as for pairs of components of $\boldsymbol{Z}$, or simple differences $Z_i - Z_j$. These pieces may then be multiplied together to form what will be called a *composite likelihood*. Maximization of composite likelihoods yields, under various conditions that are often satisfied, consistent estimators that have identifiable asymptotic behavior. Composite likelihood is another topic that will be covered in greater depth later in the course. The use composite likelihood if the integral in (6.29) cannot be evaluated presents greater challenges, and additional work remains to be conducted relative to this possibility before it can be recommended as a solid approach to estimation and inference.

6.3.2 Bayesian Analysis

As indicated, there is nothing about hierarchical models that requires a Bayesian analysis. Nevertheless, such models are usually complex enough in structure that use of a prior $[\boldsymbol{\theta}, \boldsymbol{\lambda}]$ and subsequent simulation from the posterior distribution via the methods of Markov Chain Monte Carlo (MCMC) has become the norm for analysis of hierarchical models.

Simulation from Full Posterior

The usual approach for Bayesian analysis of complex hierarchical models is to simulate values from the posterior distribution (6.20) using the methods of Markov Chain Monte Carlo (MCMC). If the full likelihood is available in closed form, there is really nothing new at this point. The primary challenges faced are often determining good jump proposals for algorithms that rely totally or in part on Metropolis-Hastings, and determining efficient techniques

to sample from other unnormalized conditional posterior distributions in algorithms that rely totally or in part on Gibbs Sampling. It is not uncommon that overall MCMC algorithms involve some Metropolis within Gibbs, some block Gibbs, and some form of rejection sampling. Exactly what type of algorithm is likely to be effective depends on many particulars of the model structure. It is often the case that the process $\boldsymbol{Y}$ corresponds to expected values of the observable data. The data model parameter $\boldsymbol{\theta}$ may then be other data model parameters, such as variances, or it may be absent. The process parameters $\boldsymbol{\lambda}$ are nearly always present, and often form the central concern for inference. If there are blocks of dependent quantities these may be most effectively handled with a random walk Metropolis that contains correlated components.

For some models the complete likelihood is not available in close form, such as if the process model $[\boldsymbol{Y}|\boldsymbol{\lambda}]$ constitutes a conditionally specified Markov random field model (other than a Gaussian conditionals or CAR model). In these cases the density corresponding to the process model is known only up to a normalizing constant,

$$f(\boldsymbol{y}|\boldsymbol{\lambda}) = \frac{1}{k(\boldsymbol{\lambda})} \exp[Q(\boldsymbol{y}|\boldsymbol{\lambda})], \tag{6.34}$$

where $k(\boldsymbol{\lambda}) = \int \exp[Q(\boldsymbol{y}|\boldsymbol{\lambda})]\, d\boldsymbol{y}$ and Q is a known function. The fundamental difficulty is that the integral defining $k(\boldsymbol{\lambda})$ is not tractable, either due to form or due to dimension. This is similar to the situation discussed in the previous subsection under the heading of Markov Chain Maximum Likelihood, and causes problems even if the other portions of the model are all known. Note that what is called a normalizing constant for $f(\boldsymbol{y}|\boldsymbol{\lambda})$ is a function of the parameters and is thus not a constant for the likelihood function. A number of methods for dealing with unnormalized likelihoods in Bayesian analysis

have been proposed, notably by Moller *et al.* (2006) and Liang (2010). These generally involve clever manipulations that result in the cancellation of the normalizing constant in the acceptance probability of a Metropolis-Hastings algorithm. The approach of Moller *et al.* requires what is called perfect sampling (as opposed to approximate sampling which is what nearly all of MCMC is about), while the algorithm of Liang does not require perfect sampling. See Hughes, Haran and Caragea (2011) for a discussion of this in the context of a purely spatial model for binary random variables. It should be recognized that while these techniques have been successfully applied to a number of particular models, they are far from being generally applicable. So, it would be a misconception to form the opinion that the methods of Moller*et al.* and/or Liang have completely solved the problem of dealing with unnormalized likelihoods in the analysis of hierarchical models. Dealing with unnormalized likelihoods in Bayesian analysis remains a challenge, for both non-Gaussian Markov random field models on their own, and especially when used as process models in a hierarchical structure.

Some Parameters Considered Fixed

It is sometimes the case that one or more parameters in the model are considered fixed, or estimated outside the analysis of the hierarchical model itself. These parameters are nearly always elements of $\boldsymbol{\theta}$, such as data model variances. Treating some data model parameters as fixed can often greatly simplify the search for efficient ways to sample from the conditional posterior distributions of other model quantities. Sometimes, this simplification in the model can be well justified, such as when the components of $\boldsymbol{Z}$ are averages of multiple measurements. Other times, it is made with little or no

justification.

Holding certain parameters fixed is also an effective approach to debugging an MCMC algorithm that appears to be exhibiting undesired behavior. Since all of the components of an MCMC algorithm are interacting pieces in determination of the joint posterior, isolating those components that are causing problems is a necessary step in improving algorithm performance. This can be true regardless of whether or not the final estimation is conducted with one or more parameters held fixed.

Sequential Monte Carlo

In any number of problems, both the observation and process models evolve over time. If the entire time frame of interest has been observed, one may choose to examine the full posterior, or the full posterior with some parameters held fixed. On the other hand, there may be interest in using the model to forecast future values, at least over short periods of time. In this case we would like an analysis that is sequential in nature. Similarly, there may be interest in viewing the problem sequentially for the purpose of detecting change points in the structure of the process model. We have already seen the idea of sequential Bayesian analysis in situations with conjugate prior and data model pairs. There, the posterior at one point in time serves as the prior for the next point in time, which is greatly facilitated mathematically by conjugacy. We would like to determine how to mimic this progression for the non-conjugate situations that arise in dynamic hierarchical models. Consider a sequence of observations of one type, $\{Z_t : t = 0, 1, \ldots\}$ and the corresponding process $\{Y_t : t = 1, \ldots\}$. We assume that primary interest lies in analyzing and forecasting the process, and in denoting distributions

with square brackets we will suppress explicit dependence on parameters, so that here we will write $[Z_t|Y_t]$ rather than $[Z_t|Y_t, \boldsymbol{\theta}]$. Our goal is to determine distributions, for $t = 2, 3, \ldots,$

$[Y_t|Z_1, \ldots, Z_{t-1}]$ which we will call the forecast distribution at time t,

$[Y_t||Z_1, \ldots, Z_t]$ which we will call the posterior distribution at time t.

Note that the forecast distribution is essentially a prior for the process at time t. The relations that allow us to upadte the forecast and posterior distributions in a sequential manner are

$$[Y_t|Z_1, \ldots, Z_{t-1}] = \int [Y_t|Y_{t-1}]\,[Y_{t-1}|Z_1, \ldots, Z_{t-1}]\,dY_{t-1}, \tag{6.35}$$

$$[Y_t|Z_1, \ldots, Z_t] \propto [Z_t|Y_t]\,[Y_t|Z_1, \ldots, Z_{t-1}]. \tag{6.36}$$

Notice that, given the left hand side of (6.35) and the data model $[Z_t|Y_t]$ we can obtain the posterior through the use of (6.36). Given this posterior and a distribution that describes the dynamic updating of the process, $[Y_t|Y_{t-1}]$ we may then obtain the next forecast distribution from (6.35).

The relations (6.35) and (6.36) underlie the methods of Sequential Monte Carlo, which also involves repeated use of the idea of importance sampling as providing weights for sampled values in a Monte Carlo approximation. Many additional details of sequential Monte Carlo are needed to put this body of methods into practice. Those details are covered in Stat 515, Advanced Bayesian Methods.

6.4 Case Study: Bias Adjusted Wind Speed Forecasts

In Iowa, as well as several other states, much attention has been focused on the production of electricity based on wind energy as a source. In 2022, wind power accounted for 62% of the generated electricity in the state. Iowa has over $6,000$ wind turbines. One energy company, MidAmerican Energy, has announced a goal of eventually being able to provide 100% of the electricity demands of its customers using renewable energy sources, primarily wind. The production of electricity from wind energy is complex and involves many topics, including several types of engineering, atmospheric science, mathematics, and statistics. Since the wind energy available depends on the weather, forecasts of weather have also become important in the overall picture of wind energy.

6.4.1 Background Information

The way that electricity is transmitted across the United States is an exceedingly complex topic, but basically an energy company must have enough electricity available to meet the needs of its customers, either from their own production, or as purchased from what is called the *grid*. At the same time, if an energy company produces more electricity than is needed by its direct customers, it may sell the excess to the grid. Wind farm operators must bid on energy to be sold and bought from the grid, and they do so for 48 hour periods of time. They can then adjust those bids for shorter time windows (e.g., 24 or 12 hours into the future), but the shorter the lead time the more it costs to purchase energy from the grid and the less is made off of selling

excess energy to the grid. Thus, having accurate forecasts of wind speed for periods of up to 48 hours is important for wind farms.

Weather models produce forecasts of wind speed along with forecasts of other meteorological variables (temperature, precipitation, etc.). These models are deterministic, but highly complex constructions of the way that energy (in the atmosphere, not electricity here) moves through pieces and components of the atmosphere. Generally, weather models are constructed on regular three dimensional grids consisting of large spatial areas (e.g., continents or hemispheres), and altitudes. These models depend on the particular sets of differential equations used to model the movement of energy and its products over space, time, and among components of the atmosphere, the particular "boundary conditions" used to constrain the addition and losses of energy (according to first principles of physics), and the initial conditions (or starting values) used to start the model. These latter are usually in terms of the variables produced by the models, such as temperature and water vapor content. The temporal frequency at which numerical weather models are capable of producing output are in terms of seconds or minutes. These temporal frequencies are called time steps in updating the sets of differential equations that govern the movement of energy and modeled variables across space. Typically, however, these high temporal frequencies are averaged over units of time, to produce values in terms of hours or days. Weather models are generally thought to give reasonable projections of what will happen over periods of four to maybe five or six days. Then they "escape initial conditions" and, although they are believed to accurately reflect the dynamics of atmospheric processes beyond that point, their output can no longer be indexed to particular points in time, and they are then forecasting climate rather than weather.

One can, of course, also use purely stochastic models to forecast weather, such as time series (ARMA and ARIMA) models. Such statistical models that contain no atmospheric science are often thought to be better than numerical weather models at forecasting current conditions into extremely short forecast windows, such as a few hours or maybe up to about a half day. Numerical weather models have been shown to be superior for forecasting time windows of more than a day. Some attention has been given to combining deterministic meteorological models and stochastic models for forecasting values into time windows of one to two days. The model of this section is an example of such a model. In these modeling exercises, because numerical weather (and climate) models contain such complicated dynamics, their output has commonly been incorporated as realizations of stochastic processes, even though they are, in fact, deterministic.

The objective of the modeling exercise discussed in this section was to take wind speed forecasts from an ensemble of weather models, incorporate those forecasts into a stochastic structure that also included observed (directly measured) wind speeds, and produce improved wind speed forecasts at a given wind farm over a period of twenty four hours. This was a preliminary exercise that was not intended to be a final product, but that could provide a framework for extension to forty eight hour periods such as those of interest to wind farm operators. A few comments are provided relative to extension of the basic structure of the model described here at the end of the section.

Data were obtained for a wind farm in Iowa, consisting of a set of cases, which is a record of 54 hours of observed (i.e., directly measured) wind speeds and the output from 8 numerical weather models with hourly values. The use of a set of weather models is called using an *ensemble* of models, and it is generally believed that the use of ensembles can provide some protection

against the bias of any particular model. The objective of analysis was to take the first 30 hours of values of both observed and forecast wind speeds to fit a model for estimating the bias of the weather model forecasts, and then to apply a correction to the forecast values for the remaining 24 hours.

6.4.2 Model Formulation

Define the following:

- Let $Y_{0,t}$ be a random variable connected with the observed wind speed at time $t = 1, \ldots, T$.
- Let $Y_{j,t}$ be a random variable connected with the forecast wind speed at time $t = 1, \ldots, T$ from weather model $j = 1, \ldots, M$.
- Denote the true wind speed as μ_t at time $t = 1, \ldots, T$.

To develop the model, suppose that observations are unbiased for true wind speed, while each weather model has its own sequence of biases over time. The observation model (what corresponds to $[\boldsymbol{Z}]$ in our general presentation) consists of two parts,

- $Y_{0,t} \sim N(\mu_t, \lambda_0^{-1})$
- $(Y_{j,t} - \mu_t) \sim N(b_{j,t}, \lambda_j^{-1})$

Note that each numerical weather model has its own precision parameter λ_j. These parameters could be taken as part of the data model parameter $\boldsymbol{\theta}$ in the general presentation, but we will instead include these values as a portion of the process model. The process $\boldsymbol{Y}$ in our general presentation thus consists of the λ_j and the $b_{j,t}$ for $j = 1, \ldots, M$ and $t = 1, \ldots, T$. We will then specify the process model to be,

- $\lambda_j \sim$ iid Gamma(α, β)
- $b_{j,t} = \gamma b_{j,t-1} + w_{j,t}; \quad w_{j,t} \sim N(0, \tau^{-1})$

and conditional independence is assumed throughout.

Some aspects of this model that are worthy of note include the following.

1. All distributions are normals, which should perhaps be modified for at least the $Y_{0,t}$.
2. $E\{Y_{j,t}\} = \mu_t + b_{j,t}$, so each forecast model is allowed its own bias at each point in time.
3. Each forecast model has its own precision λ_j, although across models these precisions are assumed to have arisen from a common mixing distribution.
4. γ is an autoregressive parameter for the bias processes $\{b_{j,t} : t = 1, \ldots, T;\ j = 1, \ldots, M\}$. If γ differs from 0 the bias follows a dynamic process such that γ is common to all forecast models.
5. In the model we will consider λ_0, and τ to be given (known). In practice, λ_0 should be able to be estimated from external information. We may consider τ a tuning parameter or eventually try to extend the model so it is estimated.

Components of the model that are of particular interest are the biases $\{b_{j,t} : t = 1, \ldots, T;\ j = 1, \ldots, M\}$, the controlling parameters of the distribution of forecast precisions α and β, and the autoregressive parameter γ. The biases are a portion of the process $\boldsymbol{Y}$ from our general presentation, while α, β, and γ are components of $\boldsymbol{\lambda}$ from the general presentation.

6.4.3 Distributions Involved in the Analysis

Prior Distributions

Prior distributions are needed for α, β, γ, $\{\mu_t : t = 1, \ldots, T\}$ and $\{b_{j,0} : j = 1, \ldots, M\}$. These distributions were specified as follows:

- $\alpha \sim \text{Uniform}(0, A)$
- $\beta \sim \text{Gamma}(a, b)$
- $\gamma \sim N(0, \lambda_g^{-1})$
- For $t = 1, \ldots, T$, $\mu_t \sim N(M_\mu, \lambda_\mu^{-1})$.
- For $j = 1, \ldots, M$, $b_{j,0} \sim N(0, \tau^{-1})$.

Note that the precision of the $b_{j,0}$ is taken as the same value as the precision of the random shock w_t in the process model of the $b_{j,t}$.

Distributions Implied by the Model

The following distributions follow from the model and will be needed in deriving full conditional posteriors:

1. For $t = 1, \ldots, T$, $Y_{0,t}$ has density

$$f_0(y_{0,t}|\mu_t, \lambda_0) = \frac{\lambda_0^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda_0}{2}(y_{0,t} - \mu_t)^2\right\}.$$

2. For $t = 1, \ldots, T$ and $j = 1, \ldots, M$, $Y_{j,t}$ has density

$$f_{j,t}(y_{j,t}|\mu_t, b_{j,t}, \lambda_j) = \frac{\lambda_j^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\lambda_j}{2}(y_{j,t} - \mu_t - b_{j,t})^2\right\}.$$

3. For $j = 1, \ldots, M$, λ_j has density

$$g_j(\lambda_j|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\lambda_j)^{\alpha-1} \exp(-\beta\lambda_j)$$

4. For $t = 1, \ldots, T$ and $j = 1, \ldots, M$, the $b_{j,t}$ have conditional densities (given $b_{j,t-1}$),

$$h_{j,t}(b_{j,t}|b_{j,t-1}, \gamma, \tau) = \frac{\tau^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\tau}{2}(b_{j,t} - \gamma b_{j,t-1})^2\right\}$$

5. The joint density of $Y_{j,1}, \ldots, Y_{j,T}$ is, for $j = 1, \ldots, M$,

$f_j(y_{j,1}, \ldots, y_{j,T}|\mu_{j,1}, \ldots, \mu_{j,T}, b_{j,1}, \ldots, b_{j,T}, \lambda_j) =$

$$\frac{\lambda_j^{T/2}}{(2\pi)^{T/2}} \exp\left\{-\frac{\lambda_j}{2}\sum_{t=1}^{T}(y_{j,t} - \mu_t - b_{j,t})^2\right\}$$

6. The joint density of $Y_{1,t}, \ldots, Y_{M,t}$ is, for $t = 1, \ldots, T$,

$f_t(y_{1,t}, \ldots, y_{M,t}|\mu_t, b_{1,t}, \ldots, b_{M,t}, \lambda_1, \ldots, \lambda_M) =$

$$\frac{\left(\prod_{j=1}^{M}\lambda_j^{1/2}\right)}{(2\pi)^{M/2}} \exp\left\{\sum_{j=1}^{M} -\frac{\lambda_j}{2}(y_{j,t} - \mu_t - b_{j,t})^2\right\}$$

7. The joint density of $\lambda_1, \ldots, \lambda_M$ is,

$$g(\lambda_1, \ldots, \lambda_M|\alpha, \beta) = \frac{\beta^{M\alpha}}{\{\Gamma(\alpha)\}^M}\left(\prod_{j=1}^{M}\lambda_j\right)^{\alpha-1} \exp\left(-\beta\sum_{j=1}^{M}\lambda_j\right)$$

8. The joint density of $b_{j,1}, \ldots, b_{j,T}$ is given by the Markov property implied by the model for $b_{j,t} = \gamma b_{j,t-1} + w_{j,t}$ as

$$\begin{aligned} h_j(b_{j,1}, \ldots, b_{j,T}|\gamma, \tau) &= \prod_{t=1}^{T} h_{j,t}(b_{j,t}|b_{j,t-1}, \gamma, \tau) \\ &= \frac{\tau^{T/2}}{(2\pi)^{T/2}} \exp\left\{-\frac{\tau}{2}\sum_{t=1}^{T}(b_{j,t} - \gamma b_{j,t-1})^2\right\} \end{aligned}$$

9. The joint density of the entire set $\{b_{j,t} : t = 1, \ldots, T; j = 1, \ldots, M\}$ is, through assumed independence of the forecast models,
$h(\{b_{j,t} : t = 1, \ldots, T; j = 1, \ldots, M\}|\gamma, \tau) =$

$$\begin{aligned} & \prod_{j=1}^{M} h_j(b_{j,1}, \ldots, b_{j,T}|\gamma, \tau) \\ = & \frac{\tau^{MT/2}}{(2\pi)^{MT/2}} \exp\left\{-\frac{\tau}{2}\sum_{j=1}^{M}\sum_{t=1}^{T}(b_{j,t} - \gamma b_{j,t-1})^2\right\} \end{aligned}$$

Conditional Posteriors

For analysis through a Gibbs sampling algorithm we need the following full conditional posteriors:

1. The full conditional posterior of α is given as,

$$\begin{aligned} p(\alpha|\cdot) &\propto \pi(\alpha) g(\lambda_1, \ldots, \lambda_M|\alpha, \beta) \\ &\propto \left\{\frac{\beta^\alpha}{\Gamma(\alpha)}\right\}^M \left(\prod_{j=1}^{M}\lambda_j\right)^{\alpha-1} \end{aligned}$$

This will prove the most challenging of the conditional posteriors and will be sampled using an adaptive ratio of uniforms algorithm.

2. The full conditional posterior of β is given as,

$$\begin{aligned} p(\beta|\cdot) &\propto \pi(\beta) g(\lambda_1, \ldots, \lambda_M|\alpha, \beta) \\ &\propto \beta^{M\alpha+a-1} \exp\left\{-\left(b + \sum_{j=1}^{M}\lambda_j\right)\beta\right\}, \end{aligned}$$

which is a Gamma distribution with parameters

$$M\alpha + a \text{ and } b + \sum_{j=1}^{M}\lambda_j$$

3. The full conditional posterior of γ is given as,

$$\begin{aligned} p(\gamma|\cdot) &\propto \pi(\gamma)h(\{b_{j,t} : t = 1, \ldots, T;\, j = 1, \ldots, M\}|\gamma, \tau) \\ &\propto \exp\left\{-\frac{\gamma^2}{2} - \frac{\tau}{2}\sum_{j=1}^{M}\sum_{t=1}^{T}(b_{j,t} - \gamma b_{j,t-1})^2\right\} \end{aligned}$$

Upon completing the square this may be seen to be a normal distribution with precision parameter

$$\lambda_\gamma + \tau\sum_{j=1}^{M}\sum_{t=1}^{T} b_{j,t-1}^2$$

and mean

$$\frac{\tau\sum_{j=1}^{M}\sum_{t=1}^{T} b_{j,t}b_{j,t-1}}{\lambda_\gamma + \tau\sum_{j=1}^{M}\sum_{t=1}^{T} b_{j,t-1}^2}$$

4. For $j = 1, \ldots, M$, the full conditional posteriors of λ_j are given as,

$$\begin{aligned} p(\lambda_j|\cdot) &\propto g_j(\lambda_j|\alpha, \beta) f_j(y_{j,1}, \ldots, y_{j,T}|\mu_1, \ldots, \mu_T, b_{j,1}, \ldots, b_{j,T}, \lambda_j) \\ &\propto (\lambda_j)^{\alpha+T/2-1}\exp\left\{-\left(\beta + \frac{1}{2}\sum_{t=1}^{T}(y_{j,t} - b_{j,t} - \mu_t)^2\right)\lambda_j\right\}, \end{aligned}$$

which is a Gamma distribution with parameters

$$\alpha + (T/2) \text{ and } \beta + \frac{1}{2}\sum_{t=1}^{T}(y_{j,t} - b_{j,t} - \mu_t)^2$$

5. For $t = 1, \ldots, T$, the full conditional posteriors of μ_t are given as,

$$\begin{aligned} p(\mu_t|\cdot) &\propto \pi(\mu_t) f(y_{0,t}|\mu_t, \lambda_0) f_t(y_{1,t}, \ldots, y_{M,t}|\mu_t, b_{1,t}, \ldots, b_{M,t}, \lambda_1, \ldots, \lambda_M) \\ &\propto \frac{\lambda_\mu^{1/2}}{(2\pi)^{1/2}}\exp\left\{-\frac{\lambda_\mu}{2}(\mu_t - M_\mu)^2\right\}\frac{\lambda_0^{1/2}}{(2\pi)^{1/2}}\exp\left\{-\frac{\lambda_0}{2}(y_{0,t} - \mu_t)^2\right\} \times \\ &\quad \prod_{j=1}^{M}\left(\frac{\lambda_j^{1/2}}{(2\pi)^{1/2}}\right)\exp\left\{\sum_{j=1}^{M} -\frac{\lambda_j}{2}(y_{j,t} - b_{j,t} - \mu_t)^2\right\} \end{aligned}$$

Upon dropping constants and completing the square this may be seen to be a normal distribution with precision parameter

$$\lambda_\mu + \lambda_0 + \sum_{j=1}^{M} \lambda_j$$

and mean

$$\frac{\lambda_\mu M_\mu + \lambda_0 y_{0,t} + \sum_{j=1}^{M} \lambda_j (y_{j,t} - b_{j,t})}{\lambda_\mu + \lambda_0 + \sum_{j=1}^{M} \lambda_j}$$

6. For $j = 1, \ldots, M$ and $t = 1, \ldots, T$, the full conditional posteriors of $b_{j,t}$ are given as,

$$\begin{aligned} p(b_{j,t}|\cdot) &\propto h_{j,t}(b_{j,t}|b_{j,t-1}, \gamma, \tau) h_{j,t+1}(b_{j,t+1}|b_{j,t}, \gamma, \tau) f_{j,t}(y_{j,t}|\mu_t, b_{j,t}, \lambda_j) \\ &\propto \exp\left\{ -\frac{\tau}{2}(b_{j,t} - \gamma b_{j,t_1})^2 - \frac{\tau}{2}(b_{j,t+1} - \gamma b_{j,t})^2 - \frac{\lambda_j}{2}(y_{j,t} - \mu_t - b_{j,t})^2 \right\} \end{aligned}$$

Upon completing the square this may be seen to be a normal distribution with precision parameter

$$\tau(1 + \gamma^2) + \lambda_j$$

and mean

$$\frac{\tau\gamma(b_{j,t-1} + b_{j,t+1}) + \lambda_j(y_{j,t} - \mu_t)}{\tau(1 + \gamma^2) + \lambda_j}$$

7. For $j = 1, \ldots, M$ the full conditional posteriors of $b_{j,0}$ are given as,

$$\begin{aligned} p(b_{j,0}|\cdot) &\propto \pi(b_{j,0}) h_{j,1}(b_{j,1}|b_{j,o}, \gamma, \tau) \\ &\propto \exp\left\{ -\frac{\tau}{2} b_{j,0}^2 - \frac{\tau}{2}(b_{j,1} - \gamma b_{j,0})^2 \right\} \end{aligned}$$

Upon completing the square this may be seen to be a normal distribution with precision parameter

$$\tau(\gamma^2 + 1)$$

and mean

$$\frac{\gamma b_{j,1}}{\gamma^2 + 1}$$

6.4.4 Results for a Case

The posterior distribution sampled from is

$$p(\alpha, \beta, \gamma, \{\lambda_j\}, \{\mu_t\}, \{b_{j,t}\}\{b_{j,0}\}, j = 1, \ldots, M; t = 1, \ldots, T | \boldsymbol{y}_0, \{\boldsymbol{y}_j : j = 1, \ldots, M\}) \tag{6.37}$$

The results presented in this section were produced by a relatively short MCMC run of length 500, of which the first 100 were discarded as burn-in. This is because the entire exercise with this model was developmental in nature – we are attempting to determine useful structures to use in this problem, not provide a final solution for one particular set of data. Considerations of efficiency of the overall Gibbs Algorithm used in this example, such as mixing rates, correlations among components of the posterior, and sensitivity to fixed parameters in the data model (λ_0 and τ) are not discussed.

Model Fitting Period

For the model fitting period (first 30 hours of data), our best estimate of the true wind speed is the process that consists of the posterior expectations of the $\{\mu_t : t = 1, \ldots, T\}$. These are our fitted values from the model. Without incorporation of observed data and modeling of biases, our best estimate of the true wind speed might be the average of the 8 numerical weather model forecasts, so these two sequences of values can be compared to both each other and the observed wind speeds. If our bias model does not substantially out-perform the meteorological model average over this period, we have certainly failed to meet our objective.

Figure 6.1 presents values for one example case. The points and solid black line in this figure are the observed wind speed values. The green line is the average of the 8 numerical weather models. The dark blue line extending over hour 0 to hour 30 is the posterior expected values of the μ_t. This figure also presents some values for the forecast period of hour 31 to hour 50, and this will be discussed in the sequel. Figures 6.2 through 6.4 present posterior distributions for α, β, and γ.

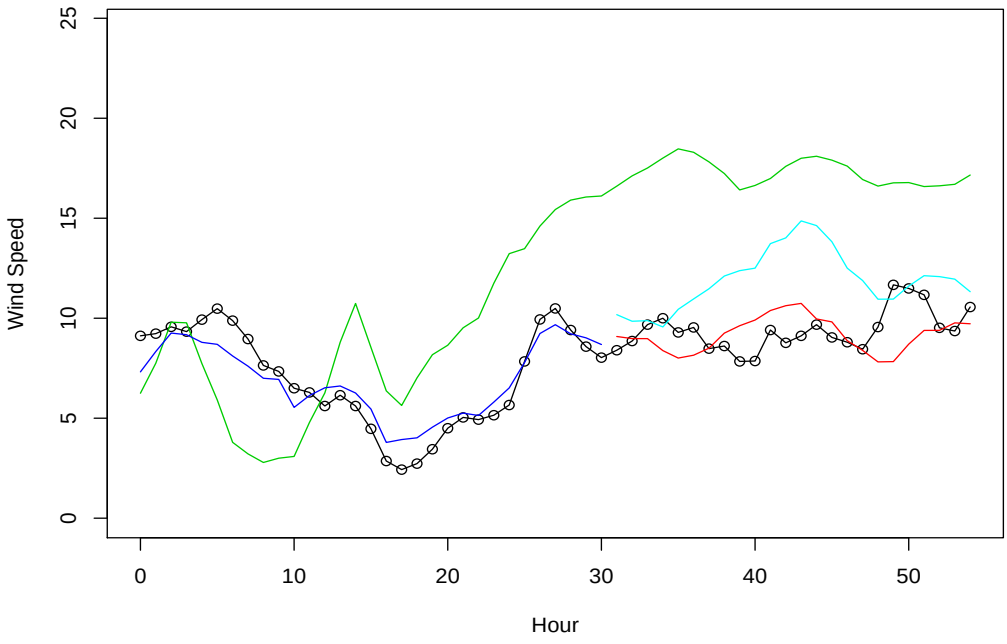

Figure 6.1: Fitted model and forecasts for one case. See text for complete explanation.

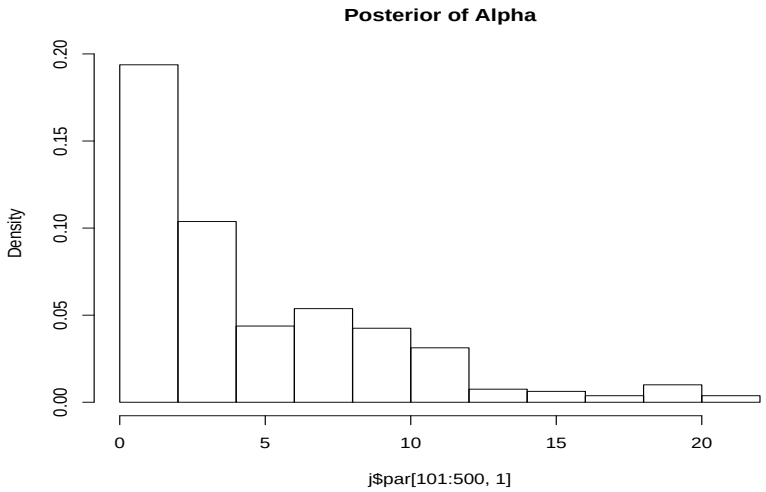

Figure 6.2: Approximate posterior of α from 400 simulated values after a burn-in of 100.

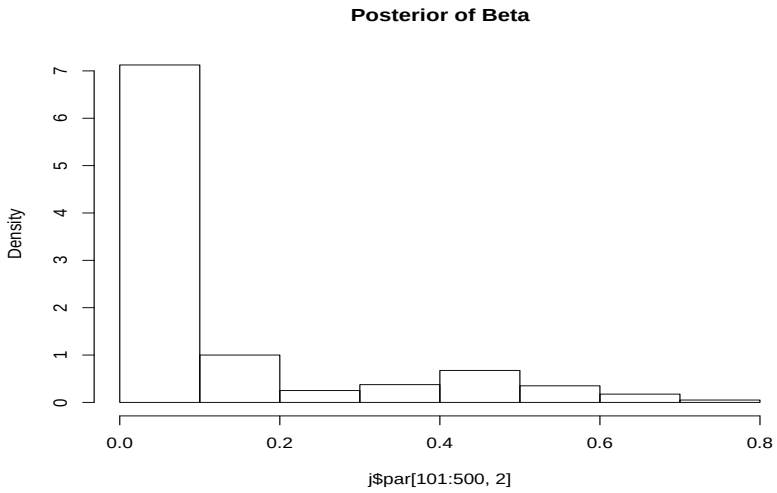

Figure 6.3: Approximate posterior of β from 400 simulated values after a burn-in of 100.

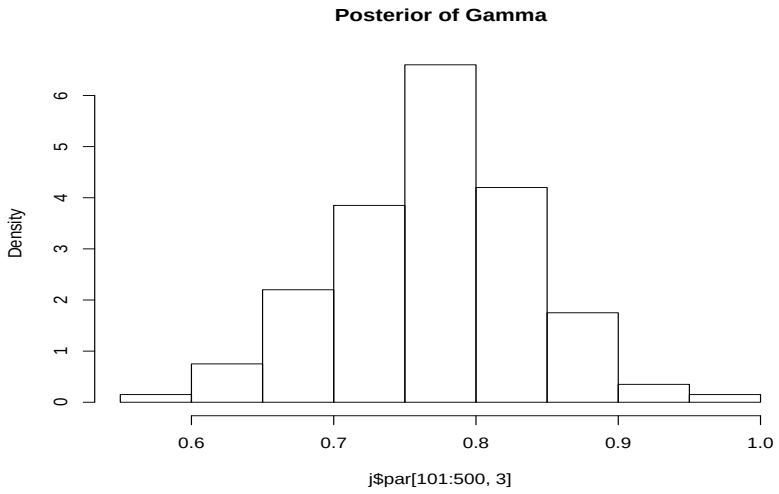

Figure 6.4: Approximate posterior of γ from 400 simulated values after a burn-in of 100.

The posterior distributions of both α and β are $J-$shaped, meaning that we place more belief on smaller values of these parameters than larger ones. The meaning of smaller here certainly depends on whether we are considering α or β, however. The entire distribution for β occurs at values of less than 1.0, with most of the probability (belief) occurring for values less than 0.2. With a value for α greater than 1 (which includes most of our belief about this quantity) this leads to distributions for the precision parameters of the numerical weather models (which is what α and β control) with high variability. The posterior distribution for γ, on the other hand, appears fairly symmetric and is centered at moderately high values $(0.75 - 0.80)$. Recall that this parameter controls autocorrelation in the numerical weather forecasts.

Forecast Period

For the forecast period (last 24 hours) we do not make use of observed wind speeds in the model, and our forecast comes from a combination of the 8 numerical weather models after bias adjustment, that is, some type of an average of bias-adjusted values from the 8 meteorological forecasts. Bias adjustment may undertaken in several different ways:

1. Use the posterior mean of the $b_{j,t}$ at $t = 30$. This is the simplest, and perhaps the crudest, adjustment we could make. Although the model has allowed the bias for each model to vary over time (according to an autoregressive process) the forecasts would presume that the bias at time hour 30 then remains fixed for the following 24 hours.

2. Simulate a bias process for each model by starting at the posterior means at $t = 30$, $\hat{b}_{j,30}$, and using the posterior expected value of γ, $\hat{\gamma}$ say, in

$$b_{j,t} = \hat{\gamma} b_{j,t-1} + w_{j,t}; \;\; t = 31, \ldots,$$

   where $w_{j,t} \sim iidN(0, \tau^2)$ and $b_{j,30} = \hat{b}_{j,30}$.

3. Simulate a bias process for each model as in 2, but instead of using $\hat{\gamma}$ (the posterior expectation of γ), use a value $\tilde{\gamma}$ simulated from the posterior distribution of γ. Note, however, that there would be only one such value applied to all models, since the model assumes this autoregressive parameter is common to all of the numerical weather models. One could extend this idea by also replacing $\hat{b}_{j,30}$ with values simulated from posterior distributions of the biases at time $t = 30$.

None of these possibilities is entirely pleasing, and how bias adjustments can be made during the forecast period can be identified as the primary concern

for further development for models of this type. For the forecasts in the example of Figure 6.1, the approach of item 2 in the immediately preceding list was used.

A single forecast was produced from the bias-adjusted weather model values by averaging in two different ways. One was to take a simple average of values from the eight models. The other was to take a weighted average with weights given by the expected values of posterior distributions for the weather model precision parameters $\lambda_j : \ j = 1, \ldots, M$. In Figure 6.1, the red line gives forecasts produced as a simple average and the aqua line gives forecasts produced as a weighted average. These lines extend over hours 31 to 54 in the figure. It appears that the simple average is somewhat better than the weighted average, at least for this one case. Importantly for this line of model development for forecasting wind speeds, both bias-adjusted forecasts are decidedly more accurate than the simple average of values from the numerical weather models.

6.4.5 Extending the Model

To improve bias adjustment in the forecast period, what is needed is some information that varies from hour to hour, is available at time $t = 30$ for the forecast period, and has some systematic relation with the manner in which model-specific bias changes over time for the set of numerical weather models. This information could then be incorporated into the model for time and model specific biases (the $b_{j,t}$) as a covariate. Suppose that such a variable is identified and denote it as $\boldsymbol{z}_{j,t}$ where $j = 1, \ldots, M$ and $t = 1, \ldots, T$ continue to index weather model and time, respectively. Then the model for

bias evolution could be changed to

$$b_{j,t} = h(\boldsymbol{z}_{j,t}),\, \boldsymbol{\beta}) + w_{j,t}, w_{j,t} = \gamma\, w_{j,t-1} + \epsilon_{j,t},$$

where $\epsilon_{j,t} \sim iidN(0, \tau^{-1})$ and $h(\cdot)$ is a known function. An extension of the model in this manner was examined in a PhD dissertation (Bramer, L.M., "Methods for modeling and forecasting wind characteristics, ISU PhD dissertation, 2013).

As a final note, it should be mentioned that modeling wind speeds and improving wind speed forecasts in particular is a challenging problem. The results presented here for the one case examined are encouraging. But while these may not have been the best results among the cases examined, they certainly were not the worst. A major difficulty for dealing with wind speeds is that there are many factors involved and, although there may be a few that are key or primary determinants, the effect of factors in determining wind speed depends on values of other factors – identification of stable relations among atmospheric variables is elusive. Thus, the search for generally applicable ways to improve wind speed forecasts continues, with the use of hierarchical models is a promising avenue for investigation.

Part III

Topics in Frequentist Analysis

Part IV

Topics in Frequentist Analysis

Chapter 7

A Primer on Asymptotic Normality

The way in which results involving asymptotic normality are presented in the literature can differ from reference to reference, and the most appropriate representation typically depends on the specific form of technical conditions under which those results are developed. In what follows we have two objectives, to present results in notation that can be related to notation used the references cited, and to make clear the general and common structure involved in results connected with asymptotic normality. Let $\boldsymbol{\theta}$ denote a true parameter value with q individual components, and let $\hat{\boldsymbol{\theta}}_n$ denote an estimator of $\boldsymbol{\theta}$ based on n random variables. Typically, $\hat{\boldsymbol{\theta}}_n$ will have been obtained by maximizing some objective function $\ell^{(n)}(\boldsymbol{\theta}|\boldsymbol{y})$. The objective function $\ell^{(n)}(\boldsymbol{\theta}|\boldsymbol{y})$ may be a log likelihood, a log quasi-likelihood, a log composite likelihood or some other objective function connected with an unbiased estimating function. These are topics that are covered in later chapters in these notes. For now, simply accept that there are some functions that we

can construct from observed data and unknown parameters that, when maximized in the parameter values, provide estimates that may have discernible properties. The most notable among such properties is asymptotic normality.

7.1 Asymptotic Context

The production of asymptotic results always involves what is known as the *asymptotic context* within which results are formulated. For models that involve independence and no grouping of variables this is not much of an issue. Sample size increases by obtaining more and more independent realizations of whatever model is under consideration. Many of the estimation methods for which we will be interested in asymptotic results, however, are motivated at least in part through the presence of additional structure in the collection of random variables involved in a model. For example, consider a longitudinal data structure in which we have m_i random variables for each of $i = 1, \ldots, n$ individuals. Here, the total number of random variables involved in the model can grow large by having $m_i \to \infty$ for some or all i, but with n fixed, by having $n \to \infty$ but with a fixed and common value $m_i = m$, or by having both $m_i = m \to \infty$ and $n \to \infty$ in some specified manner, such as with $n/m = \lambda < \infty$ for a fixed λ, or with $n/m \to \infty$ as well as each of m and n individually. Similarly, for a spatial problem defined on a finite index random field such as a lattice, there are several ways that overall sample size can grow without bound. One way that sample size can grow large is if a lattice defined by a fixed spacing and configuration of locations simply expands without bound. This is often called the *expanding lattice* asymptotic context. Another way for sample size to increase is if the number of independent realizations of a process on a given fixed finite

lattice increases. This is called the *repeating lattice* context. In principle, this might occur if a given spatial region is observed over and over again in time, with temporal spacing great enough to result in independent realizations and under the assumption that the processes involved in producing response values has not changed over time.

To a degree, physical reality does not need to match an asymptotic context in order for asymptotic results to be applied. That is, asymptotic results are about the behavior of the model and estimators as sample size increases, not about real situations. For example, just because there is an ocean or lake on one side of a spatial study region used to observe a terrestrial response does not mean that the expanding lattice context is meaningless. At the same time, asymptotic results are directly applicable only if the overall number of observations is large enough for limiting results to provide reasonable approximations to the necessarily finite problem under consideration. Thus, in the immediately preceding example, the observed spatial lattice must be large enough for the hypothetical infinitely expanding lattice results to be meaningful. Similarly, if the asymptotic context for a problem involving groups of random variables is that the number of groups grows large for a fixed number of variables within each group, then there must be enough groups observed, even if it is physically impossible that this number could grow large without bound.

Theoretical results are typically easier to obtain for an asymptotic context under which one obtains an increasing number of independent replicates of multivariate observations than they for an asymptotic context in which the dimension of a multivariate distribution increases without bound. This corresponds to settings in which the number of groups of random variables increases, be those multiple measurements on individuals or other objects,

or multiple observations of a fixed spatial region or a process of a fixed temporal extent. The context in which one obtains larger and larger collections of random variables for a single realization of a stochastic process, such as an expanding spatial lattice or a time series of increasing length, typically requires that additional structure be imposed on a model in order to obtain repeated realizations of the same statistical or probabilistic behavior.

7.2 Forms of Asymptotic Normality

There are a number of different forms, and hence notations, that statements of asymptotic normality can take. These forms differ primarily in terms of what is used as sequences of normalizing matrices for asymptotic normality, and whether they are or are not stochastic, and whether they do or do not converge to something and, if so, what that something is.

A general form for the expression of asymptotic normality of an estimator $\hat{\boldsymbol{\theta}}_n$ is

$$\mathcal{I}_n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \text{MVN}(\mathbf{0}_q, I_q) \text{ as } n \to \infty, \tag{7.1}$$

where $\mathcal{L}$ denotes convergence in law or distribution, MVN denotes a multivariate normal distribution, $\mathbf{0}_q$ is a $q-$vector of $0's$, and I_q is the $q \times q$ identity matrix with $1's$ on the diagonal and $0s$ elsewhere. Whatever the form of $\mathcal{I}_n$ involved, the implication of this result is that, for suitably large n, we may behave as if $\hat{\boldsymbol{\theta}}_n$ has a multivariate normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\mathcal{I}_n^{-1}$. In particular, $(1-\alpha)\,100\%$ confidence intervals for individual elements of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ may be formed as

$$\hat{\theta}_k \pm z_{1-\alpha/2}\,[\hat{i}^{k,k}]; \quad k = 1, \ldots, q, \tag{7.2}$$

where $\hat{\theta}_k$ is the k^{th} element of the estimator $\hat{\boldsymbol{\theta}}_n$, $\hat{i}^{k,k}$ is the k^{th} diagonal

element of the estimated version of $\mathcal{I}_n^{-1}$, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution.

A source of potential confusion for those wishing to make use of asymptotic results in applications is that there is not necessarily a unique sequence of matrices $\mathcal{I}_n$ that will cause (7.1) to hold. Different expressions giving asymptotic normality arise from differences in the forms chosen for the sequence of normalizing matrices $\mathcal{I}_n$ and the behavior of these forms as n grows large. In the general representation of asymptotic normality, there is no particular behavior required of the sequences $\{\hat{\boldsymbol{\theta}}_n : n = 1, \ldots, \}$ and $\{\mathcal{I}_n : n = 1, \ldots, \}$ other than that (7.1) hold. Technically, even $\hat{\boldsymbol{\theta}}_n$ need not converge in probability to $\boldsymbol{\theta}$, although it would seem difficult to assert that anything useful is being learned about the possible value of $\boldsymbol{\theta}$ if $\hat{\boldsymbol{\theta}}_n$ is not a consistent sequence of estimators, and we have adopted the convention here that any estimator to be considered is consistent. Given this, each element of the sequence of normalizing matrices $\mathcal{I}_n$ must grow large without bound as $n \to \infty$, will typically depend on the value of the parameter $\boldsymbol{\theta}$, and may be stochastic in that they are also functions of $\boldsymbol{Y}$.

Suitable choices for the sequence of matrices $\mathcal{I}_n$ in (7.1) largely depends on three considerations: (1) what can be shown to exist, (2) what can be explicitly identified, and (3) what can be computed on the basis of observed data. This last consideration is, of course, crucial for making practical use of a result in application. Some of the types of $\mathcal{I}_n$ and the typical notation used for each will now be given, and an example based on maximum likelihood estimation will provide more detail on possible specific forms of these types in the next section.

1. It may be that we can choose $\mathcal{I}_n = n\,\Sigma^{-1}(\boldsymbol{\theta})$, where $\Sigma^{-1}(\boldsymbol{\theta})$ is a non-

stochastic positive definite matrix that does not depend on n. Then (7.1) is often written as

$$n^{1/2}[\Sigma^{-1}(\boldsymbol{\theta})]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0}_q,\, I_q). \tag{7.3}$$

Because $\Sigma^{-1}(\boldsymbol{\theta})$ is non-stochastic and does not depend on n one could also write

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0}_q,\, \Sigma(\boldsymbol{\theta})). \tag{7.4}$$

2. Another form for the normalizing matrices that may be appropriate is $\mathcal{I}_n = c_n\, \mathcal{H}_n(\boldsymbol{\theta})$, where $\mathcal{H}_n(\boldsymbol{\theta})$ is a non-stochastic matrix depending on n but that converges (elementwise and in the ordinary sense) to a positive definite matrix $\Sigma^{-1}(\boldsymbol{\theta})$. Here, $c_n;\ n \geq 1$ denotes a sequence of constants that depend on n. Usually, $c_n = n$ but this is neither guaranteed nor necessary. Relative to the three considerations listed previously, this form for the $\mathcal{I}_n$ may be used when we can demonstrate the existence of $\Sigma^{-1}(\boldsymbol{\theta})$ but explicitly identify only the $\mathcal{H}_n(\boldsymbol{\theta})$. Here, (7.1) might be written as,

$$c_n^{1/2}[\mathcal{H}_n(\boldsymbol{\theta})]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0}_q,\, I_q). \tag{7.5}$$

It would no longer be proper to put $\mathcal{H}_n^{-1}(\boldsymbol{\theta})$ on the right-hand side as was done for $\Sigma(\boldsymbol{\theta})$ in (7.4), because of its dependence on n.

3. Another possibility is that $\mathcal{I}_n = c_n \mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ where c_n is a sequence of constants that depend on n and $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ is a stochastic matrix that converges in probability to a non-stochastic positive definite matrix $\Sigma^{-1}(\boldsymbol{\theta})$. As for the previous situation in item (2), often we will have $c_n = n$, but this is not necessary and will not always be true. In these situations we might write (7.1) as

$$c_n^{1/2}[\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0}_q,\, I_q). \tag{7.6}$$

Similar to the use of $\mathcal{H}_n(\boldsymbol{\theta})$ in item (2), this type of choice for the $\mathcal{I}_n$ arises when we are able to demonstrate the existence of $\Sigma^{-1}(\boldsymbol{\theta})$ but identify only the $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$. Although not exhaustive of the situations under which this choice for the form of the $\mathcal{I}_n$ is possible, it is not uncommon that we are in a situation in which we could use result (7.5) with $\mathcal{H}_n(\boldsymbol{\theta}) = E\{\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})\}$, but the expectation is intractable so we use $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ instead.

4. The consistent estimator $\hat{\boldsymbol{\theta}}_n$ may often be substituted for $\boldsymbol{\theta}$ in any of $\Sigma^{-1}(\boldsymbol{\theta})$, $\mathcal{H}_n(\boldsymbol{\theta})$ or $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ to produce a sequence of matrices $\mathcal{I}_n$ in (7.1), without changing the results (7.3), (7.5) or (7.6), as long as these matrices are smooth functions of $\boldsymbol{\theta}$, which holds for many models. The technical meaning of smooth varies a bit depending upon which type of normalizing matrix in which we wish to replace $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}_n$. With $\hat{\boldsymbol{\theta}}_n$ consistent for $\boldsymbol{\theta}$, the elements of $\Sigma^{-1}(\hat{\boldsymbol{\theta}}_n)$ are consistent estimators of the elements of $\Sigma^{-1}(\boldsymbol{\theta})$ when those elements are continuous functions of $\boldsymbol{\theta}$. With $\hat{\boldsymbol{\theta}}_n$ consistent for $\boldsymbol{\theta}$ and under other conditions sufficient for $\mathcal{H}_n(\boldsymbol{\theta})$ to converge to $\Sigma^{-1}(\boldsymbol{\theta})$, $\mathcal{H}_n(\hat{\boldsymbol{\theta}}_n)$ is a consistent estimator of $\Sigma^{-1}(\boldsymbol{\theta})$ if $\mathcal{H}_n(\boldsymbol{\theta})$ is equicontinuous at $\boldsymbol{\theta}$. This means that we may choose a small distance δ and a neighborhood of $\boldsymbol{\theta}$ (that may depend on δ), call it $B(\boldsymbol{\theta}, \delta)$, such that $||\mathcal{H}_n(\boldsymbol{\theta}') - \mathcal{H}_n(\boldsymbol{\theta})|| < \delta$ for all $\boldsymbol{\theta}' \in B(\boldsymbol{\theta}, \delta)$ when n is large. Similarly, with $\hat{\boldsymbol{\theta}}_n$ consistent for $\boldsymbol{\theta}$ and under other conditions sufficient for $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ to be consistent for $\Sigma^{-1}(\boldsymbol{\theta})$, $\mathcal{J}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{Y})$ is a consistent estimator of $\Sigma^{-1}(\boldsymbol{\theta})$ if $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ is stochastically equicontinuous at $\boldsymbol{\theta}$. This means that we may choose a small distance δ and a neighborhood of $\boldsymbol{\theta}$ (that may depend on δ), call it $B(\boldsymbol{\theta}, \delta)$, such that $Pr[||\mathcal{J}_n(\boldsymbol{\theta}', \boldsymbol{Y}) - \mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})|| < \delta$ for all $\boldsymbol{\theta}' \in B(\boldsymbol{\theta}, \delta)] > 1 - \epsilon$ for any

$\epsilon > 0$ when n is large. Under these conditions we would have, in addition to (7.3), (7.5) or (7.6),

$$\begin{aligned} n^{1/2}[\Sigma^{-1}(\hat{\boldsymbol{\theta}}_n)]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) &\xrightarrow{\mathcal{L}} MVN(\mathbf{0}_q, I_q) \\ (c_n)^{1/2}[\mathcal{H}_n(\hat{\boldsymbol{\theta}}_n)]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) &\xrightarrow{\mathcal{L}} MVN(\mathbf{0}_q, I_q) \\ (c_n)^{1/2}[\mathcal{J}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{Y})]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) &\xrightarrow{\mathcal{L}} MVN(\mathbf{0}_q, I_q) \end{aligned} \tag{7.7}$$

It may be that results such as (7.3), (7.5) or (7.6) are given and it is then indicated that we would estimate the large-sample covariance matrix of $\hat{\boldsymbol{\theta}}_n$ by substituting $\hat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$ in $(1/n)\Sigma(\boldsymbol{\theta})$, $(1/c_n)[\mathcal{H}_n(\boldsymbol{\theta})]^{-1}$ or $(1/c_n)[\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})]^{-1}$, but this is only justified if one of the distributional results in (7.7) holds. That is, we are always implicitly using some scaling matrix $\Sigma(\boldsymbol{\theta})$ in (7.4), even if we cannot identify it in explicit form. If we cannot identify it we replace that scaling matrix with sequences of matrices that converge to it, either directly as for $\mathcal{H}_n(\boldsymbol{\theta})$ or in probability as for $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$. We then further replace the parameter $\boldsymbol{\theta}$ in these sequences with a consistent estimator $\hat{\boldsymbol{\theta}}_n$ and, under suitable conditions such as those presented previously, limiting results continue to hold. The distinction between using $\hat{\boldsymbol{\theta}}_n$ as a "plug-in" value for $\boldsymbol{\theta}$ to estimate $\Sigma(\boldsymbol{\theta})$ and the fact that the limiting results (7.5) and (7.6) continue to hold with $\mathcal{H}_n(\hat{\boldsymbol{\theta}}_n)$ or $\mathcal{J}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{Y})$ as in (7.7) is important in understanding the final possible form of the $\mathcal{I}_n$ in (7.1), which we now present.

5. We come now to the most confusing and potentially counter-intuitive possibility for choice of the normalizing matrices $\mathcal{I}_n$ in (7.1). It may be possible to identify (almost always stochastic) matrices $\mathcal{I}_n = \mathcal{M}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{Y})$ such that (7.1) holds, but not a sequence of constants c_n such that we could take $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y}) = (1/c_n)\mathcal{M}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ and verify that these matri-

ces converge in probability to some positive definite matrix $\Sigma^{-1}(\boldsymbol{\theta})$. The issue here is not whether one could write $\mathcal{I}_n = c_n \mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y}) = c_n(1/c_n)\mathcal{M}_n(\boldsymbol{\theta}, \boldsymbol{Y})$, which is trivially true for any c_n. The issue is whether this can be done and result in convergence in probability of $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y}) = (1/c_n)\mathcal{M}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ to some non-stochastic positive definite matrix $\Sigma^{-1}(\boldsymbol{\theta})$ that does not depend on n. If this is not the case then, although as just indicated it would be possible to develop the result (7.6), that result would be of little value because the device of item (4) in the preceding material may not hold true. That is, with $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y}) = (1/c_n)\mathcal{M}_n(\boldsymbol{\theta}, \boldsymbol{Y})$, (7.6) might hold but, if $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ in (7.6) is not consistent for some $\Sigma^{-1}(\boldsymbol{\theta})$, then the third line of (7.7) would not hold. The dramatic distinction between this case and those presented previously is that a choice of $\mathcal{I}_n$ can be identified such that (7.1) holds, but we may not have even verified the existence of a $\Sigma^{-1}(\boldsymbol{\theta})$ such that (7.3) holds. A representation of asymptotic normality in this situation assumes the general form of (7.1) with $\mathcal{I}_n = \mathcal{M}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{Y})$ as,

$$[\mathcal{M}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{Y})]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0}_q,\, I_q). \tag{7.8}$$

The implication in practice is that the covariance matrix of the limiting distribution ($\Sigma(\boldsymbol{\theta})$ in (7.3) or (7.4)) cannot really be estimated because that matrix has not necessarily even been shown to exist. Nevertheless, in principle, confidence intervals for the components of $\boldsymbol{\theta}$ can still be constructed through the use of (7.2) as long as n is sufficiently large and the inverse of $\mathcal{M}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{Y})$ can be computed.

7.3 Examples of $\mathcal{I}_n$ in a Likelihood Setting

To illustrate some explicit forms of potential $\mathcal{I}_n$ matrices, consider a problem involving n independent observations of $m-$dimensional random variables with distributions that contain q scalar parameters. We will assume the repeating lattice asymptotic context of Chapter 7.1. For what are called *regular* problems with $\ell^{(n)}(\boldsymbol{\theta}|\boldsymbol{y})$ a true log likelihood, the $\boldsymbol{Y}_i$ identically distributed, and $\hat{\boldsymbol{\theta}}_n$ a maximum likelihood estimator, $\mathcal{I}_n$ could be taken as any of the forms $n\Sigma^{-1}(\boldsymbol{\theta})$ in item 1 of the previous section, $c_n\mathcal{H}_n(\boldsymbol{\theta})$ in item 2, or $c_n\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ in item 3, both of these latter two with $c_n = n$. If the $\boldsymbol{Y}_i$ are not identically distributed, then $\Sigma^{-1}(\boldsymbol{\theta})$ would most likely not be available in closed form, but either of the other two forms might suffice.

If the distributions of the $\boldsymbol{Y}_i$ are identical then $\ell^{(n)}(\boldsymbol{\theta}|\boldsymbol{Y}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}|\boldsymbol{Y}_i) = \sum_{i=1}^n \ell(\boldsymbol{\theta}|\boldsymbol{Y}_i)$. If we choose $\mathcal{I}_n = n\Sigma^{-1}(\boldsymbol{\theta})$, then $\Sigma^{-1}(\boldsymbol{\theta})$ is the Fisher (or expected) information matrix for a single observation, a $q \times q$ matrix with jk^{th} element

$$d_{j,k} = E\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{Y}_i)}{\partial \theta_j}\,\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{Y}_i)}{\partial \theta_k}\right) = E\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{Y}_i)}{\partial \theta_j\,\partial \theta_k}\right), \tag{7.9}$$

which are constants that are the same for all $\boldsymbol{Y}_i$. The first equality in (7.9) is the definition of Fisher information and the second equality is a result that follows from one of the standard regularity conditions.

In a problem with independent but not identically distributed $\boldsymbol{Y}_i$ we might use $\mathcal{I}_n = n\mathcal{H}_n(\boldsymbol{\theta})$ as in item 2 previously. Here, $\mathcal{H}_n(\boldsymbol{\theta})$ would most likely be taken as the average Fisher information matrix $\mathcal{I}_F(\boldsymbol{\theta})$, the $q \times q$ matrix with jk^{th} element

$$i_{j,k} = \frac{1}{n^2}E\left(\sum_{i=1}^n \frac{\partial \ell_i(\boldsymbol{\theta}|\boldsymbol{Y}_i)}{\partial \theta_j}\,\sum_{s=1}^n \frac{\partial \ell_s(\boldsymbol{\theta}|\boldsymbol{Y}_s)}{\partial \theta_k}\right) = \frac{1}{n}E\left(-\sum_{i=1}^n \frac{\partial^2 \ell_i(\boldsymbol{\theta}|\boldsymbol{Y}_i)}{\partial \theta_j\,\partial \theta_k}\right) \tag{7.10}$$

Note that, if we used this same form with independent and identically distributed $\boldsymbol{Y}_i$ we would have

$$\begin{aligned} i_{j,k} &= \frac{1}{n^2} E\left(\sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{Y}_i)}{\partial \theta_j} \sum_{s=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{Y}_s)}{\partial \theta_k} \right) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{s=1}^{n} E\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{Y}_i)}{\partial \theta_j} \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{Y}_s)}{\partial \theta_k} \right) \\ &= \frac{1}{n^2} \sum_{s=1}^{n} d_{i,j} = \frac{d_{i,j}}{n}. \end{aligned}$$

If we chose to use $\mathcal{I}_n = \mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ as in item 3 previously, $\mathcal{J}_n(\boldsymbol{\theta}, \boldsymbol{Y})$ would be what is called the average observed information $\mathcal{I}_n^O(\boldsymbol{\theta}, \boldsymbol{Y})$, which is usually taken to be the $q \times q$ matrix with jk^{th} element

$$i_{j,k}^o = \frac{1}{n} \sum_{i=1}^{n} -\frac{\partial^2 \ell_i(\boldsymbol{\theta}|\boldsymbol{y}_i)}{\partial \theta_j \partial \theta_k}, \tag{7.11}$$

although an alternative form is sometimes used,

$$i_{j,k}^o = \frac{1}{n^2} \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\theta}|\boldsymbol{y}_i)}{\partial \theta_j} \sum_{s=1}^{n} \frac{\partial \ell_s(\boldsymbol{\theta}|\boldsymbol{y}_s)}{\partial \theta_k}. \tag{7.12}$$

In contrast to average Fisher information, which is the same as Fisher information for a single observation if the $\boldsymbol{Y}_i$ are identically distributed, average observed information will remain distinct even in this case. But it gives the same limiting distribution as the other two and provides a suitable alternative for the $\mathcal{I}_n$ of (7.1), particularly if the expected values in (7.9) prove difficult to evaluate. In the case of independent but not identically distributed $\boldsymbol{Y}_i$, choice of average Fisher information or average observed information is typically based on ease of computation; sometimes taking expectations can reduce the complexity of expressions and $\mathcal{I}_n$ will be taken as $\mathcal{I}_F(\boldsymbol{\theta})$ while in other cases the expectations prove intractable and $\mathcal{I}_n$ will be taken as $\mathcal{I}_n^O$.

7.4 Asymptotic Context Revisited

The example of the previous section assumed the asymptotic context of a repeating lattice (and also presumed maximum likelihood estimation). The context of a repeating lattice provides replications $\boldsymbol{Y}_i$; $i = 1, \ldots, n$ of the entire process for a fixed lattice size of size m. This greatly facilitates both the production of results giving asymptotic normality and, importantly, our ability to compute the necessary quantities involved in $\mathcal{I}_n$, whatever form has been chosen for this sequence of normalizing matrices. Specifically, if independent replicates of $\boldsymbol{Y}_i$ are available, then forms of $\mathcal{I}_n$ as average information (either expected or observed) are computed as exactly that, averages over the replicate data values $\boldsymbol{y}_i$, assumed to be realizations of the $\boldsymbol{Y}_i$.

The asymptotic context of an expanding lattice, however, provides no true replicates of the process, only a single process that is continually growing in size. Recall that in this setting, i indexes only a single location, not an entire lattice and n represents the size of that lattice. Here, not only does the development of theoretical results become more challenging, but the question of how one computes realized values of the quantities involved can complicate the application of those results. Both of these difficulties are typically approached by placing restrictions on the strength of dependence allowed in the process, a Result due to Comets and Janzura presented in the chapter on composite likelihood to follow providing a notable exception. Such restrictions on the strength of dependence may be represented in terms of what are called mixing conditions, which imply essentially that two finite subregions of the lattice approach independence as the distance between them increases (cf., Gaetan and Guyon 2010, Appendix B.2). A mixing condition is often assumed as part of the conditions attached to a theoretical

result. But there is also a practical side to this. If the approach to independence implied by a mixing condition is "fast enough," then sufficiently separated subregions of a lattice could be considered as (nearly) independent replicates of the process if each of those subregions are, at the same time, large enough to reflect the dynamics of the entire process. This, of course, renders the entire idea impractical unless one has either a huge lattice or the process exhibits almost no dependence at all. Fortunately, it is not necessary that regions large enough to capture the dynamics of the entire process be (nearly) independent. Mixing conditions similar to those that allow the development of asymptotic normality produce the same type of convergence in averages familiar from the law of large numbers (i.e., averages are consistent for their expected value). What is really needed in an application, then, is to be able to define subregions or blocks, that may be overlapping, in such a way that (1) the blocks are large enough to each reflect the spatial structure of the entire process, and (2) the blocks are small enough that many replicates are available over which to take averages. This is the essence of methodologies collectively known as spatial subsampling. The interested reader is referred to, among others, Carlstein (1986), Kunsch (1989), Sherman (1996), Heagerty and Lele (1998), Nordman and Lahiri (2003, 2004), and Politis, Romano and Wolf (1999). The implementation of spatial subsampling to produce estimated values for the covariance matrix of a limiting normal distribution is an area in need of additional investigation. Lee and Lahiri (2003) and Nordman and Caragea (2009) examine the issues involved in the context of spatial subsampling for the estimation of variograms in a geostatistical setting.

Chapter 8

Quasi-Likelihood and Estimating Equations

[Quasi-Likelihood and Estimating Equations]

8.1 Quasi-Likelihood

8.1.1 Connection with Exponential Dispersion Families

To motivate quasi-likelihoods, consider maximum likelihood estimation of generalized linear models. In deriving derivatives of the log likelihood for these models we made use first of independence among response variables to write the likelihood and its derivatives in terms of sums of contributions from individual variables Y_i, and then also used the chain rule. Consider using these same techniques, but taking the derivation only to the point of obtaining a derivative of the log likelihood with respect to the expectation

μ_i,

$$\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i} = \frac{\partial \ell_i(\mu_i,\phi)}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}.$$

$$\frac{\partial \ell(\boldsymbol{\mu},\phi)}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n}\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i}.$$

An individual response variable has an exponential dispersion family form for its density or mass function,

$$f(y_i|\theta_i,\phi) = \exp[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i,\phi)].$$

As for our discussion of basic generalized linear models in Stat 520, be aware that different authors use different forms for the dispersion parameter – my ϕ may be $1/\phi$ for another author, and both of these may be $1/a(\phi)$ for yet another author. With this form, we have $E(Y_i) = \mu_i = b'(\theta_i)$ and $var(Y_i) = (1/\phi)b''(\theta_i) = (1/\phi)V(\mu_i)$. Differentiating the logarithm of this density with respect to μ_i we obtain,

$$\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i} = \frac{\partial \ell_i(\mu_i,\phi)}{\partial \theta_i}\frac{d\theta_i}{d\mu_i} = \frac{\phi\{y_i - b'(\theta_i)\}}{b''(\theta)} = \frac{\phi\{y_i-\mu_i\}}{V(\mu_i)}. \tag{8.1}$$

Notice that expression (8.1) (and thus also the sum across i) depends only on the first two moments of Y_i, $E(Y_i) = \mu_i$ and $var(Y_i) = (1/\phi)V(\mu_i)$.

Now consider a situation in which we know that the left most term in (8.1) is equal to the rightmost term for some function $V(\cdot)$, but without the knowledge that the response distributions are exponential dispersion family. Could we "recover" the log likelihood for μ_i (for fixed ϕ) through integration? Consider

$$Q_i(\mu_i|\phi) = \int_{y_i}^{\mu_i}\frac{\phi\{y_i - t\}}{V(t)}\,dt,$$

$$Q(\boldsymbol{\mu}|\phi) = \sum_{i=1}^{n}\int_{y_i}^{\mu_i}\frac{\phi\{y_i - t\}}{V(t)}\,dt. \tag{8.2}$$

The limits of integration in (8.2) are chosen based on what we hope to obtain, which is a function that is $\ell_i(\mu_i, \phi)$ up to an additive constant. Since the integrand is the derivative of $\ell_i(\mu_i, \phi)$ with respect to μ_i, the upper limit is obvious. The lower limit is to obtain a function of y_i alone, that then becomes an additive constant for the log likelihood.

Example 8.1

Suppose $Y_1, \ldots, Y_n$ have Poisson distributions with expected values $\mu_1, \ldots, \mu_n$, but we are not aware of this fact. We are, however, given that,

$$\frac{\partial \ell(\mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{\mu_i}.$$

Applying (8.2),

$$Q_i(\mu) = \int_{y_i}^{\mu_i} \frac{y_i - t}{t}\, dt = y_i \log(\mu_i) - \mu_i, + k(y_i)$$

so that,

$$Q(\boldsymbol{\mu}) = \sum_{i=1}^{n} \{y_i \log(\mu_i) - \mu_i\} + K(\boldsymbol{y}).$$

Now, if $Y_i \sim$ iid Po(μ_i) for $i = 1, \ldots, n$,

$$\ell(\boldsymbol{\mu}) = \{y_i \log(\mu_i) - \mu_i\} + w(\boldsymbol{y})$$

and, up to an additive constant, $Q(\boldsymbol{\mu}) = \ell(\boldsymbol{\mu})$.

Note that, in general, the additive constant that distinguishes $Q(\boldsymbol{\mu})$ from $\ell(\boldsymbol{\mu})$ in this situation (which is independent response variables with exponential dispersion family distributions) depends on both $\boldsymbol{y}$ and ϕ; this will be important if the value of ϕ is to be estimated.

Now, nothing in this introductory section is anything we do in practice. Its purpose was simply to show a connection between quasi-likelihood and generalized linear models, and to motivate the notion that considering the quantity (8.2) makes some sense.

8.1.2 Basic Quasi-Likelihood

The fundamental idea underlying basic quasi-likelihood is that, even in situations for which we do not have independent random variables with fully specified exponential dispersion family distributions, the function $Q(\boldsymbol{\mu}|\phi)$ in (8.2) should behave in a manner that resembles a log likelihood function for $\boldsymbol{\mu}$.

Consider independent random variables $Y_1, \ldots, Y_n$ that have expectations $\mu_1, \ldots, \mu_n$ and variances $(1/\phi)\, V_1(\mu_1), \ldots, (1/\phi)\, V_n(\mu_n)$ for a set of specified functions $V_1(\cdot), \ldots, V_n(\cdot)$. Assume that $\mu_i = h(\boldsymbol{x}_i, \boldsymbol{\beta})$ for some known function $h(\cdot)$ and unknown parameters $\boldsymbol{\beta}$ with $\dim(\boldsymbol{\beta}) = p < n$, but specify nothing additional about the model. Notice that we have allowed the functions $V_i(\cdot)$ to vary across observations. In the majority of situations it will be reasonable to take these to be the same function, but that is not necessary. What is necessary, however, is that $var(Y_i) = (1/\phi)\, V_i(\mu_i)$ where ϕ is constant and $V_i(\cdot)$ does not depend on elements of $\boldsymbol{\mu}$ other than μ_i (McCullagh and Nelder 1989, p. 324).

In this independence situation, take the quasi-likelihood to be as given in (8.2). The quasi-score function is then,

$$\begin{aligned} U_i(\mu_i|\phi) &= \frac{\phi\{y_i - \mu_i\}}{V_i(\mu_i)} \\ U(\boldsymbol{\mu}|\phi) &= \sum_{i=1}^{n} \frac{\phi\{y_i - \mu_i\}}{V_i(\mu_i)}. \end{aligned} \tag{8.3}$$

It is easy to show that the elements of $U(\boldsymbol{\mu}|\phi)$, which are first derivatives of

$Q(\boldsymbol{\mu}|\phi)$, have the following properties:

$$\begin{aligned} E\{U_i(\mu_i|\phi)\} &= 0, \\ var\{U_i(\boldsymbol{\mu}_i|\phi)\} &= \frac{\phi}{V_i(\mu_i)} \\ -E\left\{\frac{\partial}{\partial\mu_i}U_i(\mu_i|\phi)\right\} &= \frac{\phi}{V_i(\mu_i)} \end{aligned}$$

Notice that, given the first, the third of these properties constitute an condition that is analogous to the condition in regular likelihood problems that the expected information is equal to the negative expectation of second derivatives of the log likelihood. In addition, the first property given above implies that

$$\begin{aligned} E\left\{\frac{\partial Q_i(\mu_i|\phi)}{\partial\beta_k}\right\} &= E\left\{\frac{\partial Q_i(\mu_i|\phi)}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_k}\right\} \\ &= E\left\{U_i(\mu_i|\phi)\frac{\partial\mu_i}{\partial\beta_k}\right\} = 0. \end{aligned} \tag{8.4}$$

Quasi-likelihood functions share the property given in expression (8.4) with true likelihood functions, and this then suggests that maximum quasi-likelihood estimates of $\boldsymbol{\beta}$ might be found by solving these equations (absent the expectation operator) for the elements of $\boldsymbol{\beta}$.

In a similar way, the second and third properties imply that,

$$\begin{aligned} E\left\{\frac{\partial Q_i(\mu_i|\phi)}{\partial\beta_k}\frac{\partial Q_i(\mu_i|\phi)}{\partial\beta_j}\right\} &= E\left\{\{U_i(\mu_i|\phi)\}^2\frac{\partial\mu_i}{\partial\beta_k}\frac{\partial\mu_i}{\partial\beta_j}\right\} \\ &= -E\left\{\frac{\partial U_i(\mu_i|\phi)}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_k}\frac{\partial\mu_i}{\partial\beta_j}\right\} \\ &= \frac{\phi}{V_i(\mu_i)}\frac{\partial\mu_i}{\partial\beta_k}\frac{\partial\mu_i}{\partial\beta_j}. \end{aligned}$$

Given these results, we may derive a analogy to a Fisher scoring algorithm in a manner similar to that used for maximum likelihood estimation of $\boldsymbol{\beta}$ in basic generalized linear models. Now, however, we are maximizing the quasi-likelihood function rather than the log likelihood function. The result is an algorithm that updates parameter estimates as,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \delta\boldsymbol{\beta}^{(m)},$$

where,

$$\delta\boldsymbol{\beta}^{(m)} = \left(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}\right)^{-1}\boldsymbol{D}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}. \tag{8.5}$$

In (8.5) $\boldsymbol{V}$ is the $n \times n$ diagonal matrix with entries $V_i(\mu_i)$ and $\boldsymbol{D}$ is the $n \times p$ matrix with ik^{th} element $\partial\mu_i/\partial\beta_k$; recall that p is the dimension of $\boldsymbol{\beta}$. Notice that, in this algorithm, the parameter ϕ has canceled, in the same way that it did for development of the Fisher scoring algorithm for standard generalized linear models.

Inference for maximum quasi-likelihood is based on a result that if $\tilde{\boldsymbol{\beta}}$ denotes the maximum quasi-likelihood estimator of $\boldsymbol{\beta}$, then:

(i) $\tilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

(ii) $\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $(n/\phi)\left(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}\right)^{-1}$.

Given the asymptotic normality of $\tilde{\boldsymbol{\beta}}$, inference for $\boldsymbol{\beta}$ can be based on normal theory forms with the asymptotic covariance matrix

$$cov(\tilde{\boldsymbol{\beta}}) = \frac{1}{\phi}\left(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}\right)^{-1}. \tag{8.6}$$

Use of (8.6) in practice requires an estimate of ϕ. A common estimator is the moment-based estimator,

$$\hat{\phi} = \left[\frac{1}{n-p}\sum_{i=1}^{n}\frac{\{y_i-\hat{\mu}_i\}^2}{V(\hat{\mu}_i)}\right]^{-1}.$$

When might one consider the use of quasi-likelihood in estimation and inference? One situation in which quasi-likelihood presents a viable option for estimation and inference is if a standard generalized linear model has been used, but the resulting fit exhibits a less than totally adequate description of the observed variances. That is, we formulate a model with $Y_1, \ldots, Y_n$ distributed according to an exponential dispersion family, which then dictates the variance function $V(\mu_i)$. Diagnostics, such as described in McCullagh and Nelder (1989) may reveal that this assumed variance function does not describe the data in a completely satisfactory manner. We may then choose to envisage a model with essentially the same type of random component only with a modified variance function.

Example 8.2

McCullagh and Nelder (1989, Chapter 9.2.4) present an example in which the initial model was a binomial random component with logistic link, and hence variance function $V(\mu_i) = \mu_i(1 - \mu_i)$ (the response variables Y_i were taken as proportions rather than counts). Diagnostic plots (McCullagh and Nelder 1989, p. 331) showed that the variances appeared to decrease too rapidly (as a function of μ_i) at the extremes (for μ_i close to 0 and 1). They then suggested a variance function $V(\mu_i) = \mu_i^2(1 - \mu_i)^2$ as an alternative. This variance function does not correspond to any of the standard exponential dispersion family distributions and, hence, a fully specified random component is no longer available. A quasi-likelihood analysis was then conducted of these data with

$$E(Y_i) = \mu_i; \qquad var(Y_i) = \frac{1}{\phi}V(\mu_i) = \frac{1}{\phi}\mu_i^2(1 - \mu_i)^2,$$

and with $\log\{\mu_i/(1-\mu_i)\} = \boldsymbol{x}_i^T\boldsymbol{\beta}$. This then leads to the quasi-likelihood (in μ_i),

$$Q_i(\mu_i|\phi) = (2y_i - 1)\log\left(\frac{\mu_i}{1-\mu_i}\right) - \frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}.$$

The potential drawbacks to quasi-likelihood are relatively obvious, and include:

1. The asymptotic result given for maximum quasi-likelihood estimators in this section was presented as analogous to *part* of the result of Likelihood Theorem 2 in Chapter 5 of the Stat 520 notes. The part that is missing in the quasi-likelihood result is the statement of asymptotic efficiency (part (iii) of Likelihood Theorem 2).

2. While inferential procedures are available for parameters in the systematic model component, represented in this section as $\boldsymbol{\beta}$, quasi-likelihood no longer offers a vehicle by which to make inferences about any other portion of the model such as quantiles or other functionals of the distribution.

3. Related to item 2, interpretation of results relative to the underlying scientific phenomenon or mechanism of interest becomes less well-defined. Consider Example 8.2 given previously. By replacing the variance function of a binomial with something else, we have admitted that we do not understand the observation process, since we no longer have an actual model. This should (in my opinion) cast substantial doubt on whether we have much of a grasp on modeling the scientific phenomenon of interest through the systematic model component.

Drawing on the third comment above, quasi-likelihood appears in many cases to be an attempt to account for the additional variance in an observable

process without having to go to the trouble of modeling it in an adequate manner; effort is diverted to estimation rather than modeling.

8.1.3 Extended Quasi-Likelihood

It was noted immediately following Example 8.1 that, in exponential dispersion family situations, the additive constant that separates a quasi-likelihood and the true log likelihood will depend on both $\boldsymbol{y}$ and ϕ, and we have been careful to write the quasi-likelihood and quasi-score functions as conditional on the dispersion parameter ϕ. With this conditioning, the quasi-likelihood method produces estimates for systematic model component parameters that behave in a manner similar to that of maximum likelihood estimates, minus asymptotic efficiency. If, however, we would like a quasi-likelihood function that is a sort-of likelihood in terms of ϕ as well as in terms of $\mu_i; \quad i = 1, \ldots, n$, something else is needed than what has been developed to this point. That development is the intent of what is typically called *extended quasi-likelihood* (Nelder and Pregibon 1987).

Extended quasi-likelihood functions are derived in McCullagh and Nelder (1989) by essentially pre-supposing the end product. Barndorff-Nielsen and Cox (1989) give a progression in which extended quasi-likelihood is derived as a *tilted Edgeworth* expansion (also often called a saddlepoint approximation). In either case, what results is the extended quasi-likelihood of Nelder and Pregibon (1987), written for a single random variable Y_i as,

$$\begin{aligned} Q_i^+(\mu_i, \phi) &= \int_{y_i}^{\mu_i} \frac{\phi\{y_i - t\}}{V(t)}\, dt - \frac{1}{2}\log\{2\pi(1/\phi)V(y_i)\} \\ &= Q_i(\mu_i, \phi) - \frac{1}{2}\log\{2\pi(1/\phi)V(y_i)\}. \end{aligned} \tag{8.7}$$

The only difference between $Q_i(\mu_i|\phi)$ in (8.2) and what has been written as $Q_i^+(\mu_i, \phi)$ in (8.7) is whether ϕ is considered a fixed value in the function, or part of the argument. As pointed out by McCullagh and Nelder (1989, p. 350) the justification of extended quasi-likelihood as a saddlepoint approximation depends on some type of assumption that renders the contribution of higher-order cumulants to the Edgeworth expansion used negligible. Typically, in such an expansion for a density, cumulants higher than order 4 are dropped (e.g., Stuart and Ord 1987); in derivation of extended quasi-likelihood we make this assumption for cumulants of greater than order 2.

The use of extended quasi-likelihood is perhaps epitomized by models of the generalized linear model form in which we attempt to model both the expectation and variance as functions of covariates. Note that this is the generalized linear model counterpart of additive error models with variance that depends on unknown parameters. One way to specify such models is given in McCullagh and Nelder (1989, Chapter 10.2) as,

$$E(Y_i) = \mu_i; \quad g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta};$$

$$var(Y_i) = \frac{1}{\phi_i} V(\mu_i),$$

and,

$$\phi_i = E\{d_i(Y_i, \mu_i)\}; \quad h(\phi_i) = \boldsymbol{u}_i^T \boldsymbol{\gamma};$$

$$var\{d_i(Y_i, \mu_i)\} = \frac{1}{\tau}\phi^2.$$

In the above model formulation, $d_i(Y_i, \mu_i)$ is chosen as a measure of dispersion, often,

$$d_i(Y_i, \mu_i) = \frac{(Y_i - \mu_i)^2}{V(\mu_i)}.$$

The function $h(\cdot)$ is a "dispersion link function" that connects the expected value of $d_i(Y_i, \mu_i)$, namely ϕ_i, with covariates $\boldsymbol{u}_i$ (often some or all of $\boldsymbol{x}_i$, just

as in additive error models), and the relation between ϕ and $var\{d_i(Y_i, \mu_i)\}$ is dictated if one uses extended quasi-likelihood for estimation. McCullagh and Nelder (1989, Chapter 10.5) give several additional adjustments to the extended quasi-likelihood procedure that may be desirable.

8.2 Generalized Estimating Equations

Thus far in this chapter our discussion has been entirely in the context of independent response variables Y_i; $i = 1, \ldots, n$. One of the primary areas of application for quasi-likelihood and similar ideas, however, has certainly been longitudinal studies in which random variables correspond to repeated observation of particular sampling units or individuals (e.g., Zeger and Liang 1986; Liang and Zeger 1986; Zeger, Liang, and Albert 1988). This type of situation is often dealt with using a model that implies correlated random variables within individuals. A fundamental property of this setting is that we do, in fact, have independent realizations across groups or individuals of some type of marginal model depending on the same parameters.

In the dependence setting, quasi-likelihood blends into what we will call *Estimating Functions* because the starting point is really the quasi-score rather than the quasi-likelihood. In fact, the quasi-likelihood itself is rarely mentioned or defined (but see McCullagh and Nelder 1989, Chapter 9.3.2). The term *Generalized Estimating Equations* has also been used, notably by Zeger and Liang (1986). The fundamental concept follows from noting two things about what was presented previously as the quasi-score function, which may be written, for $k = 1, \ldots, p$, as,

$$\sum_{i=1}^{n} U_i(\mu_i|\phi)\frac{\partial \mu_i}{\partial \beta_k} = \sum_{i=1}^{n} \frac{\phi\{y_i - \mu_i\}}{V_i(\mu_i)}\frac{\partial \mu_i}{\partial \beta_k}.$$

Note that, again, the roots of these equations in $\boldsymbol{\beta}$ will not involve ϕ. Secondly, note that we could use the same form of these equations with y_i and μ_i replaced by vectors $\boldsymbol{y}_i \equiv (y_{i,1}, \ldots, y_{i,m_i})^T$ and $\boldsymbol{\mu}_i \equiv (\mu_{i,1}, \ldots, \mu_{i,m_i})^T$, and $V_i(\mu_i)$ replaced by an $m_i \times m_i$ matrix $\boldsymbol{V}_i$ which gives, up to the scalar multiple ϕ, the covariance matrix of $\boldsymbol{Y}_i$. Let n now denote the number of independent vectors $\{\boldsymbol{y}_i : i = 1, \ldots, n\}$ (e.g., number of individuals in a longitudinal study). The above equations may then be written as what Zeger and Liang (1986) called generalized estimating equations,

$$\sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = 0, \tag{8.8}$$

where $\boldsymbol{S}_i \equiv (\boldsymbol{y}_i - \boldsymbol{\mu}_i)$ is $m_i \times 1$, $\boldsymbol{D}_i$ is $m_i \times p$ with jr^{th} element $\partial \mu_{i,j}/\partial \beta_r$ and $\boldsymbol{V}_i$ is $m_i \times m_i$ with structure to be given directly. For a particular model, an appropriate structure must be chosen for $\boldsymbol{V}_i$; $i = 1, \ldots, n$. Liang and Zeger (1986) suggested using a *working correlation matrix* $\boldsymbol{R}_i(\alpha)$ to help define $\boldsymbol{V}_i$ as,

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\alpha) \boldsymbol{A}_i^{1/2}, \tag{8.9}$$

where $\boldsymbol{A}_i$ is diagonal with elements given by what, in the independence case were the variance functions $V(\mu_i)$ and are now functions $w(\mu_{i,j})$ such that $var(Y_{i,j}) \propto w(\mu_{i,j})$.

Estimates of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are then found by solving the estimating equations (8.8) with the definition of $\boldsymbol{V}_i$ as in (8.9). In order to do this, a value is needed for the correlation matrix parameter α. We will discuss this, and estimation of ϕ shortly. But first, it can be helpful to describe the types of matrices $R_i(\alpha)$ that might be used in some structures.

1. Independence.

   If $R_i(\alpha) = I_{m_i}$, then the estimating equations (8.8) reduce to those

of quasi-score functions in the independence case. Nevertheless, this choice of $R(\alpha)$ is sometimes used to obtain starting values for iterative algorithms to solve the full equations (8.8) with a different choice of $R(\alpha)$.

2. Random Effects Structure.

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & \dots & \alpha \\ \vdots & \vdots & \dots & \vdots \\ \alpha & \alpha & \dots & 1 \end{pmatrix}_{m_i \times m_i} .$$

This is the correlation structure of a linear random effects model in which all variables in the same group or within the same individual (here the variables contained in $\boldsymbol{Y}_i$) have the same correlation.

3. Autoregressive Process of Order k.

Here, we would take the uv^{th} element of the matrix $\boldsymbol{R}_i$ to be

$$[\boldsymbol{R}_i]_{u,v} = \begin{cases} \alpha^{|t_{i,u} - t_{i,v}|}, & |t_{i,u} - t_{i,v}| \leq k \\ 0 & |t_{i,u} - t_{i,v}| > k, \end{cases}$$

where $t_{i,u}$ and $t_{i,v}$ are the u^{th} and v^{th} observation times of the group or individual indexed by i.

For estimation of $\boldsymbol{\beta}$, α, and ϕ, the general prescription is to utilize a Gauss-Newton type of algorithm in which one iteration for $\boldsymbol{\beta}$ is conducted using current estimates of α as fixed, then estimating new α for the current β and so forth. At each iteration, α is updated via a moment-based estimator, and ϕ is also estimated based on a moment-type estimator. We will not present the details here (since we've covered the basic ideas already), but see Zeger and Liang (1986) or Pawitan (2001) for all the particulars.

Inference about $\boldsymbol{\beta}$ is made using asymptotic normality of the GEE estimates. The form of the asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ is given in Zeger and Liang (1986) and Liang and Zeger (1986), and is estimated by using moment-based estimators $\hat{\alpha}$ and $\hat{\phi}$ as plug-in values.

8.3 Estimating Functions

The quasi-score function and generalized estimating equations of the previous sections are special cases of what are known as *estimating functions*. Given random variables $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)^T$ with a density $f_Y(\boldsymbol{y}|\boldsymbol{\theta})$, an unbiased estimating function for $\boldsymbol{\theta}$ can be defined quite simply as any function $v(\boldsymbol{Y}, \boldsymbol{\theta})$ such that, for all values of $\boldsymbol{\theta} \in \Theta$,

$$E\{v(\boldsymbol{Y}, \boldsymbol{\theta})\} = 0. \tag{8.10}$$

If we are interested in only a part of $\boldsymbol{\theta}$, then the estimating function is defined relative to only that component or components.

Example 8.3

Let $Y_1, \ldots, Y_n$ be $iid\, N(\mu, \sigma^2)$. Possible estimating functions for μ include,

$$\begin{aligned} v(\boldsymbol{y}, \mu) &= \sum_{i=1}^{n} (y_i - \mu) = 0, \\ v(\boldsymbol{y}, \mu) &= \sum_{i=1}^{n} sgn(y_i - \mu) = 0, \\ v(\boldsymbol{y}, \mu) &= \sum_{i=1}^{n} \psi(y_i - \mu) = 0, \end{aligned}$$

where in this third possibility, for some real number r,

$$\psi(x) = x(r^2 - x^2)^2 I(-r < x < r).$$

Estimates resulting from solving these three possible estimating functions for μ are the sample mean, the sample median, and a robust estimator that corresponds to use of the function $\psi(\cdot)$ which in the above is known as the *biweight* or *bisquare* function, attributed to Tukey (e.g., Hampel *et al.* 1986).

Asymptotic inference for estimating functions is developed in much the same way as for maximum likelihood estimators and, although we did not cover it in any detail, maximum quasi-likelihood estimators. Specifically, in the case of a scalar parameter θ, suppose conditions sufficient to show the following are assumed.

1. The solution to an unbiased estimating function $\tilde{\theta}$ may be expanded as

$$v(\boldsymbol{y}, \tilde{\theta}) \approx v(\boldsymbol{y}, \theta) + \left. \frac{\partial v(\boldsymbol{y}, \tilde{\theta})}{\partial \tilde{\theta}} \right|_{\tilde{\theta}=\theta} (\tilde{\theta} - \theta),$$

   from which we have, since $v(\boldsymbol{y}, \tilde{\theta}) = 0$ by definition,

$$\tilde{\theta} - \theta = -v(\boldsymbol{y}, \theta) \left\{ \left. \frac{\partial v(\boldsymbol{y}, \tilde{\theta})}{\partial \tilde{\theta}} \right|_{\tilde{\theta}=\theta} \right\}^{-1}.$$

2. The random version of $v(\cdot)$ satisfies,

$$n^{-1/2} v(\boldsymbol{Y}, \theta) \;\text{ is }\; AN(0, q(\theta)),$$

   for some function $q(\cdot)$.

3. The random version of the derivative of $v(\cdot)$ satisfies,

$$\frac{1}{n} \frac{\partial v(\boldsymbol{Y}, \theta)}{\partial \theta} \xrightarrow{p} s(\theta),$$

   for some function $s(\cdot)$.

Then the result is (Barndorff-Nielsen and Cox 1994, p.303),

$$\tilde{\theta} \;\text{ is }\; AN\left(\theta, \frac{q(\theta)}{n\, s^2(\theta)}\right).$$

There appear to be two major areas in which estimating functions surface. The first is an an alternative theory of estimation, in which estimating functions are presented as an alternative to both least squares (viewed as an exact theory procedure) and maximum likelihood. This is, for example, the view offered by Godambe and Kale (1991, p. 17) in which the authors claim to demonstrate that estimating functions "unifies the method of maximum likelihood and the method of minimum variance unbiased estimation in the case of parametric models". Under this viewpoint we must deal with the development of criteria under which one of a variety of possible estimating functions can be deemed optimal. There will be any number of unbiased estimating functions that can be developed in most situations (as illustrated in Example 8.3). A variety of possible optimality criteria are offered by Godambe and Heyde (1987). A slightly different view is offered by Barndorff-Nielsen and Cox (1994) who point out that, in the case of $Y_1, \ldots, Y_n \sim iid$ with common density $f(y|\theta)$ for a scalar θ, the two matrices $q(\cdot)$ and $s(\cdot)$ in conditions 2 and 3 listed above for asymptotic normality of $\tilde{\theta}$ become

$$q(\theta) = var\left\{\frac{\partial \log\{f(y_i|\theta)}{\partial\theta}\right\},$$

$$s(\theta) = E\left\{\frac{\partial^2 \log\{f(y_i|\theta)}{\partial\theta^2}\right\},$$

so that $q(\theta) = -s(\theta)$ and the asymptotic result leads to the solution to the likelihood equation (an estimating function) being asymptotically normal with variance given by the inverse information. This then motivates them to suggest a criterion

$$\rho_v(\theta) = \frac{[E\{\partial v(\boldsymbol{Y},\theta)/\partial\theta\}]^2}{var\{v(\boldsymbol{Y},\theta)\}},$$

as a measure of the lost efficiency of the solution to any estimating function $v(\boldsymbol{Y},\theta)$; this is because $\rho_v(\theta)$ is maximized by the likelihood score function.

The second major area in which estimating functions play a role is in the development of robust estimators, as indicated by the third possible estimating function given in Example 8.3 for the one-sample normal problem. This is the flavor of the presentation by, for example, Pawitan (2001, Chapter 14). In this context, estimating functions are also often called M-estimators. See, for example, Hampel *et. al.* (1986), and Carroll and Ruppert (1988, Chapter 7).

One aspect of estimating functions that deserves mention is that, if $v_1(\cdot), \ldots, v_p(\cdot)$ are unbiased estimating functions, then any linear combination, $\sum a_j v_j$ is also an unbiased estimating function. This property can sometimes be utilized to form simple estimating functions in even complex situations. For example, in the autoregressive process

$$Y(t) = \theta Y(t-1) + \epsilon(t),$$

where $\epsilon(t) \sim iid\, N(0,1)$, we could take estimating functions,

$$v_t = y(t) - \theta y(t-1),$$

and form the combination,

$$v(\boldsymbol{y}, \theta) = \sum_t y(t-1)\, v_t.$$

An example of such development is given in McCullagh and Nelder (1989, p.341).

Chapter 9

Composite Likelihood

Composite likelihoods appear to have originated with Lindsay (1988). This idea attracted a great deal of attention in the early 2000's, and remains an area of active research today. Overviews of this topic can be found in Varin (2008) and Varin, Reid and Firth (2011).

9.1 Definition

Consider an $m-$dimensional random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$ that has a joint probability density or mass function $f(\boldsymbol{y}|\boldsymbol{\theta})$ for some parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T \in \Theta$. The random vector $\boldsymbol{Y}$ may have component variables corresponding to observation of the same quantity on m sampling units or may have component variables corresponding to observations of different quantities on the same sampling unit. Situations in which composite likelihoods are valuable typically involve models that specify dependence structures among the component variables of $\boldsymbol{Y}$. For example, $\boldsymbol{Y}$ may correspond to observation of a single quantity at m locations in a geographical region (e.g., field,

county, or state), or observation of a single quantity at m time points for an individual patient in a clinical trial. Alternatively, $\boldsymbol{Y}$ may correspond to observations of m different attributes of a sampling unit, such as total income, total debt, and assessed value for households. Even more complex situations are possible to consider, such as observation of income, debt, and assessed value for households at given geographic locations in a city, county, or state.

Given a random vector $\boldsymbol{Y}$, consider a set of marginal and/or conditional events $\{\mathcal{A}_k : k = 1, \ldots, K\}$. For example, if the components of $\boldsymbol{Y}$ are discrete, a given event $\mathcal{A}_k$ might be the event that $Y_1 = y$ or the event $(Y_1 = y_1) \cap (Y_2 = y_2)$ or the event that $Y_1 = y_1$ given that $Y_2 = y_2$, or even the event that $(Y_1 = y_1) \cap (Y_2 = y_2) \cap \ldots \cap (Y_m = y_m)$. Let $L_k(\boldsymbol{\theta}\boldsymbol{y})$ denote the likelihood of event $\mathcal{A}_k$ that is dictated by the joint pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$. A composite likelihood is then defined as

$$L_c(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{k=1}^{K} [L_k(\boldsymbol{\theta}|\boldsymbol{y})]^{w_k}, \tag{9.1}$$

where w_k may be a non-negative weight for $k = 1, \ldots, K$. The log composite likelihood is then

$$\ell_c(\boldsymbol{\theta}|\boldsymbol{y}) = \log\{L_c(\boldsymbol{\theta}|\boldsymbol{y})\} = \sum_{k=1}^{K} w_k \, \log\{L_k(\boldsymbol{\theta}|\boldsymbol{y})\} = \sum_{k=1}^{K} w_k \, \ell_k(\boldsymbol{\theta}|\boldsymbol{y}). \tag{9.2}$$

We will assume for the remainder of this chapter that the weights w_k are all the same and may be set identically equal to 1. A value $\hat{\boldsymbol{\theta}}$ is a maximum composite likelihood estimate if, for all $\boldsymbol{\theta} \in \Theta$,

$$L_c(\boldsymbol{\theta}|\boldsymbol{y}) \leq L_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \quad \text{or} \quad \ell_c(\boldsymbol{\theta}|\boldsymbol{y}) \leq \ell_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}). \tag{9.3}$$

Clarification of what structures qualify as composite likelihoods is perhaps best obtained through examples. Typically, the events $\{\mathcal{A}_k : k = 1, \ldots, K\}$

correspond to either all marginal or all conditional events, although mixtures of marginal and conditional events would be possible.

1. Independence Likelihood

   What is often called an independence likelihood results from using k to index individual elements of $\boldsymbol{Y}$ so that $K = m$, $\mathcal{A}_k$ is the event that $Y_k = y_k$ (or $y_k - \delta < Y_k < y_k + \delta$ in the continuous case) and

$$L_c(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{k=1}^{m} f(y_k|\boldsymbol{\theta}), \tag{9.4}$$

   where $f(y_i|\boldsymbol{\theta})$ is the marginal pdf or pmf of Y_i dictated by the model $f(\boldsymbol{y}|\boldsymbol{\theta})$. This is called an independence (composite) likelihood for fairly obvious reasons – it is the likelihood we would construct if the model assumed the elements of $\boldsymbol{Y}$ were independent.

2. Pairwise Marginal Likelihood

   A pairwise marginal likelihood is constructed by taking k to index distinct pairs of elements of $\boldsymbol{Y}$ so that $K = m(m-1)/2$, $\mathcal{A}_k$ is the event that $(Y_i = y_i) \cap (Y_j = y_j)$; $(i, j) = k$ (or its continuous analog), and

$$\begin{aligned} L_c(\boldsymbol{\theta}|\boldsymbol{y}) &= \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} f(y_i, y_j|\boldsymbol{\theta}) \\ &= \prod_{1 \le i < j \le m} f(y_i, y_j|\boldsymbol{\theta}), \end{aligned} \tag{9.5}$$

   where $f(y_i, y_j|\boldsymbol{\theta})$ is the bivariate marginal pdf or pmf dictated for Y_i and Y_j by the joint model $f(\boldsymbol{y}|\boldsymbol{\theta})$. It should be clear that such a pairwise marginal likelihood could be extended to be a composite likelihood formed from triples or quadruples or margins of any dimension.

3. Pairwise Conditional Likelihood

   A pairwise conditional likelihood is constructed by taking k to index all pairs of elements of $\boldsymbol{Y}$ so that $K = m(m-1)$, $\{\mathcal{A}_k$ is the event that $Y_i = y_i$ given that $Y_j = y_j$ for $(i,j) = k$, and

$$L_c(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} \prod_{j \neq i} f(y_i|y_j, \boldsymbol{\theta}), \tag{9.6}$$

   where $f(y_i|y_j, \boldsymbol{\theta})$ are conditional pdfs or pmfs dictated by $f(\boldsymbol{y}|\boldsymbol{\theta})$.

4. Pairwise Difference Likelihood

   A pairwise difference likelihood is constructed by taking k to index distinct pairs of elements of $\boldsymbol{Y}$ so that $K = m(m-1)/2$, $\mathcal{A}_k$ is the event that $(Y_i - Y_j) = (y_i - y_j)$; $(i,j) = k$, and

$$\begin{aligned} L_c(\boldsymbol{\theta}|\boldsymbol{y}) &= \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} f(y_i - y_j|\boldsymbol{\theta}) \\ &= \prod_{1 \leq i < j \leq m} f(y_i - y_j|\boldsymbol{\theta}), \end{aligned} \tag{9.7}$$

   where $f(y_i, -y_j|\boldsymbol{\theta})$ is the pdf or pmf dictated for $(Y_i - Y_j)$ by the joint model $f(\boldsymbol{y}|\boldsymbol{\theta})$.

5. Besag's Original Pseudo-likelihood

   In the context of spatial problems formulated on a discrete-index random field, Besag (1975) suggested the use of what he called a pseudo-likelihood function for which k indexes individual elements of $\boldsymbol{Y}$ so that $K = m$, $\mathcal{A}_k$ is the event that $Y_k = y_k$ given values for $\{Y_j : j \neq k\}$ and

$$L_c(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{k=1}^{m} f(y_k|\{y_j : j \neq k\}, \boldsymbol{\theta}), \tag{9.8}$$

   where $f(y_k|\{y_j : j \neq k\})$ is the full conditional pdf or pmf dictated by $f(\boldsymbol{y}|\boldsymbol{\theta})$.

9.2 Composite Scores as Estimating Functions

Beginning with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ and $w_k = 1$ for all k in (9.2), define the composite score function as

$$U_c(\boldsymbol{\theta}|\boldsymbol{y}) = (U_1(\boldsymbol{\theta}|\boldsymbol{y}), \ldots, U_p(\boldsymbol{\theta}|\boldsymbol{y}))^T, \tag{9.9}$$

where, for $j = 1, \ldots, p$,

$$U_j(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial}{\partial \theta_j}\ell_c(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{k=1}^{K} \frac{\partial}{\partial \theta_j}\ell_k(\boldsymbol{\theta}|\boldsymbol{y}). \tag{9.10}$$

It will be useful in what follows to note that $U_c(\boldsymbol{\theta}|\boldsymbol{y})$ may also be written as,

$$U_c(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{k=1}^{K} U_k(\boldsymbol{\theta}|\boldsymbol{y}), \tag{9.11}$$

where,

$$U_k(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\frac{\partial}{\partial \theta_1}\ell_k(\boldsymbol{\theta}|\boldsymbol{y}), \ldots, \frac{\partial}{\partial \theta_p}\ell_k(\boldsymbol{\theta}|\boldsymbol{y})\right)^T. \tag{9.12}$$

The components of a composite likelihood are true likelihood "pieces" so that, as long as we can interchange differentiation and integration,

$$E\{U_c(\boldsymbol{\theta}|\boldsymbol{y})\} = \mathbf{0}. \tag{9.13}$$

Verifying that this is the case is not entirely trivial. Expression (9.1) is a general definition of a composite likelihood in which $L_k(\boldsymbol{\theta}|\boldsymbol{y})$ is either a marginal or conditional distribution dictated by the joint model $f(\boldsymbol{y}|\boldsymbol{\theta})$. Consider the representation of $U_c(\boldsymbol{\theta}|\boldsymbol{y})$ given in (9.9) and (9.10). Suppose first that $L_k(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}_s|\boldsymbol{\theta})$ where $f(\boldsymbol{y}_s|\boldsymbol{\theta})$ is the joint marginal pdf or pmf for a subset of random variables $\boldsymbol{Y}_s \subset \boldsymbol{Y}$. Then, for any $j = 1, \ldots, p$ and dominating

measure μ,

$$\begin{aligned}
E\{U_j(\boldsymbol{\theta}|\boldsymbol{y})\} &= E\left[\sum_{k=1}^{K}\frac{\partial}{\partial\theta_j}\log\{L_k(\boldsymbol{\theta}|\boldsymbol{y})\}\right] \\
&= \sum_{k=1}^{K}E\left[\frac{\partial}{\partial\theta_j}\log\{f(\boldsymbol{y}_s|\boldsymbol{\theta})\}\right] \\
&= \sum_{k=1}^{K}E\left[\frac{1}{f(\boldsymbol{y}_s|\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_j}f(\boldsymbol{y}_s|\boldsymbol{\theta})\right\}\right] \\
&= \sum_{k=1}^{K}\int_{-\infty}^{\infty}\frac{1}{f(\boldsymbol{y}_s|\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_j}f(\boldsymbol{y}_s|\boldsymbol{\theta})\right\}f(\boldsymbol{y}|\boldsymbol{\theta})\,d\mu(\boldsymbol{y}) \\
&= \sum_{k=1}^{K}\int_{-\infty}^{\infty}\frac{1}{f(\boldsymbol{y}_s|\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_j}f(\boldsymbol{y}_s|\boldsymbol{\theta})\right\}f(\boldsymbol{y}_s|\boldsymbol{\theta})\,d\mu(\boldsymbol{y}_s) \\
&= \sum_{k=1}^{K}\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta_j}f(\boldsymbol{y}_s|\boldsymbol{\theta})\,d\mu(\boldsymbol{y}_s) \\
&= \sum_{k=1}^{K}\frac{\partial}{\partial\theta_j}\int_{-\infty}^{\infty}f(\boldsymbol{y}_s|\boldsymbol{\theta})\,d\mu(\boldsymbol{y}_s) = 0. \qquad (9.14)
\end{aligned}$$

Now suppose that $L_k(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}_s|\boldsymbol{y}_r,\boldsymbol{\theta})$, where $f(\boldsymbol{y}_s|\boldsymbol{y}_r,\boldsymbol{\theta})$ is the joint conditional pdf or pmf for $\boldsymbol{Y}_s \subset \boldsymbol{Y}$ given $\boldsymbol{Y}_r \subset \boldsymbol{Y}$. Then, for any $j = 1,\ldots,p$,

$$\begin{aligned}
E\{U_j(\boldsymbol{\theta}|\boldsymbol{y})\} &= E\left[\sum_{k=1}^{K}\frac{\partial}{\partial\theta_j}\log\{L_k(\boldsymbol{\theta}|\boldsymbol{y})\}\right] \\
&= \sum_{k=1}^{K}E\left[\frac{\partial}{\partial\theta_j}\log\{f(\boldsymbol{y}_s,\boldsymbol{y}_r|\boldsymbol{\theta})\} - \frac{\partial}{\partial\theta_j}\log\{f(\boldsymbol{y}_r|\boldsymbol{\theta})\}\right]. \qquad (9.15)
\end{aligned}$$

Applying (9.14) to each term in (9.15) (starting at the second line of 9.14) shows that these expectations are also equal to zero. The result is that composite likelihoods have the form of unbiased estimating functions (e.g., Godambe 1991).

We know that for true likelihood functions $L(\boldsymbol{\theta}|\boldsymbol{y})$ and $\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \log\{L(\boldsymbol{\theta}|\boldsymbol{y})\}$,

and any $u, v = 1, \ldots, p$,

$$E\left[\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_u}\,\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_v}\right] = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_u\,\partial \theta_v}\right]. \tag{9.16}$$

as long as dominated convergence holds. In a manner similar to that given above in expressions (9.14) and (9.15), this equality will also hold for each component likelihood piece of a log composite likelihood, $\ell_k(\boldsymbol{\theta}|\boldsymbol{y})$,

$$E\left[\frac{\partial \ell_k(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_u}\,\frac{\partial \ell_k(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_v}\right] = -E\left[\frac{\partial^2 \ell_k(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_u\,\partial \theta_v}\right], \tag{9.17}$$

but it does not hold in general for the entire log composite likelihood $\ell_C(\boldsymbol{\theta}|\boldsymbol{y})$. An important quantity for composite likelihoods is what is often called Godambe Information, defined as

$$G(\boldsymbol{\theta}) = H(\boldsymbol{\theta})J^{-1}(\boldsymbol{\theta})H(\boldsymbol{\theta}). \tag{9.18}$$

In (9.18) $J(\boldsymbol{\theta})$ is a $p \times p$ matrix with uv^{th} element

$$J_{u,v}(\boldsymbol{\theta}) = E\left[\left\{\sum_{k=1}^{K}\frac{\partial \ell_k(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_u}\right\}\left\{\sum_{k=1}^{K}\frac{\partial \ell_k(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_v}\right\}\right], \tag{9.19}$$

and $H(\boldsymbol{\theta})$ is a $p \times p$ matrix with uv^{th} element

$$H(\boldsymbol{\theta}) = E\left[-\sum_{k=1}^{K}\frac{\partial^2 \ell_k(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_u\,\partial \theta_v}\right]. \tag{9.20}$$

In much of the literature on composite likelihoods these matrices are written as $J(\boldsymbol{\theta}) = E[U_C(\boldsymbol{\theta}|\boldsymbol{y})\,U_C^T(\boldsymbol{\theta}|\boldsymbol{y})]$ and $H(\boldsymbol{\theta}) = E[-\nabla U_C(\boldsymbol{\theta}|\boldsymbol{y})]$, where ∇ is the gradient operator. Notice that $J(\boldsymbol{\theta})$ gives the variance of $\ell_C(\boldsymbol{\theta})$ while $H(\boldsymbol{\theta})$ is related to the curvature of the log composite likelihood surface. For $\ell_C(\boldsymbol{\theta})$ a full likelihood these are the same, $H(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$ and (9.18) is the usual the Fisher information.

9.3 Composite Likelihood Asymptotics

We have emphasized several times the importance of asymptotic context, and that is certainly true in the case of results for maximum composite likelihood estimators. Some relevant results are sketched here for two asymptotic contexts in which composite likelihoods see use.

9.3.1 Asymptotic Context I

Here, let $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,m})^T$ denote one random variable of the type discussed up to this point, and assume that independent and identically distributed copies of this variables will be available as $\{\boldsymbol{Y}_i : i = 1, \ldots\}$. Let the common joint pdfs or pmfs of the Y_i be $f(\boldsymbol{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ is a $p-$dimensional parameter, $\boldsymbol{\theta} \in \Theta$. Continuing to take weights w_k in (9.1) to all have the value 1, the composite and log composite likelihoods based on n random variables are

$$\begin{aligned} L_C^{(n)}(\boldsymbol{\theta}|\boldsymbol{y}) &= \prod_{i=1}^{n} L_C(\boldsymbol{\theta}|\boldsymbol{y}_i) = \prod_{i=1}^{n}\prod_{k=1}^{K} L_k(\boldsymbol{\theta}|\boldsymbol{y}_i) \\ \ell_C^{(n)}(\boldsymbol{\theta}|\boldsymbol{y}) &= \sum_{i=1}^{n} \log\{L_C(\boldsymbol{\theta}|\boldsymbol{y}_i)\} = \sum_{i=1}^{n}\sum_{k=1}^{K} \log\{L_k(\boldsymbol{\theta}|\boldsymbol{y}_i)\}. \end{aligned} \tag{9.21}$$

Let $\hat{\boldsymbol{\theta}}_C^{(n)}$ denote the parameter value that maximizes $L_C^{(n)}(\boldsymbol{\theta}|\boldsymbol{y})$ or $\ell_C^{(n)}(\boldsymbol{y})$. In this first asymptotic context we suppose that m remains fixed as $n \to \infty$, so we are collecting more and more copies of the $m-$dimensional random variables. Under regularity conditions similar to those used in determining asymptotic results for full maximum likelihood estimators, we have that

$$\sqrt{n}\,[G(\boldsymbol{\theta})]^{1/2}\,(\hat{\boldsymbol{\theta}}_C^{(n)} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0},\, I_{p\times p}), \tag{9.22}$$

where $I_{p\times p}$ is the $p \times p$ identity matrix and $G(\boldsymbol{\theta})$ is given in (9.18). Notice here that $G(\boldsymbol{\theta})$ is a non-stochastic matrix that does not depend on n (the sums in (9.19) and (9.20) are over components of the composite likelihood for a single observation, not over replicate observations). We may then consider replacing $G(\boldsymbol{\theta})$ with a consistent estimator – actually, we will use estimators for $J(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ separately.

A sample version of $H(\boldsymbol{\theta})$ is

$$\hat{H}_n(\boldsymbol{\theta}) = -\frac{1}{n}\nabla U_C(\boldsymbol{\theta}|\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_C^{(n)}}, \tag{9.23}$$

which means that $\hat{H}_n(\boldsymbol{\theta})$ is $p \times p$ with uv^{th} element

$$h_{u,v} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\frac{\partial^2 \ell_k(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\theta_u\,\partial\theta_v}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_C^{(n)}}. \tag{9.24}$$

Another possible replacement for $H(\boldsymbol{\theta})$ arises because (9.17) does hold for each component piece of a composite likelihood, and we could take the elements of $\hat{H}_n(\boldsymbol{\theta})$ to be given by

$$h_{u,v} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\frac{\partial \ell_k(\boldsymbol{\theta}|\boldsymbol{y}_i)}{\partial\theta_u}\,\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y}_i)}{\partial\theta_v}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_C^{(n)}}. \tag{9.25}$$

A sample version of $J(\boldsymbol{\theta})$ is

$$\hat{J}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[U_C(\boldsymbol{\theta}|\boldsymbol{y}_i)\,U_C^T(\boldsymbol{\theta}|\boldsymbol{y}_i)\right]\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_C^{(n)}}, \tag{9.26}$$

so that $\hat{J}_n(\boldsymbol{\theta})$ is a $p \times p$ matrix with uv^{th} element

$$j_{u,v} = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}\frac{\partial \ell_k(\boldsymbol{\theta}|\boldsymbol{y}_i)}{\partial\theta_u}\sum_{k=1}^{K}\frac{\partial \ell_k(\boldsymbol{\theta}|\boldsymbol{y}_i)}{\partial\theta_v}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_C^{(n)}}. \tag{9.27}$$

There are several potential difficulties with the use of sample versions of $H(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ in practice. First, how large n needs to be relative to p in

order for (9.22) to provide a good approximation when sample versions of the Godambe information are used is not easily determined. In addition, and importantly, even though a composite likelihood may have a simpler form than a full likelihood, derivatives needed to determine computational versions of $H(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ can be quite complex, leaving computation a still challenging task. One alternative to the use of sample versions of $G(\theta)$ is to estimate the covariance matrix of $\hat{\boldsymbol{\theta}}_C^{(n)}$ through the use of a jackknife. Let $\hat{\boldsymbol{\theta}}$ denote a maximum composite likelihood estimate for a given set of data $\boldsymbol{y} = \{\boldsymbol{y}_i : i = 1, \ldots, n\}$. Let $\boldsymbol{y}_{-i} = \{\boldsymbol{y}_j : j \neq i\}$ be the data with the i^{th} observation removed and $\hat{\boldsymbol{\theta}}_{-i}$ the composite likelihood based on $\boldsymbol{y}_{-i}$. A jackknife estimate of the covariance of the limiting normal distribution is

$$\hat{C}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\boldsymbol{\theta}}_{-i} - \hat{\boldsymbol{\theta}}\right) \left(\hat{\boldsymbol{\theta}}_{-i} - \hat{\boldsymbol{\theta}}\right)^T \tag{9.28}$$

Another alternative that is common in applications is to produce interval estimates directly through the use of parametric bootstrap, rather than rely on asymptotic normality of the maximum composite likelihood estimator with some estimated or computed covariance matrix. Part of the motivation for use of composite likelihoods is usually computational simplicity in the first place, so producing sets of bootstrap estimates is frequently not a major problem.

9.3.2 Asymptotic Context II

In the second asymptotic context we will assume that we have only one random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$ with joint pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ is a $p-$dimensional parameter, $\boldsymbol{\theta} \in \Theta$. Continuing to take weights w_k in (9.1) to all have the value 1, the composite and log composite

likelihoods are

$$\begin{aligned} L_C^{(m)}(\boldsymbol{\theta}|\boldsymbol{y}) &= \prod_{k=1}^{K} L_k(\boldsymbol{\theta}|\boldsymbol{y}) \\ \ell_C^{(m)}(\boldsymbol{\theta}|\boldsymbol{y}) &= \sum_{k=1}^{K} \log\{L_k(\boldsymbol{\theta}|\boldsymbol{y})\}. \end{aligned} \tag{9.29}$$

Let $\hat{\boldsymbol{\theta}}_C^{(m)}$ denote the parameter value that maximizes $L_C^{(m)}(\boldsymbol{\theta}|\boldsymbol{y})$ or $\ell_C^{(n)}(\boldsymbol{y})$ and consider $m \to \infty$. Replicate observations of $\boldsymbol{Y}$ are not available, and asymptotic results are more difficult to arrive at that for Asymptotic Context I. Essentially what is needed are restrictions on the model that produce internal replications of the same statistical behaviors, such as are produced by assumptions of stationarity. The two fundamental problems that arise are (1) determining the existence of a sequence of normalizing matrices $\mathcal{I}_m$ such that, as $m \to \infty$, $[\mathcal{I}_m]^{-1/2}(\hat{\boldsymbol{\theta}}_C^{(m)} - \boldsymbol{\theta})$ converges in distribution to a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance given by the $p-$dimensional identity matrix $I_{p\times p}$, and (2) determining how to compute $\mathcal{I}_m$.

There are some general results available for certain groups of regularity conditions, but in general the situation is more fragmented than previously, that is, there is less common structure to be exploited and hence a much greater variety of results under a greater variety of conditions. We will give here one example, that being for a composite likelihood given as Besag's original pseudo-likelihood estimator. Let $\boldsymbol{y}_{-i}$ denote the variables $\{y_j : j \neq i\}$ (don't get the index i here confused with the use of i in the previous section – here, i denotes a particular component of the $m-$vector $\boldsymbol{y}$. Besag's log pseudo-likelihood can be defined as

$$\ell_C^{(m)} = \sum_{i=1}^{m} \log\{f(y_i|\boldsymbol{y}_{-i}, \boldsymbol{\theta})\} = \sum_{i=1}^{m} \ell_i(\boldsymbol{\theta}|\boldsymbol{y}). \tag{9.30}$$

Comets and Janzura (1998) present asymptotic normality of maximum pseudo-likelihood estimators for a spatial setting on a regular lattice. Here, let $N_i;\ i = 1, m$ denote the neighborhoods of the m components of $\boldsymbol{Y}$. Impose a condition of "translation-invariant" conditional specification, but without placing conditions on joint distributions or the strength of statistical dependence associated with a model (more on this to follow). Notationally, let $\ell_i(\boldsymbol{\theta}|\boldsymbol{y}) = \log\{f(y_i|\boldsymbol{y}(N_i))$ be the i^{th} term of Besag's original log pseudo-likelihood (9.30). With all vectors being considered column vectors and the superscript T denoting transpose, the gradient of the log pseudo-likelihood can be written as an element-wise sum of vectors as,

$$U_n^T = \sum_{i=1}^{n} U_i^T = \sum_{i=1}^{n} \left(\frac{\partial \ell_i(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1}, \ldots, \frac{\partial \ell_i(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_p} \right). \tag{9.31}$$

Let $H_i(\boldsymbol{\theta}, \boldsymbol{Y})$ denote the $p \times p$ matrix of negative second derivatives of $\ell_i(\boldsymbol{\theta}|\boldsymbol{y})$ with uv^{th} element

$$h_{u,v} = -\frac{\partial^2 \ell_i(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_u \, \partial \theta_v}; \quad u, v = 1, \ldots, p. \tag{9.32}$$

Note that, in contrast to the previous section, the elements of H_i are not expected values. The result of Comets and Janzura (see also Gaetan and Guyon, 2010, Theorem 5.5) may now be stated as follows.

Result (Comets and Janzura)

Suppose that $\boldsymbol{Y}$ is a Markov random field on a regular integer lattice with translation-invariant conditional specification and let N_i denote the neighborhood of Y_i. If $\hat{\boldsymbol{\theta}}$ is the original maximum pseudo-likelihood estimator of $\boldsymbol{\theta}$, then

$$[\mathcal{J}_m(\boldsymbol{\theta}, \boldsymbol{Y})]^{-1/2} \, H_m(\boldsymbol{\theta}, \boldsymbol{Y})[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] \xrightarrow{\mathcal{L}} \text{MVN}\ (\mathbf{0}, I_{p \times p}), \tag{9.33}$$

where $H_m(\boldsymbol{\theta}, \boldsymbol{Y}) = \sum_{i=1}^{m} H_i(\boldsymbol{\theta}, \boldsymbol{Y})$ with these terms as given in (9.32), and

$\mathcal{J}_m(\boldsymbol{\theta}, \boldsymbol{Y})$ is

$$\mathcal{J}_m(\boldsymbol{\theta}, \boldsymbol{Y}) = \sum_{i=1}^{m} U_i \sum_{j \in \{i \cup N_i\}} U_j, \tag{9.34}$$

with U_i as in (9.31).

There are two aspects of this result that make it particularly worthy of note. First is the condition of translation-invariant conditional specification, which does not involve conditions on the joint distribution or, notably, the strength of dependence, and this is in contrast to most asymptotic results for pseudo-likelihood estimation. For a formal statement of this condition see Gaetan and Guyon (2010, pp. 63-64). The condition of translation-invariance for conditional specifications is, however, rather strong in its own way. Essentially, translation-invariance of a conditional specification implies that the conditional distributions specified in the model are identical for each location. Note that this eliminates non-constant mean structure in space, or dependence that varies over time for a lattice that includes locations defined in both space and time. A second aspect of the Result that is interesting is the form of the inner summation in (9.34). An algebraic identity gives that

$$U^T U = \sum_{i=1}^{m} U_i \sum_{j=1}^{m} U_j, \tag{9.35}$$

and $U^T U$ or its average $(1/m)\, U^T U$ appears in any number of asymptotic results in the literature; in a straight likelihood problem, for example, $(1/m)\, U^T U$ is one possible estimator of Fisher information. In (9.34), however, the inner summation is over only the union of each site with its neighbors, not over all sites. To my knowledge, this makes the result of Comets and Janzura unique in the identification of sequences of normalizing matrices $\mathcal{I}_m$. As described previously, this result implies that, if m is sufficiently large, we may behave as if $\hat{\boldsymbol{\theta}}_C^{(m)}$ has a normal distribution with covariance

matrix $\mathcal{I}_m^{-1} = H_m^{-1}(\boldsymbol{\theta}, \boldsymbol{Y})\, \mathcal{J}_m(\boldsymbol{\theta}, \boldsymbol{Y})\, H_m^{-1}(\boldsymbol{\theta}, \boldsymbol{Y})$. The matrices J_m and H_m can be estimated by evaluating them at the maximum pseudo-likelihood (or composite likelihood) estimator.

Another result for Besag's original pseudo-likelihood is given by Guyon (1995). In this result stronger conditions are assumed than in the result of Comets and Janzura. The conditions assumed by Guyon lead to maximum pseudo-likelihood estimators such that

$$\hat{\boldsymbol{\theta}}_C^{(m)} \;\; AN\ (\mathbf{0}, \Sigma), \tag{9.36}$$

where

$$\Sigma^{-1} = H(\boldsymbol{\theta})\, J^{-1}(\boldsymbol{\theta})\, H(\boldsymbol{\theta}),$$

with these matrices defined as for composite likelihood in the previous section. Note, however, that evaluation of $H(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ may prove difficult.

9.3.3 The Importance of Asymptotic Results

One could argue that difficulties in applying asymptotic results, primarily in computing relevant covariance matrices, have made the topic of asymptotic inference for composite likelihood largely one of greater importance for statistical development than for application, at least for problems involving the asymptotic context of a single multivariate observation that expands in dimension. Experience with composite likelihoods even in the simpler situation in which replicate observations can be obtained of the same multivariate distribution seems to suggest that direct computation of sample versions of covariances may not be immediate in these cases either. Certainly for problems that would fall under the heading of Asymptotic Context II in these notes, the majority of applications have relied on the production of interval

estimates of parameter values through the use of parametric bootstrap procedures. If this is the case, one may wonder why I have devoted so much space to a discussion of asymptotic results for composite likelihood estimation. The basic reason is that the existence of asymptotic results provides some necessary justification to the use of parametric bootstrap procedures. Parametric bootstrap methods are based on simulation with the objective of learning about the sampling distribution of an estimator for the finite sample size that exists in a given problem (through the sampling distribution of some function of the estimator and the true parameter). In order for this to be meaningful we need some assurance that such a distribution exists. Any finite collection of values will give a set of relative frequencies that looks like a distribution, regardless of whether those values have been produced from a mechanism that approximates a true distribution or not. If we assume that our simulation procedure is approximating a distribution when in fact it is not, inferences drawn on the basis of the empirical relative frequencies in the set of simulated values are meaningless and potentially seriously misleading, even if we are not aware of the fact because our simulated values look like they are describing a distribution. That this is more than just a superfluous technical fine point may be seen by considering a simulation intended to illustrate the basic central limit theorem as taught in statistical methods courses. If multiple samples are simulated from a Cauchy distribution (that has no expected value or variance), centered by the location parameter (mode or median) and scaled by the square root of the sample variances divided by the square root of sample size, the resulting collection of values will describe an empirical distribution, but an empirical distribution that is approximating nothing. It is true that the empirical distribution will be quite unattractive in shape, but if we did not know what was supposed to happen we might

naively assume that this empirical distribution is approximating the sampling distribution of standardized sample means. Any inferential quantities formed on the basis of this empirical distribution would be quite misleading indeed. The same danger exists in settings with much greater complexity, and our ability to intuitively or empirically detect when things have gone awry is severely limited. Thus, the existence of asymptotic results that apply to estimators are necessary to lend reasonable assurance that a simulated empirical distribution is, in fact, approximating a distribution that actually exists and may be used to construct inferential quantities. A result that leads to a limiting distribution implies that there exists a sequence of distributions for finite sample sizes, and it is one of these that an empirical distribution of simulated values is then approximating.

Chapter 10

Parametric Bootstrap

What is known as a *parametric bootstrap* is a widely applicable method for assessing uncertainty in parameter estimates, often in the form of interval estimation, although the formation of prediction regions or intervals is also a clear area of application. We will present the parametric bootstrap in the case of a scalar parameter θ, although it appears that the same essential ideas could be easily extended to confidence regions for vector-valued parameters. In particular, since the parametric bootstrap is a version of simulation-based inference, using a model with $\boldsymbol{\theta} \equiv (\theta_1, \boldsymbol{\theta}_2)$ and simulating in the manner to be described below results in an assessment of the marginal distribution of an estimator $\hat{\theta}_1$. Parametric bootstrap may certainly be applied to problems in which we are able to locate simultaneous maximum likelihood estimates, but more often it sees use with estimators of alternative forms such as those we have discussed in the past several chapters.

10.1 Basic Bootstrap Estimators

To set the basic notation for this section, let $Y_1, \ldots, Y_n$ be random variables that follow a model giving joint density or mass function $f(y_1, \ldots, y_n|\theta)$. Suppose that an estimator of θ, $\hat{\theta}$, is available by some means (maximum likelihood, least squares, quasi-likelihood, an estimating function, or composite likelihood). We assume that $\hat{\theta}$ is a function of the observations, $\hat{\theta} \equiv \hat{\theta}(\boldsymbol{y})$, whether or not that function can be expressed in closed form. Substitution of $\hat{\theta}(\boldsymbol{y})$ for θ in the model gives the *fitted model* with density or mass function $f(y_1, \ldots, y_n|\hat{\theta}(\boldsymbol{y}))$ and distribution function $F(y_1, \ldots, y_n|\hat{\theta}(\boldsymbol{y}))$. In these expressions it must be understood that the arguments $y_1, \ldots, y_n$ are arbitrary, while $\boldsymbol{y}$ in $\hat{\theta}(\boldsymbol{y})$ is the observed data from a particular study. The basic simulation process is to generate observations $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)$ from the fitted model and then calculate the estimator of θ from these simulated values, which we will denote $\hat{\theta}(\boldsymbol{y}^*)$. If this process is repeated M times, we obtain the bootstrap estimates $\{\hat{\theta}(\boldsymbol{y}_m^*) : \ m = 1, \ldots, M\}$. The underlying idea is that the distribution of a function of the true parameter and the actual estimate, $h(\theta, \hat{\theta}(\boldsymbol{y}))$, can be approximated by the empirical distribution of that same function of $\hat{\theta}(\boldsymbol{y})$ and the bootstrap estimates $\hat{\theta}(\boldsymbol{y}_m^*)$.

What we have is essentially a Monte Carlo procedure with a twist. Rather than computing a Monte Carlo approximation to the expected value of a simple function of $\hat{\theta}$, we are attempting to obtain a Monte Carlo approximation to the expected value of a function of $\hat{\theta}$ and the true parameter value θ. Because the true parameter is unknown, we use the actual estimate $\hat{\theta}(\boldsymbol{y})$ as the true parameter in simulation. The reason for using a function of estimator and parameter $h(\theta, \hat{\theta}(\boldsymbol{y}))$ is to mitigate the effect of the true parameter on the sampling distribution of the estimator. A simple example will help make

this clear. Consider estimating the mean of a normal distribution μ based on a data set $\boldsymbol{y}$ assumed to be a realization of a one-sample normal model. The estimator, no matter how you approach the problem, is the sample mean $\hat{\mu}(\boldsymbol{Y}) = \bar{Y}$. The sampling distribution of $\bar{Y}$ depends on the parameter μ, but the distribution of $h(\mu, \bar{Y}) = \bar{Y} - \mu$ does not. Thus, the empirical distribution of $\{h(\bar{y}, \bar{y^*}_m) = (\bar{y^*}_m - \bar{y});\ m = 1, \ldots, M\}$ will not depend on the actual estimate $\bar{y}$ and can be taken as an approximation to the distribution of $h(\mu, \bar{Y})$. Now, it is not always possible to determine a function $h(\theta, \hat{\theta}(\boldsymbol{Y}))$ that is a pivotal quantity, although that is the ideal. The concept is that the use of this comparison or discrepancy function can help reduce, if not eliminate, the dependence of the sampling distribution of $\hat{\theta}(\boldsymbol{Y})$ on the parameter θ.

The point of the previous paragraph is essential but is also easily forgotten or ignored because Monte Carlo estimators using one of the most common discrepancy functions, the simple difference, often appear to simply use the empirical distribution of bootstrap estimates at the true parameter value $\hat{\theta}(\boldsymbol{Y})$ to approximate the sampling distribution of the estimator $\hat{\theta}(\boldsymbol{Y})$ at the actual true parameter θ. Although it increases our notational burden a bit, a practical mechanism to avoid making errors is to only bootstrap quantities that have asymptotic distributions. In our context this means that disrepancy functions chosen for use in applications should always correspond to appropriately centered and scaled versions of estimators for which a limit distribution is known to exist. This serves as a very effective way to avoid committing potentially serious mistakes. To simplify notation at this point we will write $\hat{\theta}_n$ rather than $\hat{\theta}(\boldsymbol{y})$ and $\theta^*_{n,m}$ rather than $\hat{\theta}(\boldsymbol{y}^*_m)$. Consider a maximum likelihood estimator $\hat{\theta}(\boldsymbol{y})$ that satisfies typical regularity conditions. Then we know that $n^{1/2}(\hat{\theta}_n - \theta)$ has a limit distribution and we might

take

$$h(\theta, \hat{\theta}_n) = n^{1/2}(\hat{\theta}_n - \theta).$$

Consider approximating the expected value of this discrepancy function,

$$\begin{aligned} E_M\{h(\theta, \hat{\theta})\} &= E_M\{n^{1/2}(\hat{\theta}_n - \theta)\} \\ &= \frac{1}{M}\sum_{m=1}^{M} n^{1/2}(\theta^*_n - \hat{\theta}_n)_m \\ &= n^{1/2}(\bar{\theta^*}_n - \hat{\theta}_n). \end{aligned} \tag{10.1}$$

An approximation to the variance of $h(\theta, \hat{\theta}_n)$ is,

$$\begin{aligned} V_M\{h(\theta, \hat{\theta}_n)\} &= V_M\{n^{1/2}(\hat{\theta}_n - \theta)\} \\ &= \frac{n}{M}\sum_{m=1}^{M}\left[(\theta^*_n - \hat{\theta}_n) - (\bar{\theta^*}_n - \hat{\theta}_n)\right]^2_m \\ &= \frac{n}{M}\sum_{m=1}^{M}(\theta^*_{n,m} - \bar{\theta^*}_n)^2. \end{aligned} \tag{10.2}$$

Now an approximation to bias$(\hat{\theta}_n) = E\{\hat{\theta}_n - \theta)\}$ is easily obtained from (10.1) and an approximation to $var(\hat{\theta}_n - \theta) = var(\hat{\theta}_n)$, may be obtained from (10.2).

The fact that we are not using the empirical distribution of bootstrap estimates as a *direct* approximation to the sampling distribution of $\hat{\theta}_n$ may become more clear if we consider using the empirical distribution function of $\{h(\hat{\theta}_n, \theta^*_{n,m};\ m = 1, \ldots, M\}$ to approximate the distribution function of $h(\theta, \hat{\theta}_n)$. In this case,

$$\begin{aligned} G_M(u) &= \frac{1}{M}\sum_{m=1}^{M} I[h(\hat{\theta}_n, \theta^*_n)_m \le u] \\ &= \frac{1}{M}\sum_{m=1}^{M} I[n^{1/2}(\theta^*_{n,m} - \hat{\theta}_n) \le u] \\ &= \frac{1}{M}\sum_{m=1}^{M} I[\theta^*_{n,m} \le \hat{\theta}_n + u/n^{1/2}]. \end{aligned} \tag{10.3}$$

In (10.3) $I(A)$ is the indicator function that assumes a value of 1 if A is true and a value of 0 otherwise. It should now be clear that in parametric bootstrap methods the empirical distribution of bootstrap estimates is not being used to approximate the distribution of $\hat{\theta}_n$. In fact, the sampling distribution of $\hat{\theta}_n$ is not being approximated at all by these methods. It is the sampling distribution of $h(\theta, \hat{\theta}_n)$ that is the target of approximation.

In order to form bootstrap interval estimates of θ we will need to approximate specific quantiles of the distribution of $h(\theta, \hat{\theta}_n)$. To do so we use the general result that, if $X_1, \ldots, X_N$ are independently distributed with distribution function F_X, and if $X_{[1]}, \ldots, X_{[N]}$ denote the ordered values, then

$$E\{X_{[k]}\} \approx F^{-1}\left(\frac{k}{N+1}\right),$$

leading to the estimate of the q^{th} quantile of F_X, which is $x_q = F_X^{-1}(q)$ as,

$$x_q = X_{[(N+1)q]}.$$

Thus, the estimated value of the q^{th} quantile of $h(\theta, \hat{\theta}_n)$ is the $(M+1)q^{th}$ largest value of $\{h(\hat{\theta}_n, \theta^*_{n,m}) : m = 1, \ldots, M\}$, assuming that $(M+1)q$ is an integer, which we can always arrange, or we can use the largest integer less than or equal to $(M+1)q$ in its place.

10.2 Bootstrap Confidence Intervals

In this section we will consider three types of bootstrap confidence intervals the most important being what are called basic bootstrap intervals. Basic intervals developed using different discrepancy functions $h(\theta, \hat{\theta}_n)$ result in what are sometimes considered different types of intervals, but it is only the particular form chosen for this function that actually changes.

10.2.1 Normal Approximation Intervals

Consider a situation in which we would be willing to form intervals based on asymptotic normality of $\hat{\theta}_n$, but the variance or standard error of the limit distribution is unavailable or extremely difficult to compute. If $\hat{\theta}_n$ is AN with mean $\theta + B(\hat{\theta}_n)$ and variance $V(\hat{\theta}_n)$ then an approximate normal interval can be derived in the usual way from,

$$Pr\{L_\alpha \leq \theta \leq U_\alpha\} = 1 - \alpha,$$

which leads to

$$\begin{aligned} L_\alpha &= \hat{\theta_n} - B(\hat{\theta}_n) - V^{1/2}(\hat{\theta}_n) z_{1-\alpha/2} \\ U_\alpha &= \hat{\theta}_n - B(\hat{\theta}_n) + V^{1/2}(\hat{\theta}_n) z_{1-\alpha/2}, \end{aligned} \tag{10.4}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. To see this, note that it is the quantity

$$\frac{\hat{\theta}_n - B(\hat{\theta}_n) - \theta}{V^{1/2}(\hat{\theta}_n)},$$

that is approximately $N(0,1)$. The derivation then proceeds exactly as in elementary statistics courses using $\hat{\theta}_n - B(\hat{\theta}_n)$ in place of an unbiased estimator of θ.

To use the interval (10.4) in practice requires only estimation of $B(\hat{\theta}_n)$ as given in (10.1) and $V(\hat{\theta}_n)$ as given in (10.2). A bootstrap normal approximation interval for θ is then,

$$\left(\hat{\theta}_n - B_M(\hat{\theta}_n) - V_M^{1/2}(\hat{\theta}_n) z_{1-\alpha/2},\ \ \hat{\theta}_n - B_M(\hat{\theta}_n) + V_M^{1/2}(\hat{\theta}_n) z_{1-\alpha/2}\right). \tag{10.5}$$

10.2.2 Basic Bootstrap Intervals

Now suppose we are in a situation in which we are reluctant to make direct use of an asymptotic result to compute intervals. We can derive an interval

with equal probability in each tail, that is, equal left-tail and right-tail errors by finding L_α and U_α such that

$$\begin{aligned} Pr\{\theta \leq L_\alpha\} &= \alpha/2 \\ Pr\{\theta \leq U_\alpha\} &= 1-\alpha/2. \end{aligned} \tag{10.6}$$

If $h(\theta, \hat{\theta}_n)$ does not reverse inequalities then

$$\begin{aligned} Pr\{h(\theta, \hat{\theta}_n) \leq h(L_\alpha, \hat{\theta}_n)\} &= \alpha/2 \\ Pr\{h(\theta, \hat{\theta}_n) \leq h(U, \hat{\theta}_n)\} &= 1-\alpha/2. \end{aligned} \tag{10.7}$$

Then $h(L_\alpha, \hat{\theta}_n)$ is the $\alpha/2$ quantile and $h(U_\alpha, \hat{\theta}_n)$ is the $1-\alpha/2$ quantile of the distribution of $h(\theta, \hat{\theta}_n)$. Let these be dentoed as $v_{\alpha/2}$ and $v_{1-\alpha/2}$, respectively. If $h(\theta, \hat{\theta}_n)$ does reverse inequalities, then

$$\begin{aligned} Pr\{h(\theta, \hat{\theta}_n) \leq h(L_\alpha, \hat{\theta}_n)\} &= 1-\alpha/2 \\ Pr\{h(\theta, \hat{\theta}_n) \leq h(U, \hat{\theta}_n)\} &= \alpha/2. \end{aligned} \tag{10.8}$$

In this case, $h(L_\alpha, \hat{\theta}_n) = v_{1-\alpha/2}$ is the $1-\alpha/2$ quantile and $h(U_\alpha, \hat{\theta}_n) = v_{\alpha/2}$ is the $\alpha/2$ quantile of the distribution of $h(\theta, \hat{\theta}_n)$.

The q^{th} quantile of the distribution of $h(\theta, \hat{\theta}_n)$ is approximated by the parametric bootstrap as

$$\tilde{v}_q = h(\hat{\theta}_n, \theta^*_n)_{[(M+1)q]}. \tag{10.9}$$

To obtain endpoints of an interval for θ then requires solving $h(L_\alpha, \hat{\theta}_n) = \tilde{v}_q$ for L and $h(U_\alpha, \hat{\theta}_n) = \tilde{v}'_q$ for U, where q and q' may be $\alpha/2$ or $1-\alpha/2$ depending on the situation.

Difference Discrepancy Function

Consider the discrepancy function $h(\theta, \hat{\theta}_n) = n^{1/2}(\hat{\theta}_n - \theta)$. This function reverses the inequalities in (10.6) so that

$$\begin{aligned} \hat{\theta}_n - L_\alpha &\approx (\theta^*_{n,m} - \hat{\theta}_n)_{[(M+1)(1-\alpha/2)]} \\ \hat{\theta}_n - U_\alpha &\approx (\theta^*_{n,m} - \hat{\theta}_n)_{[(M+1)(\alpha/2)]} \end{aligned}$$

which results in

$$\begin{aligned} L_\alpha &\approx \hat{\theta}_n - (\theta^*_{n,m} - \hat{\theta}_n)_{[(M+1)(1-\alpha/2)]} \\ &= \hat{\theta}_n - (\theta^*_{n,m})_{[(M+1)(1-\alpha/2)]} + \hat{\theta}_n \\ &= 2\hat{\theta}_n - (\theta^*_{n,m})_{[(M+1)(1-\alpha/2)]}, \end{aligned}$$

and,

$$\begin{aligned} U_\alpha &\approx \hat{\theta}_n - (\theta^*_{n,m} - \hat{\theta}_n)_{[(M+1)(\alpha/2)]} \\ &= \hat{\theta}_n - (\theta^*_{n,m})_{[(M+1)(\alpha/2)]} + \hat{\theta}_n \\ &= 2\hat{\theta}_n - (\theta^*_{n,m})_{[(M+1)(\alpha/2)]}. \end{aligned}$$

A $(1-\alpha)100\%$ interval estimate for θ is then,

$$\left(2\hat{\theta}_n - (\theta^*_{n,m})_{[(M+1)(1-\alpha/2)]},\, 2\hat{\theta}_n - (\theta^*_{n,m})_{[(M+1)(\alpha/2)]}\right). \tag{10.10}$$

The interval (10.10) is called the *basic bootstrap confidence interval* for θ by Davison and Hinkley (1997), but it is really the basic interval that results from the discrepancy function $h(\theta, \hat{\theta}_n) = n^{1/2}(\hat{\theta}_n - \theta)$. Note that this interval estimate assumes that things have been arranged so that $(M+1)(\alpha/2)$ and $(M+1)(1-\alpha/2)$ are integers. This is not difficult if M is large; for example, take $\alpha/2 = 0.05$ and $M = 9999$ for a 90% interval.

The normal approximation interval given in expression (10.4) has the property of being symmetric, in this case about $\hat{\theta}_n - B(\hat{\theta}_n)$; it will be symmetric about $\hat{\theta}_n$ if that estimator is known to be unbiased, in which case we take $B(\hat{\theta}_n) = 0$ rather than estimating it. The interval (10.10) however, is not necessarily symmetric. It will be symmetric or close to symmetric as the empirical distribution of $\{\theta^*_{n,m} : m = 1, \ldots, M\}$ is symmetric or close to symmetric.

Ratio Discrepancy Function

Now consider using the discrepancy function $h(\theta, \hat{\theta}_n) = n^{1/2}(\theta/\hat{\theta}_n - 1)$. We would consider this function only for θ that are restricted to a strictly positive parameter space, in which case it does not reverse the inequality of (10.6) so that,

$$\begin{aligned}
\left[n^{1/2}\left(\frac{L_\alpha}{\hat{\theta}_n} - 1\right)\right] &\approx \left[n^{1/2}\left(\frac{\hat{\theta}_n}{\theta^*_n} - 1\right)\right]_{[(M+1)(\alpha/2)]} \\
\left[n^{1/2}\left(\frac{U_\alpha}{\hat{\theta}_n} - 1\right)\right] &\approx \left[n^{1/2}\left(\frac{\hat{\theta}_n}{\theta^*_n} - 1\right)\right]_{[(M+1)(1-\alpha/2)]}
\end{aligned}$$

These relations imply that

$$\begin{aligned}
\frac{L_\alpha}{\hat{\theta}_n} &\approx \left(\frac{\hat{\theta}_n}{\theta^*_n}\right)_{[(M+1)(\alpha/2)]} \\
\frac{U_\alpha}{\hat{\theta}_n} &\approx \left(\frac{\hat{\theta}_n}{\theta^*_n}\right)_{[(M+1)(1-\alpha/2)]}
\end{aligned}.$$

Now,

$$\begin{aligned}
\left(\frac{\hat{\theta}_n}{\theta^*_n}\right)_{[(M+1)(\alpha/2)]} &= \hat{\theta}_n \frac{1}{\theta^*_{n,[(M+1)(1-\alpha/2)]}} \\
\left(\frac{\hat{\theta}_n}{\theta^*_n}\right)_{[(M+1)(1-\alpha/2)]} &= \hat{\theta}_n \frac{1}{\theta^*_{n,[(M+1)(\alpha/2)]}}
\end{aligned}$$

resulting in

$$\begin{aligned} L_\alpha &\approx \hat{\theta}_n^2 \frac{1}{\theta^*_{n,[(M+1)(1-\alpha/2)]}} \\ U_\alpha &\approx \hat{\theta}_n^2 \frac{1}{\theta^*_{n,[(M+1)(\alpha/2)]}} \end{aligned}$$

and a $(1-\alpha)100\%$ interval estimate for θ is

$$\left(\hat{\theta}_n^2 \frac{1}{\theta^*_{n,[(M+1)(1-\alpha/2)]}},\ \hat{\theta}_n^2 \frac{1}{\theta^*_{n,[(M+1)(\alpha/2)]}} \right). \tag{10.11}$$

Studentized Discrepancy Function

As a final example, suppose we use the discrepancy function

$$h(\theta, \hat{\theta}_n) = \frac{\hat{\theta}_n - \theta}{[V(\hat{\theta}_n)]^{1/2}},$$

where $V(\hat{\theta}_n)$ denotes the estimated variance of $\hat{\theta}_n$. Notice that to make use of this disrepancy function in practice we need to be able to produce an estimated variance for $\hat{\theta}_n$ from any given sample. This function reverses the inequality of (10.6) so that,

$$\begin{aligned} \frac{\hat{\theta}_n - L_\alpha}{[V(\hat{\theta}_n]^{1/2})} &\approx \left(\frac{\theta^*_n - \hat{\theta}_n}{[V(\theta^*_n)]^{1/2}} \right)_{[(M+1)(1-\alpha/2)]} \\ \frac{\hat{\theta}_n - U_\alpha}{[V(\hat{\theta}_n]^{1/2})} &\approx \left(\frac{\theta^*_n - \hat{\theta}_n}{[V(\theta^*_n]^{1/2})} \right)_{[(M+1)(\alpha/2)]}. \end{aligned}$$

In this case,

$$\begin{aligned} L_\alpha &\approx \hat{\theta}_n - [V(\hat{\theta}_n)]^{1/2} \left(\frac{\theta^*_n - \hat{\theta}_n}{[V(\theta^*_n)]^{1/2}} \right)_{[(M+1)(1-\alpha/2)]} \\ U_\alpha &\approx \hat{\theta}_n - [V(\hat{\theta}_n)]^{1/2} \left(\frac{\theta^*_n - \hat{\theta}_n}{[V(\theta^*_n)]^{1/2}} \right)_{[(M+1)(\alpha/2)]} \end{aligned}$$

and a $(1-\alpha)100\%$ interval estimate for θ is

$$\left(\hat{\theta}_n - [V(\hat{\theta}_n)]^{1/2}\left(\frac{\theta_n^* - \hat{\theta}_n}{[V(\theta_n^*)]^{1/2}}\right)_{[(M+1)(1-\alpha/2)]}, \hat{\theta}_n - [V(\hat{\theta}_n)]^{1/2}\left(\frac{\theta_n^* - \hat{\theta}_n}{[V(\theta_n^*)]^{1/2}}\right)_{[(M+1)(\alpha/2)]}\right). \tag{10.12}$$

The interval in (10.12) is often called a *studentized* bootstrap confidence interval. Studentized intervals have been shown in simulation studies to have quite nice coverage properties, and are generally considered to be superior to basic intervals using a difference discrepancy function. The drawback to these intervals is that we must be able to compute an estimated variance for the estimator from the actual data and also each bootstrap data set. This is a drawback, of course, because the motivation for using a bootstrap procedure to obtain an interval estimate is decreased if we have other options available, which we almost certainly will if we can compute $V(\hat{\theta}_n)$ from either an exact or asymptotic result.

10.3 Percentile Bootstrap Intervals

We will briefly mention one other approach to the formulation of bootstrap confidence intervals, called *percentile methods* by Davison and Hinkley (1997, Chapter 5.3). The origin of the name is not intuitively obvious but presumably comes from the fact that we end up using the $(\alpha/2)$ percentile and $(1-\alpha/2)$ percentile of the empirical distribution of bootstrap estimates θ^* as interval endpoints, as will be shown below. Percentile methods have been modified (or adjusted) in a number of ways that seem to offer some improvement over normal approximation or basic bootstrap intervals in meeting nominal coverage goals, although assessments have been made primarily through the use of nonparametric bootstrap sampling (see, e.g., Davison and Hinkley

1997, Chapter 5.4 and Chapter 5.7). The performance of percentile methods in parametric bootstrap, which is our concern here, is less well understood.

Suppose that there exists some transformation of the estimator $\hat{\theta}_n$, say $\hat{\phi}_n \equiv W(\hat{\theta}_n)$, such that the distribution of $\hat{\phi}_n$ is known to be symmetric; for the moment the existence of such a function $W(\cdot)$ is all that matters, not its identity. Consider, then, applying this transformation to $\hat{\theta}_n$ and then using the basic bootstrap method with difference discrepancy function to form an interval for $\phi = W(\theta)$, with the following modifications. First, notice that we now have estimates of functionals of the distribution of $h(\phi, \hat{\phi}_n)$ through the bootstrap simulations $\{h(\hat{\phi}, \phi_n^*)_m : m = 1, \ldots, M\}$. In the basic bootstrap method we take $h(\phi, \hat{\phi}_n) = \hat{\phi}_n - \phi$ and $h(\hat{\phi}, \phi_n^*)_m = \phi_{n,m}^* - \hat{\phi}_n$. The development of the basic bootstrap interval for ϕ proceeds in the same way as before and we want $h(L_\alpha - \hat{\phi}_n) = \hat{\phi}_n - L_\alpha = v_{1-\alpha/2}$ and $h(U_\alpha - \hat{\phi}_n) = \hat{\phi}_n - U_\alpha = v_{\alpha/2}$ Now, however, the symmetry of the distribution of $\hat{\phi}_n - \phi$ indicates that $v_{1-\alpha/2} = -v_{\alpha/2}$ so that,

$$\begin{aligned} \hat{\phi} - L_\alpha &= -v_{\alpha/2} \\ \hat{\phi} - U_\alpha &= -v_{1-\alpha/2}, \end{aligned}$$

where v_α is now a quantile from the distribution of $\hat{\phi}_n - \phi$. Then,

$$\begin{aligned} L_\alpha &= \hat{\phi}_n + (\phi_n^* - \hat{\phi}_n)_{[(M+1)(\alpha/2)]} \\ &= (\phi_n^*)_{[(M+1)(\alpha/2)]} \\ U_\alpha &= \hat{\phi}_n + (\phi_n^* - \hat{\phi}_n)_{[(M+1)(1-\alpha/2)]} \qquad (10.13) \\ &= (\phi_n^*)_{[(M+1)(1-\alpha/2)]}. \qquad (10.14) \end{aligned}$$

Now, if the transformation $W(\cdot)$ that produced ϕ from θ was monotone, then $\phi_{[k]}^*$ corresponds to $\theta_{[k]}^*$ for any integer $k \in \{1, \ldots, M\}$. Transforming

the interval endpoints (10.13) back to the θ scale then results in the bootstrap percentile interval for θ of,

$$\left((\theta_n^*)_{[(M+1)(\alpha/2)]},\ \ (\theta_n^*)_{[(M+1)(1-\alpha/2)]}\right). \tag{10.15}$$

What is surprising, then, is that such an interval can be formulated (and computed) for θ without ever determining what the transformation $\phi = W(\theta)$ might be.

10.4 Predication Intervals

Another use of parametric bootstrap is in forming prediction intervals for a new random variable Y^0, assumed to follow the same model as $Y_1, \ldots, Y_n$. Let the model evaluated at a possible value y^0 be denoted as $F(y^0|\theta)$. Given an estimated parameter $\hat{\theta}_n$ the natural starting point is an interval with endpoints given by the $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of the estimated model $F(y^0|\hat{\theta})$. Denote these values as

$$\begin{aligned} q(\hat{\theta}_n)_{\alpha/2} &= F^{-1}(\alpha/2\,|\,\hat{\theta}_n) \\ q(\hat{\theta}_n)_{1-\alpha/2} &= F^{-1}(1-\alpha/2\,|\,\hat{\theta}_n). \end{aligned} \tag{10.16}$$

The interval $q_(\hat{\theta}_n)\alpha/2$, $q(\hat{\theta}_n)_{1-\alpha/2})$ will be overly optimistic (i.e., too short) because it does not take into account uncertainty in the estimation of θ by $\hat{\theta}_n$. That is, if we knew the true value θ, it would be the case that

$$Pr\left[q(\theta)_{\alpha/2} \le Y^0 < q(\theta)_{1-\alpha/2}\,|\,\theta\right] = 1-\alpha.$$

Since we do not know θ but are estimating it with $\hat{\theta}_n$ we need to assess the actual coverage rate,

$$Pr\left[q(\hat{\theta}_n)_{\alpha/2} \le Y^0 < q(\hat{\theta}_n)_{1-\alpha/2}\,|\,\theta\right] = 1-c(\alpha). \tag{10.17}$$

If there is a functional relation between $c(\alpha)$ and α, then we could "adjust" the procedure to use $q(\hat{\theta}_n)_{\alpha'/2}$ and $q(\hat{\theta}_n)_{1-\alpha'/2}$, where α' is chosen such that $c(\alpha') = \alpha$. The essential problem, then, is estimation of $c(\alpha)$ in expression (10.17). A parametric bootstrap may be used for this estimation in the following way.

The function of θ and $\hat{\theta}_n$ to be estimated is,

$$\begin{aligned} h(\theta, \hat{\theta}_n) &= Pr\left[q(\hat{\theta}_n)_{\alpha/2} \leq Y^0 < q(\hat{\theta}_n)_{1-\alpha/2} \,|\, \theta\right] \\ &= E\left\{I\left[q(\hat{\theta}_n)_{\alpha/2} \leq Y^0 < q(\hat{\theta}_n)_{1-\alpha/2} \,|\, \theta\right]\right\}. \end{aligned}$$

Given a fitted model through $\hat{\theta}_n$, simulate bootstrap data sets $\boldsymbol{y}^*_m \equiv (y^*_1, \ldots, y^*_n)^T_m$ in the usual way from $F(y_1, \ldots, y_n|\hat{\theta})$ to obtain bootstrap estimates $\theta^*_{n,m}$; $m = 1, \ldots, M$. Also simulate values of the predictand y^0_m; $m = 1, \ldots, M$ from the fitted model $F(y^0|\hat{\theta}_n)$. Compute the intervals $(q^*(\theta^*_n)_{(\alpha/2)},\ q^*(\theta^*_n)_{(1-\alpha/2)})_m$ with nominal coverage $1-\alpha$ for each bootstrap data set as in 10.16 with θ^*_n in place of $\hat{\theta}_n$.

Estimate the actual coverage of the interval as,

$$1 - c_M(\alpha) = \frac{1}{M}\sum_{m=1}^{M} I\left[\left(q^*(\theta^*_n)_{(\alpha/2)} \leq y^0_m < q^*(\theta^*_n)_{(1-\alpha/2)}\right)_m\right]. \qquad (10.18)$$

Expression (10.18) is then a bootstrap estimate of the probability (10.17). There are then two options. One might simply report $1 - \hat{c}(\alpha)$ as the actual coverage, or one might relate $\hat{c}(\alpha)$ to α through some empirical model (e.g., a quadratic regression of $\hat{c}(\alpha)$ on α might provide a good description of the relation). In the latter case, we can attempt to select an appropriate value α' to use in expression (10.16) in calculating $q(\hat{\theta}_n)_{\alpha'/2}$ and $q(\hat{\theta}_n)_{1-\alpha'/2}$ to provide an actual coverage at level $1-\alpha$.

10.5 Dependence and Other Complications

The usefulness of parametric bootstrap is perhaps the greatest in situations for which we have an estimator $\hat{\theta}$ but it is difficult to derive the variance or distribution of $\hat{\theta}$. At the same time, we have presented parametric bootstrap methods for sets of independent random variables $Y_1, \ldots, Y_n$. This does seem somewhat incongruous, since it is situations in which we fail to have independence among response variables that most often leads to the inability to make use of distributional results for the purposes of inference. As pointed out by Davison and Hinkley (1997) the theoretical underpinnings of using bootstrap methods with models that contain complex dependence structures (e.g., spatial models) are both unresolved and an area of intensive research. This still remains true today, although any number of advances have been made since the late 1990s, largely in the area of nonparametric bootstrap. Nevertheless, the use of simulation from fitted parametric models seems to hold great potential for a large number of problems.

Consider problems which might be amenable to asymptotic inference. Underlying the development of theoretical properties of bootstrap estimators (either parametric or nonparametric) are two levels of asymptotics. At one level is the convergence of the distribution of $h(\hat{\theta}_n, \theta^*_m)$ computed from bootstrap data sets to the distribution of $h(\theta, \hat{\theta}_n)$. Here, we must recall that θ^*_m is a function of the bootstrap sample, $\boldsymbol{y}^*_m \equiv (y^*_1, \ldots, y^*_n)^T_m$, so that suppressed in this notation is the fact that each $\boldsymbol{y}^*_m$ is of dimension n and each θ^*_m is based on $\boldsymbol{y}^*_m$. The fact that bootstrap samples $\boldsymbol{y}^*_m$ have been independently generated certainly helps in demonstrating this convergence as M increases. A difficulty is if the distribution of $h(\theta, \hat{\theta}_n)$ depends heavily on the value of θ, since we are using $\hat{\theta}_n$ in the role of the true value of θ for bootstrap simula-

tions. The ideal situation is if $h(\theta, \hat{\theta}_n)$ is a *pivotal* quantity; recall this means that the distribution of $h(\theta, \hat{\theta}_n)$ is independent of θ. Unfortunately, we can often only demonstrate this in a definite manner for fairly simple models, in which case we may have alternative procedures than bootstrap for computing inferential quantities. We may, however, always examine the dependence of the distribution of $h(\theta, \hat{\theta}_n)$ on θ in the following way. Let $Q_M(h|\theta^{(k)})$ denote the estimated q^{th} quantile of $h(\theta, \hat{\theta}_n)$ based on a bootstrap simulation of size M with data generated from the model with parameter value $\theta^{(k)}$. That is, $Q_M(h|\theta^{(k)}) = h(\theta^{(k)}, \theta^*)_{[(M+1)q]}$. If values $Q_M(h|\theta^{(k)})$ are computed for a range of values, $\theta^{(k)} \in \{\hat{\theta}_n \pm k\delta : k = 1, \ldots, K\}$ for some δ, then a simple plot of $Q_M(h|\theta^{(k)})$ against $\theta^{(k)}$ may demonstrate the degree of dependence of $h(\theta, \hat{\theta}_n)$ on θ, or at least the relative dependence for several possible choices of $h(\cdot)$.

The second level of convergence needed arises because, in order for inference about θ based on $\hat{\theta}_n$ to have discernible properties, it is necessary that the distribution of $h(\theta, \hat{\theta}_n)$ allows some description of its probabilistic behavior as n grows large. This issue involves the proverbial *not loosing sight of the forest for the trees* as it applies to bootstrap methods, even outside of an asymptotic context. Consider, for example, estimating μ from a model that gives $Y_1, \ldots, Y_n \sim iid\, N(\mu, \sigma^2)$. My estimator of choice will be $\hat{\mu}_n = 0.1Y_{[2]} + 0.9Y_{[n-3]}$, where $Y_{[k]}$ denotes the k^{th} largest value of the set $\{Y_i : i = 1, \ldots, n\}$. Note, among other things, that here $E\{\hat{\mu}_n\} = \mu$ so $\hat{\mu}_n$ is an unbiased estimator. By taking $h(\mu, \hat{\mu}_n) = (\hat{\mu}_n - \mu)$ I will be perfectly capable of estimating the distribution of $h(\mu, \hat{\mu}_n)$ through a parametric bootstrap, forming bootstrap intervals, and so forth. This clearly does not, however, offer any justification for my choice of $\hat{\mu}_n$ in the first place.

An additional issue that can sometimes arise in applications of parametric

bootstrap involves the combined issues of *simulation error* and *statistical error*. We present one hypothetical example to illustrate this.

A Multinomial Model

An important environmental characteristic of riverine ecosystems is the distribution of "sediment types" over the bottom of the river. Sediment types are often placed into fairly broad categories such as sand, silt, clay, and gravel. These categories have to do with a combination of particle size and organic matter content of the substrate. Sediment type is one of the factors that determine the abundances of many types of aquatic invertebrates and plants and, consequently, things that depend on them such as fish. Sediment type is also related to various characteristics of water quality, such as clarity and the availability of dissolved oxygen to aquatic life. Ecologists are interested in these relations both from the viewpoint of scientific understanding and to improve prediction of productivity. Consider small portion of a large river such as the Mississippi or Yangtze from which we have obtained n sediment samples from main channel, side channel and backwater habitats, each of which will be categorized as either clay, sand, silt, or gravel, depending on the predominant composition of the sample. Our primary objective is to compare the distributions of sediment types among the habitats. We might well approach this problem based on a multinomial data model with four categories. Suppose that we would like to use a parametric bootstrap to construct confidence intervals for the probabilities of sediment types. Maximum likelihood estimates of the multinomial parameters based on $n = 100$ samples were $\hat{\theta}_1 = 0.19$, $\hat{\theta}_2 = 0.33$, $\hat{\theta}_3 = 0.41$, and $\hat{\theta}_4 = 0.07$ for clay, sand, silt and gravel, respectively. In a parametric bootstrap, many simulated data

sets will result in 0 values for gravel, which makes estimation for these simulated realizations of the model difficult or impossible. It is not clear how this problem should be dealt with. One approach, which has been used in any number of similar situations, is to condition the simulation estimates of uncertainty on estimable data sets by simply discarding any simulated data sets that do not allow estimation. This must result in an underestimation of uncertainty, but uncertainty about *what*; the values of parameters, the model itself, or the error of simulation? Put another way, should the simulation of unestimable data sets from an otherwise perfectly acceptable model impact error of simulation, statistical error in estimation, or error in model selection?

Chapter 11

Simulation Based Model Assessment

11.1 Fundamental Concepts

One of the goals of statistical modeling is to capture the key elements of a scientific mechanism in a small number of model parameters. Estimation of those parameters yields the fitted model. The fundamental concept of simulation based model assessment is that if a model provides a sound conceptualization of the scientific mechanisms that led to the actual data, then the fitted model should be capable of generating data that behave in a manner similar to the actual data that were generated by those scientific mechanisms. What are needed to put this concept into practice are:

1. Data realizations from a fitted model.

2. A test statistic. This can be a measure of discrepancy between either a fitted model and a data set, a measure of discrepancy between two

data sets, or a quantification of some aspect of interesting behavior in a set of data.

3. A reference distribution for the test statistic chosen. We will achieve this through simulation.

The second and third items in this list are inter-twined with each other, and require additional discussion. Consideration of these aspects of simulation-based model assessment will lead to three situations; (1) a discrepancy measure between a data set and a fitted model is available but either a theoretical reference distribution is not available, or we choose not to use such a reference distribution, (2) a discrepancy measure between two data sets is available but either a theoretical reference distribution is not available or we choose not to use one if it is available, and (3) a measure of some interesting aspect of data behavior is available but we have no theoretical reference distribution available. We will consider the first two situations in the case of independent response variables. The third situation lends itself readily to either independence cases or to models with more complex dependence structures such as longitudinal settings or spatial models.

11.2 Formulating Test Statistics

11.2.1 Discrepancy Between a Data Set and a Fitted Model

Most of the available quantities used as goodness of fit statistics constitute measures of discrepancy between a data set and a model fitted to the data set. A few of the more commonly used statistics for independent random

variables are briefly reviewed here. The setting for all that follows in this subsection is that we have a model fitted to independent random variables $Y_1, \ldots, Y_n$ that has resulted in a set of estimated expected values $\hat{\mu}_1, \ldots, \hat{\mu}_n$, or a set of fitted probability mass or density functions $f_1(y|\hat{\boldsymbol{\theta}}), \ldots, f_n(y|\hat{\boldsymbol{\theta}})$.

1. Chi-Squared Statistic.

   A traditional goodness of fit measure is the Chi-squared statistic,

   $$D = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{var}(\hat{\mu}_i)}, \tag{11.1}$$

   which has, under the hypothesized model, a limiting χ^2 distribution with $n - p$ degrees of freedom where p is the number of parameters estimated.

2. Deviance.

   As was noted in the development of deviance residuals, the overall deviance for exponential dispersion families can sometimes be used as a goodness of fit statistic. If the dispersion parameter ϕ is known (e.g., $\phi = 1$ for binomial or Poisson random components), the deviance has a limiting χ^2 distribution with $n - p$ degrees of freedom, where p is the number of estimated parameters in a model. For many models this distributional result does not hold, but the scaled deviance using an estimated value of the dispersion parameter $\hat{\phi}$ still provides a measure of discrepancy between the data and a fitted model.

3. Power-Divergence Statistics.

   An entire family of goodness of fit statistics was proposed by Read and Cressie (1988) as the family of power divergence statistics. Suppose that the model under consideration is either for discrete random variables with possible values $y_i \in \{C_1, \ldots, C_k\}$, or that we have binned

a model for continuous random variables into k categories $C_1, \ldots, C_k$. For $j = 1, \ldots, k$, let X_j denote the observed frequency with which the response variables $Y_1, \ldots, Y_n$ take on the value or belong to the category C_j. Suppose further that the model may be used to calculate marginal probabilities for either the possible data values (discrete case) or category membership (continuous case). In either case, let these estimated probabilities be denoted as $\hat{\pi}_j; \;\; j = 1, \ldots, k$. The family of power divergence statistics is defined as, for $-\infty < \lambda < \infty$,

$$D_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} X_j \left[\left(\frac{X_j}{n\hat{\pi}_j} \right)^\lambda - 1 \right]. \tag{11.2}$$

In many cases, an asymptotic χ^2 distribution is available for D_λ under the hypothesized model. This may not always be the case, however, particularly in models for continuous variables in which parameters are estimated from the density form of the model (e.g., with a likelihood or log likelihood defined in terms of densities) and D_λ is applied to categories from a subsequent binning procedure (e.g., Read and Cressie, 1998, Chapter 4.1).

The family of power divergence statistics is indexed by the parameter λ and includes a number of traditional statistics such as Pearson's Chi-square ($\lambda = 1$) and the likelihood ratio statistic for multinomial data (limit as $\lambda \to 0$). Read and Cressie (1998) suggest a generally useful value of $\lambda = 2/3$, but it seems to me that one of the strengths of the power divergence statistic is what it might reveal as λ varies. This family of statistics increases in power against *bump* alternatives as λ gets larger and positive, and increases in power for *dip* alternatives as λ gets larger and negative. A bump alternative corresponds to one or

more cell frequencies substantially larger than under the hypothesized (or fitted) model, and a dip alternative corresponds to a cell frequency substantially smaller than under the hypothesized model. Thus, computing D_λ over a range of values for λ would seem to provide valuable information (see, e.g., Kaiser and Finger, 1996).

4. Kolmogorov-Smirnov Statistics.

   Consider here a set of random variables $\{Y_i : i = 1, \ldots, n\}$ that are not only independent but also identically distributed according to a theoretical distribution with density $f(y|\boldsymbol{\theta})$. Kolmogorov-Smirnov statistics are based on the empirical distribution function, defined for $-\infty < y < \infty$ as,

$$G_n(y) \equiv \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq y), \tag{11.3}$$

   where $I(A)$ is the identity function that takes on a value of 1 if A is true and 0 otherwise. Any model for *iid* random variables produces a theoretical distribution function, usually with parameters that are to be estimated. Within the context of this subsection, the estimated distribution function can be written as

$$F(y|\hat{\boldsymbol{\theta}}) = \int_{-\infty}^{y} f(t|\hat{\boldsymbol{\theta}})\, dt.$$

   The Kolmogorov-Smirnov statistics are,

$$\begin{aligned} D^+ &= \sup\left\{G_n(y) - F(y|\hat{\boldsymbol{\theta}})\right\} \\ D^- &= \sup\left\{F(y|\hat{\boldsymbol{\theta}}) - G_n(y)\right\} \\ D &= \max\{D^+, D^-\} \\ D' &= D^+ + D^- \end{aligned} \tag{11.4}$$

   The last quantity in (11.4) is often called Kuiper's statistic.

A good deal of work has been conducted on determining the distributions of these and other statistics based on the empirical distribution function under various settings (see Chapter 4 of D'Agostino and Stephens, 1986, for a review). Our concern, as with the other discrepancy measures presented, will be to make use of these statistics in a simulation-based assessment procedure.

5. Cramer-von Mises Statistic.

   The Cramer-von Mises statistic is also based on the empirical distribution function (11.3) and is generally presented as a statistic useful in testing a hypothesized distribution F_0 as,

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(y) - F_0(y)]^2 \, dF_0(y)$$

   A computational form of this statistic for ordered data observations $y_{[1]} \leq y_{[2]} \leq \ldots, \leq y_{[n]}$ is

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_0(y_{[i]}) - \frac{2i-1}{2n} \right)^2 . \tag{11.5}$$

   As for Kolmogorov-Smirnov and related statistics, a good deal of work has been conducted on distributional theory for the Cramer-von Mises statistic, generally in an asymptotic framework. Again, as for the other statistics presented here, our concern is simply the possible use of this statistic as a measure of discrepancy.

6. Generalized Residuals.

   An extremely flexible procedure for measuring the discrepancy between a set of data and a fitted model is based on what may be called generalized residuals; this terms comes from a paper by Cox and Snell

(1968). Without going into detail on the concept of generalized residuals proposed by Cox and Snell, we will simply define what we will call generalized residuals for sets of independent continuous and discrete random variables. Given continuous independent random variables $Y_1, \ldots, Y_n$ with a model that implies theoretical distribution functions $F_i(y_i|\boldsymbol{\theta})$; $i = 1, \ldots, n$ having a common parameter $\boldsymbol{\theta}$, define the residual quantities,

$$r_i \equiv F_i(y_i|\hat{\boldsymbol{\theta}}) = \int_{-\infty}^{y_i} f_i(t|\hat{\boldsymbol{\theta}})\, dt; \quad i = 1, \ldots, n. \tag{11.6}$$

If the model is representative of the data, the $\{r_i : i = 1, \ldots, n\}$ should behave in a manner similar to a sample from the uniform distribution on $(0, 1)$. The probability integral transform would hold if the parameter $\boldsymbol{\theta}$ were used in (11.6) rather than an estimate $\hat{\boldsymbol{\theta}}$. While this result does not actually hold when we use $\hat{\boldsymbol{\theta}}$ as in (11.6), we expect that the residuals r_i; $i = 1, \ldots, n$ should provide diagnostic quantities useful to detect gross discrepancies between the model and observed responses.

To define similar residuals corresponding to discrete random variables, let $Y_1, \ldots, Y_n$ be independent random variables with a common set of possible values, ordered as $\Omega \equiv \{y_{[1]}, y_{[2]}, \ldots, y_{[m]}\}$. Take the ordered value of observation i to be the q^{th} ordered value, that is, $y_i = y_{[q]}$. Define the (random) generalized residual r_i' to be the realized value of a random variable with distribution uniform on the interval $(F_i(y_{[q-1]}|\hat{\boldsymbol{\theta}}),\ F_i(y_{[q]}|\hat{\boldsymbol{\theta}}))$, that is,

$$r_i' \equiv u_i; \quad \text{where } U_i \sim iid\ U\left(F_i(y_{[q-1]}|\hat{\boldsymbol{\theta}}),\ F_i(y_{[q]}|\hat{\boldsymbol{\theta}})\right). \tag{11.7}$$

Similar to the residuals of expression (11.6), these residuals should behave in the same manner as a sample of *iid* uniform variables on the

interval (0, 1). A set of residuals $\{r_i' : i = 1, \ldots, n\}$ will not, however, be unique for a given set of observations $\{y_i : i = 1, \ldots, n\}$.

Any of the goodness of fit statistics presented previously (e.g., Kolmogorov-Smirnov or Cramer-von Mises) could then be used to measure the discrepancy of these generalized residuals with a uniform distribution. With estimated parameters in a fitted model, the distribution of generalized residuals will not be uniform, and this has stymied their use in model assessment in the past. If the number of observations is large relative to the number of estimated parameters, however, the distribution of generalized residuals should be quite similar to a uniform distribution, and statistics that compare their empirical distribution to a theoretical uniform distribution on the unit interval should provide useful measures of discrepancy in a simulation-based procedure.

11.2.2 Discrepancy Between Two Data Sets

Overall discrepancy between two sets of data may be quantified using two-sample versions of some of the goodness of fit statistics presented previously as discrepancy measures between a set of data and a fitted model. Among these are the two-sample versions of Kolmogorov-Smirnov statistics and the Cramer-von Mises statistic. To formalize, let $G_n(y)$ and $H_m(y)$ denote the empirical distribution functions of two sets of data, one of size n and the other of size m; in simulation-based procedures we will typically have $n = m$ but that is not strictly necessary. The two-sample Kolmogorov-Smirnov statistics

are then,

$$\begin{aligned} D^+ &= \sup\{G_n(y) - H_m(y)\} \\ D^- &= \sup\{H_m(y) - G_n(y)\} \\ D &= \max\{D^+, D^-\} \\ D' &= D^+ + D^- \end{aligned} \tag{11.8}$$

Let $\boldsymbol{y} = \{y_i : i = 1, \ldots, n\}$ denote the observations from one set of data, and $\boldsymbol{y}^* = \{y_j^* : j = 1, \ldots, m\}$ the observations from the other set of data. The two-sample Cramer-von Mises statistic can be written as (c.f., Conover, 1980),

$$D = \frac{mn}{(m+n)^2} \sum_{x \in \boldsymbol{y}} \sum_{x \in \boldsymbol{y}^*} \left[G_n(x) - H_m(x)\right]^2 . \tag{11.9}$$

For comparison of a set of data with a fitted model the Kolmogorov-Smirnov and Cramer-von Mises statistics are typically presented in the context of independent and identically distributed random variables, as was done previously in this subsection. But any set of observed data can be used to construct a marginal empirical distribution function, regardless of whether those data are assumed to have arisen from a model with *iid* random variables, independent but not identically distributed random variables, or even dependent random variables. Conditional on any factors that result in non-identical or non-independent distributions, such as covariates in a regression model, any theoretical model can be used to simulate sets of data that reflect the observed levels of those factors. This then provides a vehicle for comparison of the marginal data distribution with the marginal distribution reflected by a given fitted model.

11.2.3 Quantification of Data Behavior

In many situations we may have interest in a particular aspect of the pattern of observed data, and whether a fitted model provides a good representation of that behavior. For example, if a set of data contains a small number of extremely large values that are separated from the bulk of the observations and with fairly great spacing among themselves, we may have attempted to account for those values by using a distributional form with a long right tail in the model. We might then reasonably question whether that distributional form is sufficient to represent the observed pattern, in the size of samples we actually have. That is, a long right tail can lead to large values, but does it do so with about the correct frequency, and with the type of spacing observed in the actual data. To quantify this data pattern, we might use the difference between the largest and next-to-largest values in the data set, or we might use the average spacing among the three or four largest values.

There is great flexibility in choice of an appropriate quantification of various aspects of data patterns. The goal, of course, is to define a measure or measures that reflect behaviors we believe are important to a given problem. While general prescriptions are elusive, we can list some of the more common issues with which we might be concerned. Appropriate quantifications that reflect the aspects of data patterns listed here are largely problem specific, although they can be motivated by features of the data identified during exploratory analysis.

1. Extreme Values.

   As illustrated in the introductory remarks to this chapter, we are often concerned with data observations that differ from modeled expected values. Even with a model we are generally satisfied with as a description of the overall data pattern we may wish to assess the frequency with which extreme observations occur. Such observations in a data set may reflect unusual circumstances, oddities or errors such that the distributions of those observations differ from that of the remaining set; this is the traditional sense of data values labeled as *outliers*. But extreme observations may also reflect situations that arise somewhat infrequently, but should not be considered unusual or entirely unexpected under a given model. Our intent in assessing patterns of extreme values may be to guide model improvement, may be simply to identify an aspect of the observed situation our (fitted) model is not entirely adequate to describe, or may be to identify cases in the data that deserve closer inspection from a scientific viewpoint.

2. Unusual Data Value Frequencies.

   In some cases a set of data appears to exhibit a high relative frequency of one or two particular values. Perhaps the most common occurrence of this phenomenon is with count data having a large frequency of 0 values. We may well have modeled such a situation through use of a mixture, such as a gamma-Poisson or lognormal-Poisson mixture model. Often the only situations in which one is able to distinguish between these model forms are those that have a high frequency of zero values, because lognormal and gamma distributions can often be matched up to the first two moments, except for $J-$shaped gamma dis-

tributions. A relevant question is then whether both, one, or neither of these models has adequately captured the frequency of zero observations, or whether a more severe model, such as a two-stage model of a binary process combined with a conditional count process, such as a zero-inflated Poisson, is called for.

3. Need for Additional Random Terms.

   It is not always clear when random effects or random data model parameters are beneficial in describing a problem. While the underlying subject matter or science can provide the strongest motivation for such terms in a model, we do not always have such guidance available. There has been, in my opinion, an unfortunate tendency among statisticians to assign the label of *overdispersion* to any situation involving large and perhaps complex patterns of variances, and to respond by including various overdispersion parameters in a model without giving interpretation to such parameters within the context of the problem. Such additional random terms in a model are not infrequently added for any number of rather arbitrarily chosen groups to account for overdispersion. An excellent question in many cases is whether random parameters or effects are truly needed, or whether an alternative approach to modeling variances might be preferred.

4. Mean-Variance Relations.

   Aside from those based on the normal distribution, most random model components imply a particular form of relation between expected values and variances. In some ways this can be thought of as a systematic portion of the random model component. It may well be possible to determine the type of mean-variance relation exhibited by a data set,

and to assess potential models relative to this aspect of the observed data pattern.

5. Marginal versus Conditional Structures.

   As we have seen, many complex models involve conditioning on either data model parameters (e.g., mixed linear models, hierarchical models) or on portions of the entire set of observable random variables (e.g., Markov random field models). When this is the case we have referred to conditional and marginal model structures. In determining an appropriate quantification or quantifications of data pattern, we need to keep in mind whether the pattern or patterns we have interest in are connected with marginal or conditional model structures. A model that is fully adequate to describe a particular problem should, of course, correctly reflect both of these structures. Our ability to assess these parts of overall model structure may, however, be limited by data availability. For example, it is not uncommon in fitting Markov random field models that estimates can be obtained using composite likelihoods (e.g., Besag's original psuedo-likelihood), and that the estimates indicate the presence of substantial dependence among the random field locations. But, if one simulates from the fitted model, it may occur that the simple average over all locations (as an estimate of the marginal mean) is no where near the observed value. One would then need to seriously question whether the data generating mechanism embodied in the fitted statistical model provides an adequate description of the actual scientific mechanisms that led to the observed data.

11.3 Simulation of Reference Distributions

Given the selection of one or more measures of discrepancy and/or quantifications of data pattern, we are prepared to simulate reference distributions for those measures or quantities. We will consider, in turn, the three situations discussed previously.

11.3.1 Discrepancy Between a Data Set and a Fitted Model

If the assessment is to be based on a measure of discrepancy between the actual data and the fitted model, a value of the measure is available for the actual analysis. The chosen measure may be thought of as a function of the estimated parameter $\hat{\theta}$ and the true but unknown parameter θ_0, so let this value of the discrepancy measure be denoted as $D(\hat{\theta}, \theta_0)$. Let the joint distribution implied by the fitted model be denoted $F(\boldsymbol{y}|\hat{\theta})$. Simulation of a reference distribution against which to assess the value $D(\hat{\theta}, \theta_0)$ may be produced by what is essentially a parametric bootstrap in the following manner:

1. For $k = 1, \ldots, M$, simulate data sets $\boldsymbol{y}^{(k)}$ from $F(\boldsymbol{y}|\hat{\theta})$.

2. For each simulated data set estimate $\hat{\theta}$ as $\theta^{(k)}$ and compute the chosen discrepancy measure as $D(\theta^{(k)}, \hat{\theta})$.

3. The empirical distribution function of the M values $\{D(\theta^{(k)}, \hat{\theta}) : k = 1, \ldots, M\}$ forms a reference distribution against which to assess the actual value $D(\hat{\theta}, \theta_0)$. In particular, a simulation-based $p-$value can be

computed as,

$$p = \frac{1}{M} \sum_{k=1}^{M} I\{D(\hat{\theta}, \theta_0) \geq D(\theta^{(k)}, \hat{\theta})\}, \tag{11.10}$$

where I is the indicator function that assumes a value of 1 if its argument is true and a value of 0 otherwise.

11.3.2 Discrepancy Between Two Data Sets

If the assessment is to proceed based on a measure of discrepancy between two data sets, a test quantity or test statistic is not available from only the actual data set and the model estimated from it. In this case, we need to obtain through simulation both the test quantity and its reference distribution. Let the actual data set be denoted as $\boldsymbol{y}^a$, the distribution implied by the fitted model be denoted as before by $F(\boldsymbol{y}|\hat{\theta})$, and the chosen measure of discrepancy of $\boldsymbol{y}^a$ with any other set of data be denoted as $D(\boldsymbol{y}^a, \boldsymbol{y}^{(m)})$. We assume here that $D(\boldsymbol{y}^a, \boldsymbol{y}^{(m)})$ is a summary measure that compares data sets in total, such as Kolmogorov-Smirnov or Cramer-Von Mises statistics discussed previously, and we assume that this measure can assume only non-negative values. A procedure to accomplish simulation-based assessment is as follows:

1. For $m = 1, \ldots, M$, simulate data sets $\boldsymbol{y}^{(m)}$ from $F(\boldsymbol{y}|\hat{\theta})$.

2. For each simulated data set, compute the discrepancy between it and the actual data resulting in the set of measures $\{D(\boldsymbol{y}^a, \boldsymbol{y}^{(m)}) : k = 1, \ldots, M\}$.

3. Compute the average of these discrepancy measures as a reflection of the difference between the actual data and the fitted model as

$$T = (1/M) \sum_{m=1}^{M} D(\boldsymbol{y}^a, \boldsymbol{y}^{(m)}).$$

The statistic T will play the role of a test statistic for a hypothesis that the model with estimated parameter $\hat{\theta}$ provides an adequate fit to the data.

4. For each simulated data set $\boldsymbol{y}(m)$, repeat this entire process as if it were the actual data. That is, for $m = 1, \ldots, M$,

   4.1 Estimate the parameter as $\hat{\theta}^{(m)}$ from the simulated data $\boldsymbol{y}^{(m)}$.

   4.2 For $j = 1, \ldots, M$, simulate data sets $\boldsymbol{y}^{(m,j)}$ from $F(\boldsymbol{y}|\hat{\theta}^{(k)})$.

   4.3 For each second-level simulated data set $\boldsymbol{y}^{(m,j)}$, compute the discrepancy between it and the original simulated data resulting in the set of measures $\{D(\boldsymbol{y}^{(m)}, \boldsymbol{y}^{(m,j)}) : j = 1, \ldots, M\}$.

   4.4 Compute the average of these discrepancy measures as

$$T^{(m)} = (1/M) \sum_{j=1}^{M} D(\boldsymbol{y}^{(m)}, \boldsymbol{y}^{(m,j)}).$$

   The statistic $T^{(m)}$ plays the role of a test statistic for a hypothesis that the model with estimated parameter $\hat{\theta}^{(m,j)}$ provides an adequate fit to the data $\boldsymbol{y}^{(m)}$.

5. The result of step 4 is a set of values $\{T^{(m)} : m = 1, \ldots, M\}$. The empirical distribution of these M values represents a reference distribution against which to compare the actual test statistic T from step 3 of the procedure. If desired, this comparison may be represented in the form of a $p-$value as

$$p = \frac{1}{M} \sum_{m=1}^{M} I(T \leq T^{(m)}), \tag{11.11}$$

   where $I(\cdot)$ is the usual indicator function.

There are possible modifications of this procedure that may be useful in particular situations, most of which have an objective of reducing the computational burden of this procedure. Whatever modifications are contemplated, however, it is important is that the original test statistic T and the simulated test statistics $T^{(m)}$ be produced in precisely the same manner. That is, whatever is done to the actual data should also be done to each of the simulated data sets from step 1.

11.3.3 Quantification of Data Patterns

Suppose that model assessment is to be based on one or more given quantifications of the behavior of data that might be generated from the fitted model. Let $Q(\boldsymbol{y}^a)$ denote the value of such a quantity for the actual data set. Really, the only distinction between this situation and that for comparison of two data sets is that we assume $Q(\boldsymbol{y})$ can be computed from a single data set rather than requiring a pair of data sets. This does, however, greatly simplify the procedure needed to produce a reference distribution from that required for discrepancy measures between data sets, as it eliminates the need for estimation using each data set simulated from the actual fitted model. A simulation-based procedure for arriving at a reference distribution for $Q(\boldsymbol{y}^a)$ can be outlined as follows.

1. For $m = 1, \ldots, M$, simulate data sets $\boldsymbol{y}^{(m)}$ from $F(\boldsymbol{y}|\hat{\theta})$.

2. For each simulated data set, compute the quantity $Q(\boldsymbol{y}^{(m)})$, resulting in the set of quantities $\{Q(\boldsymbol{y}^{(m)} : k = 1, \ldots, M\}$.

3. A simulation-based $p-$value for $Q(\boldsymbol{y}^a)$ may then be computed as

$$\begin{aligned} p_\ell &= \frac{1}{M}\sum_{m=1}^{M} I\left(Q(\boldsymbol{y}^{(m)}) \leq Q(\boldsymbol{y}^a)\right) \\ p_u &= \frac{1}{M}\sum_{m=1}^{M} I\left(Q(\boldsymbol{y}^{(m)}) \geq Q(\boldsymbol{y}^a)\right) \\ p &= \min\{p_\ell, p_u\}. \end{aligned} \tag{11.12}$$

   where again $I(\cdot)$ is the usual indicator function.

In (11.12) there are both lower and upper sided $p-$values and both include equality with the actual test statistic $Q(\boldsymbol{y}^a)$. The reason for this is that some important data characteristics can be highly discrete and have a relatively small number of possible values. In these situations, a fitted model that generates a large number of data sets with test quantities $Q(\boldsymbol{y}^{(m)})$ equal to $Q(\boldsymbol{y}^a)$ should not be judged as discrepant with the actual data. At the same time, some data characteristics lead to test quantities that indicate discrepancy with the actual data if they are either much smaller or much larger than $Q(\boldsymbol{y}^a)$.

11.4 Case Study: Sales of Green Beans

A grocery chain in the Midwestern United States was interested in factors that affect sales of basic food items, including price. Here, we will consider the relation between price and number of units sold in one grocery store for green beans. The data did not indicate what a unit was or how price was recorded, but from the values in the data it is not unreasonable to suppose that units were cans and price was in dollars.

Our objective is to formulate a regression model to relate number of units sold, as response variables, to price, as covariates. A scatterplot of the data from the store under consideration is shown in Figure 11.1.

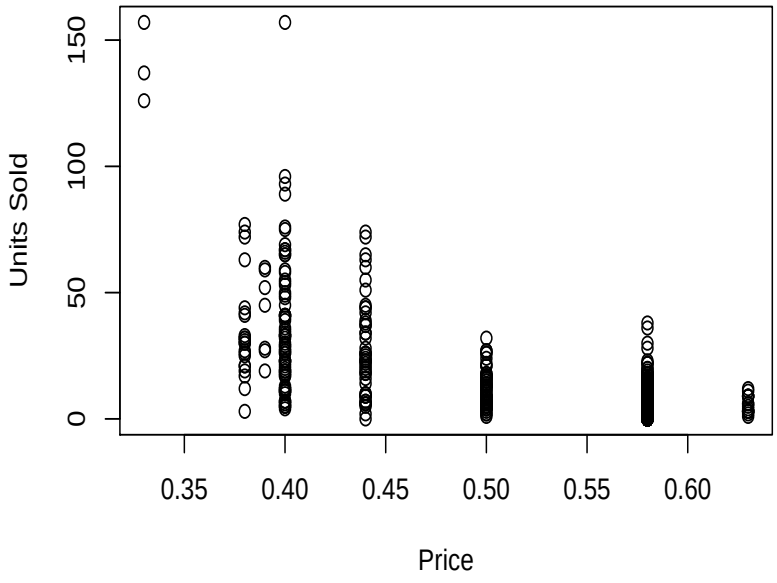

Figure 11.1: Sales of green beans in a grocery store as a function of price.

Without details being presented, two models were arrived at as possible representations of the data, a regression with zero-inflated Possion response distributions (ZIP) and a regression with random component given by gamma-Poisson mixutre distributions (GP). Both models used a systematic model component given as $\mu_i = 1/x_i^2$. In the model with GP random component, the gamma parameter α was held constant across levels of the covariate. The selection of these models made use of typical residuals plots

and model diagnostics. We determined four characteristics of the data that were of interest in modeling; (1) the frequency of 0 values, (2) the mean-variance relation, (3) the magnitude of variances at given covariate values, and (4) the frequency of extreme values.

11.4.1 Formulating Test Quantities

We denote response random variables as $\{Y_i : i = 1, \ldots, n\}$ and covariate values as $\{x_i : i = 1, \ldots, n\}$. Note that there are a fixed number of distinct covariate values and $x_i \in \{w_k : k = 1. \ldots, K\}$ for each i. For each of the two models, we simulate M sets of response values $\boldsymbol{y}_m = \{y_i^* : i = 1, \ldots, n\}$ using the same covariate values as in the actual data. To make use of our four criteria in a model assessment procedure then, we conducted the following for each of the two models, ZIP and GP.

1. Zero Frequencies

   In the actual data, let the number of zero response values at level w_k of the covariate be denoted as z_k^a; $k = 1, \ldots, K$. The expected frequencies of $Y_i = 0$ at each covariate value were computed from the model and evaluated with the estimated parameter values. Let $\{e_k^a : k = 1, \ldots, K\}$ denote these values. The assessment quantity or test statistic for frequency of zero reponses was,

   $$T_1^a = \sum_{k=1}^{K} I(z_k^a - e_k^a > 0), \tag{11.13}$$

   where $I(cdot)$ is the indicator function.

   For each simulated data set, the fitted model was first estimated from those values. Then the number of zero responses and the expected number under the fitted model were computed at each level of the

covariate as $z_{m,k}$, and $\{e_{m,k}$, both for $k = 1, \ldots, K\}$. The test quantity was computed for each simulated data set in the same way as for the actual data to give $\{T_1^{(m)} : m = 1, \ldots, M\}$. Because the test quantities are quite discrete, having possible values $\{0, 1, \ldots, K\}$ both lower and upper $p-$values were computed as in (11.12).

2. Magnitude of Variances.

   To assess the magnitude of variances a procedure similar to that used to assess zero frequencies was employed, where the sample variance at each covariate level replaced the number of 0 responses, and model variances replaced the expected number of 0 responses. Otherwise, the procedure was as described for assessing the frequency of 0 responses, giving test quantities for the actual data T_2^a and each simulated data set $\{T_2^{(m)} : m = 1, \ldots, M\}$. Lower and Upper $p-$values were again computed as in (11.12).

3. Mean-Variance Relation.

   To assess how well the modela reflect the relation between means and variances, we made use of the Box-Cox procedure for both actual and simulated data sets. Compute the sample means and variances at each level of the covariate in the actual data that contains at least 6 observations as $\{(b_k, s_k^2) : k = 1, \ldots, K\}$, regress $0.5 \log(s_k^2)$ on $\log(b_k)$ and compute the slope T_3^a using ordinary least squares and a straight line regression model. Repeat this process for each of the simulated data sets to arrive at values $T_3^{(m)}; m = 1, \ldots, M$. Here, the test quantities are not highly discrete and we need compute only the upper $p-$value as in (11.11). The lower $p-$ value will be 1 minus the upper value.

4. Large Values. As the scatterplot of Figure 11.1 shows, there are some extreme large responses. Grocers are naturally interested in these particular data points and, even if our current modeling exercise cannot explain them, the frequency with which they might occur. To assess the abilities of our models to account for large values we used the following procedure. First, raw residuals were computed for the actual data as $r_i = y_i - \hat{\mu}_i$, where $\hat{\mu}_i$ are the estimated expected values from the regression of sales on price. Let $R_{neg} = \max\{|r_i| : r_i < 0\}$. The test quantity was then

$$T_4^a = \sum_{i=1}^{n} I(r_i \geq R_{neg}),$$

where $I(\cdot)$ is the usual indicator function. T_4 gives the number of positive residuals that are at greater than or equal to the largest magnitude negative residual. The model is fit to each simulated data set, residuals are computed and the test quantity is computed for each set, giving $\{T_4^{(m)} : m = 1, \ldots, M\}$. For this criteria there are a fairly small number of positive residuals (large observations) that are extreme, so both lower and upper $p-$values were again computed.

11.4.2 Results

Table 11.1 gives the $p-$values that resulted from the assessment criteria using $M = 5,000$ simulated data sets.

Test quantity T_1 is large if the number of 0 responses in the data are greater than expected under the (fitted) model. Small (large) values of the lower $p-$value indicate that the model over-represents (under-represents) the frequency of 0 values relative to the actual data. Small (large) values of the

	ZIP		GP	
Criterion	Lower	Upper	Lower	Upper
Zero Frequency (T_1)	0.01232	1	0.7339	0.6243
Variance Magnitude (T_2)	0.9994	0.0058	0.2238	0.9604
Mean-Variance Relation (T_3)	0.8606	—	0.3061	—
Large Values (T_4)	1	0	0.7750	0.2618

Table 11.1: Simulation based $p-$values for four assessment criteria applied to green bean regression models.

upper $p-$value indicates that the model under-represents (over-represents) the frequency of 0 response values. Based on the values of Table 11.1 it appears that the ZIP model produces a greater frequency of 0 values than supported by the data. The GP model, on the other hand, appears to be an adequate representation of the data based on this criterion.

Test quantity T_2 is large if the theoretical variance under the (fitted) model is greater than exhibited by the data. Small (large) values of the lower $p-$value indicate that the model tends to produce fewer (more) instances (groups of values at given covariate levels) in which observed variances are greater than theoretical than observed in the actual data. That is, the model over-represents (under-represents) variances relative to the actual data. Small (large) upper $p-$values indicate that the model produces data in which a larger (smaller) number of instances have variances greater than the theoretical model value. That is, the model under-represents (over-represents) variances relative to the actual data. The $p-$values for the ZIP model indicate that this model reflects less variance within covariate groups than seen in the actual data. The GP model appears superior in this regard, although

based on the upper $p-$value one may wonder if it has a slight tendency to produce data with variances that are too small relative to the actual data. Test quantity T_3 needs to be used to produce only lower or only upper $p-$values. This quantity is larger when variances increase more rapidly with means and smaller when variances increase less rapidly in relation to means. In Table 11.1 the lower $p-$value is given. Small values of the lower $p-$value indicate that the model produces data for which variances increase more rapidly with the mean than do the actual data, while large $p-$values indicate that the model produces data for which variances increase less rapidly with the mean than do the actual data. The $p-$values of Table 11.1 show that both ZIP and GP models represent the mean-variance relation exhibited by the actual data in a reasonable manner.

Test quantity T_4 is larger when a greater number of positive residuals qualify as extreme and smaller when a lesser number of positive residuals qualify as extreme. Small (large) lower $p-$values indicate that the model generates data for which fitting the model results in more (less) extreme positive residuals than resulted from fitting the model to the actual data. Small (large) upper $p-$values indicate that the model generates data for which fitting the model results in fewer (more) extreme positive responses. The values of Table 11.1 show that the ZIP model under-represents extreme large responses relative to the data, while the GP model appears to be more in concert with the data in terms of this criterion.

Overall, both models represent the mean-variance relation exhibited by the data in a reasonable manner. The ZIP model produces 0 responses too great a frequency, under-represents variances within levels of the covariate, and also under-represents extreme large values. The GP mixture model does not suffer these same deficiencies, although it may over-represent variances within

covariate levels to some degree. The fitted gamma-Poisson mixture regression model is shown in Figure 11.2 along with 95% confidence bands. Estimation was simultaneous maximum likelihood and confidence bands computed using asymptotic normality and the delta method.

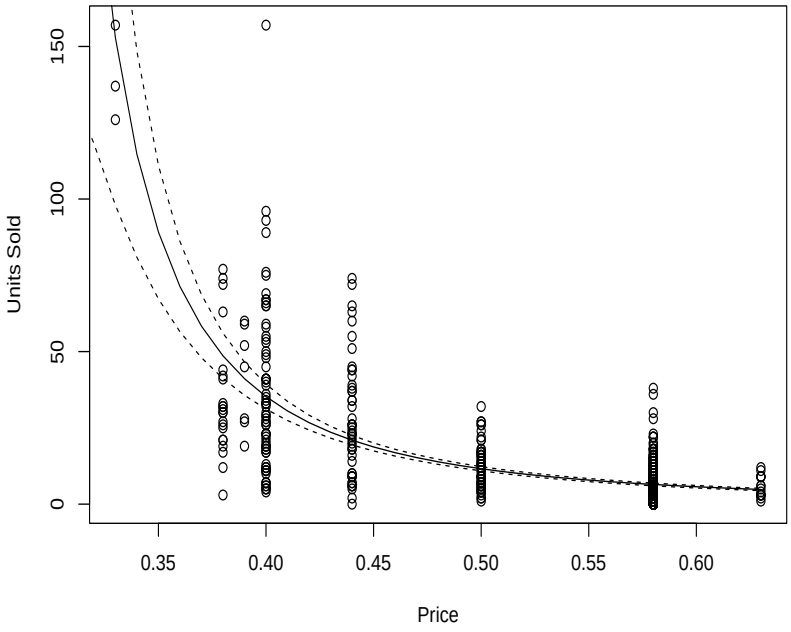

Figure 11.2: Estimated expectation function and confidence bands for gamma-Poisson mixture regression model fit to green bean data.

Chapter 12

Prediction and Prediction Error

12.1 Roles and Implications of Prediction in Statistical Analysis

Prediction enters a statistical analysis in two distinct ways. First, prediction may be the primary objective of an analysis, where we distinguish prediction from forecasting. If prediction is the primary objective of an analysis we may not be greatly concerned with including model parameters that have no clear plausible interpretation within the scientific context of the problem under investigation. Choosing to include a poorly understood trend in large-scale (or mean) structure in spatial or temporal models, as opposed to allowing the small-scale (or dependence) structure to account for trend is a classic example. If our primary objective were to conceptualize the scientific phenomenon in a statistical model we might be reluctant to include trend in the large-scale portion of the model structure unless there is a scientifi-

cally plausible explanation for that trend. On the other hand, if our primary objective were prediction of unobserved locations or times internal to the extent of the available data then we would likely be well advised to incorporate trend into the mean function, even if we don't understand what might be causing it. Even in situations for which prediction is not a primary objective of analysis, however, a fundamental notion is that an adequate model should be able to predict well. The distinction here is one of primacy or importance. The best prediction ability in a given problem might be shown by a model that includes terms without scientific motivation, and I might choose not to consider that model if my primary objective is conceptualization of the scientific problem. But, given two competing models with similar levels of scientific support we might choose between the two based at least in part on their relative predictive performances.

12.2 A General Notational Framework

In portions of this chapter we will rely on a general notational framework for models that is built around the concept of a random field. Let $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ denote a set of random variables connected with observable quantities, with $\boldsymbol{s}_i$ a non-random "location variable". Several possibilities for the location variables $\boldsymbol{s}_i$ are:

1. Independent random variables.

   Here, we would natrually take $\boldsymbol{s}_i = i$ and simplify notation by referring to $Y(\boldsymbol{s}_i)$ as just Y_i.

2. Groups of random variables.

   Here, we might define $\boldsymbol{s}_i = (k, j)$ where k indexes group and j indexes

observation within group; $k = 1, \ldots, K$ and $j = 1, \ldots, n_k$.

3. Geographic random variables.

   Here, we might take $\boldsymbol{s}_i = (u_i, v_i)$, where u_i denotes latitude and v_i longitude, or u_i denotes horizontal coordinate on a grid and v_i denotes vertical coordinate on a grid.

4. Time series of random variables.

   Here we might take $\boldsymbol{s}_i = t$ where t is time, if each $Y(\boldsymbol{s}_i)$ occurs at a unique time, or $\boldsymbol{s}_i = (t, j)$ where t is time and j is observation number at time t; $t = 1, \ldots, T$ and $j = 1, \ldots, n_t$.

We assume that each $Y(\boldsymbol{s}_i)$ is modeled through a parametric distribution having a density (or mass function) f_i, depending on parameter $\psi(\boldsymbol{s}_i)$ through the data model,

$$f_i(y(\boldsymbol{s}_i)|\psi(\boldsymbol{s}_i)); \quad y(\boldsymbol{s}_i) \in \Omega_i. \tag{12.1}$$

Here, the densities f_i are indexed by i to allow for the possibility of differing covariates $\boldsymbol{x}_i$ or auxilliary information (e.g., binomial sample size).

We will assume that the parameters $\{\psi(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ represent "minimal" parameters in the sense that any other parameters used in writing the densities f_i; $i = 1, \ldots, n$ are functions of the $\psi(\boldsymbol{s}_i)$, and also that we may write,

$$\psi(\boldsymbol{s}_i) = (\psi_f(\boldsymbol{s}_i), \psi_r(\boldsymbol{s}_i)), \tag{12.2}$$

where $\psi_f(\boldsymbol{s}_i)$ represents parameters that are fixed in the data model and $\psi_r(\boldsymbol{s}_i)$ denotes parameters that are random in the data model. We take $\psi_r(\boldsymbol{s}_i)$ to have a distribution with parameterized density $g_i(\psi_r(\boldsymbol{s}_i)|\boldsymbol{\lambda})$, where this density may result from marginalization over any additional levels of random terms in the model. For examle, if $\psi_r(\boldsymbol{s}_i)$ is modeled directly in

terms of a distribution $g_{1,i}(\psi_r(\boldsymbol{s}_i)|\lambda_1(\boldsymbol{s}_i))$ with $\lambda_1(\boldsymbol{s}_i)$ having a distribution with density $g_2(\lambda_1(\boldsymbol{s}_i)|\boldsymbol{\lambda})$, then,

$$g_i(\psi_r(\boldsymbol{s}_i)|\boldsymbol{\lambda}) = \int g_{1,i}(\psi_r(\boldsymbol{s}_i)|\lambda_1(\boldsymbol{s}_i))g_2(\lambda_1(\boldsymbol{s}_i)|\boldsymbol{\lambda})\, d\lambda_1(\boldsymbol{s}_i). \tag{12.3}$$

Finally, we then take the marginal density of $Y(\boldsymbol{s}_i)$ to be given by

$$h_i(y(\boldsymbol{s}_i)|\psi_f(\boldsymbol{s}_i),\, \boldsymbol{\lambda}) = \int f_i(y(\boldsymbol{s}_i)|\psi_f(\boldsymbol{s}_i),\, \psi_r(\boldsymbol{s}_i))\; g(\psi_r(\boldsymbol{s}_i)|\boldsymbol{\lambda})\; d\psi_r(\boldsymbol{s}_i). \tag{12.4}$$

Example 12.1

Consider a typical linear regression model with independent response variables, written as

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma\epsilon_i; \quad i = 1, \ldots, n.$$

This model fits into our general notation by defining $\boldsymbol{s}_i \equiv i$ and $\psi_f(\boldsymbol{s}_i) \equiv (\beta, \sigma^2)$ and dropping remaining elements of the structure; there is no $\psi_r(\boldsymbol{s}_i)$ or densities g_i, $g_{1,i}$ or $g_{2,i}$.

Example 12.2

We have written a standard generalized linear model with responses independent and,

$$\begin{aligned} f(y_i|\theta_i, \phi) &= \exp\left[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\right], \\ \mu_i &= b'(\theta_i) \\ \eta_i &= \boldsymbol{x}_i^T\,\boldsymbol{\beta} \\ g(\mu_i) &= \eta_i \end{aligned}$$

which fits into our general notation with $\boldsymbol{s}_i \equiv i$ and $\psi_f(\boldsymbol{s}_i) \equiv (\boldsymbol{\beta},\, \phi)$. Note here that all intermediate parameters can be written in terms of these fixed

values as

$$\eta_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^T \boldsymbol{\beta}; \quad \mu_i(\boldsymbol{\beta}) = g^{-1}(\eta_i(\boldsymbol{\beta})); \quad \theta_i(\boldsymbol{\beta}) = b'^{-1}(\mu_i(\boldsymbol{\beta})).$$

Example 12.3

A beta-binomial mixture model for a set of independent random variables may be written as,

$$\begin{aligned} f_i(y_i|\theta_i) &= \frac{n_i!}{y_i!\,(n_i - y_i)!}\,\theta_i^{y_i}\,(1-\theta_i)^{n_i - y_i}, \\ g(\theta_i|\alpha,\,\beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}, \\ h_i(y_i|\alpha,\beta) &= \frac{n_i!}{y_i!\,(n_i - y_i)!}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}. \end{aligned}$$

This fits into our general notation with $\boldsymbol{s}_i \equiv i$, $\psi_r(\boldsymbol{s}_i) = \theta_i$ and $\boldsymbol{\lambda} \equiv (\alpha, \beta)$.

Example 12.4

A model for data with random cluster or group effects is,

$$Y_{g,i} = \boldsymbol{x}_{g,i}^T \boldsymbol{\beta} + \sum_{g=1}^{G} \delta_g I(i \in \mathcal{C}_g) + \sigma\epsilon_{g,i}; \quad \delta_g \sim iidN(0, \tau^2); \quad \epsilon_{g,i} \sim iidN(0,1),$$

where g indexes group for $g = 1, \ldots, G$, and i indexes observation within group for $i = 1, \ldots, n_g$. To put this into our general notation, define $\boldsymbol{s}_i \equiv (g, i)$, $\psi_f(\boldsymbol{s}_i) \equiv (\boldsymbol{\beta}, \sigma^2)$, $\psi_r(\boldsymbol{s}_i) \equiv \delta_g$, and $\boldsymbol{\lambda} \equiv \tau^2$.

This general framework for notation is sufficient to cover most of the models we have discussed, as illustrated in the preceding examples. We will present general results in terms of this random field notation, but will

revert back to simpler and more typical indexing in examples and particular sections of the chapter where the generality of the random field location $\boldsymbol{s}_i$ is not needed.

12.3 Decision Theory and Prediction

The most common approach to considering prediction problems is to use the framework of decision theory. To develop this approach more formally, let $\boldsymbol{Y} \equiv \{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ be a set of random variables that correspond to observable data at random field locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. Let $\boldsymbol{\theta}_r$ be composed of the unique elements in the set of random data model parameters $\{\psi_r(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ and let $\boldsymbol{\theta}_f$ be composed of the unique elements in the union of the set of fixed data model parameters and fixed parameters of the mixing distribution of random data model parameters (if any), namely $\{\psi_f(\boldsymbol{s}_i) : i = 1, \ldots, n\} \cup \boldsymbol{\lambda}$.

Our goal is to predict the value of a random variable we will denote Z. This random variable may be any of the elements of $\boldsymbol{\theta}_r$ at either observed or unobserved locations, or may be an unobserved response variable at an unobserved location, namely $Y(\boldsymbol{s}_0)$. We assume that Z follows the same process that generated $\boldsymbol{Y}$. Assume that a model has been specified for these random variables that results in the joint distribution function for $\boldsymbol{y}$ and Z, $F(\boldsymbol{y}, z|\boldsymbol{\theta}_f)$. We may or may not need to derive this distribution in explicit form in order to make use of the following development.

Let $p_z(\boldsymbol{Y})$ denote any potential predictor of Z that is a function of the observed $\boldsymbol{Y}$; note that $p_z(\boldsymbol{Y})$ may depend on $\boldsymbol{Y}$ through the estimated parameter $\hat{\boldsymbol{\theta}}_f$ (which will be a function of $\boldsymbol{Y}$), through the observed values $\boldsymbol{y}$, or through both. In the decision theoretic framework we define a loss func-

tion for any possible predictor as $L\{p_z(\boldsymbol{Y}), Z\}$ and the expected loss or risk as the expectation $R\{p_z(\boldsymbol{Y}), Z\} = E[L\{p_z(\boldsymbol{Y}), Z\}]$, where the expectation is taken with respect to the joint distribution of $\boldsymbol{Y}$ and Z. Choosing an optimal predictor under a given loss function corresponds to minimizing the associated risk. By far the most common choice of loss function is squared error loss, $L\{p_z(\boldsymbol{Y}), Z\} = \{p_z(\boldsymbol{y}) - Z\}^2$. Under squared error loss, risk is also often called the mean squared prediction error,

$$mspe(p_z) \equiv E\left[\{p_z(\boldsymbol{Y}) - Z\}^2\right]. \tag{12.5}$$

The usual method for minimizing (12.5) is to use a conditioning argument as follows. We have that

$$mspe(p_z) = E\left(E\left[\{p_z(\boldsymbol{Y}) - Z\}^2 | \boldsymbol{Y}\right]\right). \tag{12.6}$$

The inner (conditional) expectation in (12.6) is minimized for each $\boldsymbol{Y}$ by taking $p_z(\boldsymbol{Y}) = E\{Z|\boldsymbol{Y}\}$ which must then also lead to the minimum over the outer (marginal) expectation. Thus, the minimum $mspe$ predictor of Z is

$$p_z(\boldsymbol{Y}) = E\{Z|\boldsymbol{Y}\}, \tag{12.7}$$

which has $mspe$,

$$\begin{aligned} mspe(p_z) &= E\left\{E\left(\left[E\{Z|\boldsymbol{Y}\} - Z\right]^2 |\boldsymbol{Y}\right)\right\} \\ &= E\left[\text{var}\left\{Z|\boldsymbol{Y}\right\}\right] \\ &= var\{Z\} - \text{varvar}\left[E\left\{Z|\boldsymbol{Y}\right\}\right] \\ &= var\{Z\} - \text{var}\left\{p_z(\boldsymbol{Y})\right\}. \end{aligned} \tag{12.8}$$

At first glance, expression (12.8) might seem counter-intuitive in that it indicates the total error of prediction $mspe(p_z)$ decreases as the variance

of the predictor $p_z(\boldsymbol{Y})$ increases. To understand why this is not counter-intuitive, consider that the predictor of (12.7) is the conditional expectation $E\{Z|\boldsymbol{Y}\}$. If this conditional expectation has variance 0, then the values of $\boldsymbol{Y}$ have no influence on the predictor, that is, the predictor of Z is constant over all possible values for $\boldsymbol{Y}$, and in this case the predictor will be just the marginal expectation of Z. If, however, the possible values of $\boldsymbol{Y}$ are allowed to influence the predictor $p_z(\boldsymbol{Y})$ then the variance of this predictor will increase. Thus, expression (12.8) indicates the degree to which the conditional expectation $E\{Z|\boldsymbol{Y}\}$ differd from the marginal expectation $E\{Z\}$ as the values of the conditioning variables $\boldsymbol{Y}$ change.

Expression (12.8) gives the *marginal* mean squared prediction error for $p_z(\boldsymbol{Y})$. Another development would focus on minimization of the the *conditional* mean squared prediction error. Now, we know that minimization of the conditional *mspe* will again result in the predictor $p_z(\boldsymbol{Y}) = E\{Z|\boldsymbol{Y}\}$, a fact that we actually have already used in moving from (12.6 to (12.7). It will be useful in what follows, however, to determine an expression for the conditional mean squared prediction error for any predictor $p_z(\boldsymbol{Y})$. Specifically, for any predictor $p_z(\boldsymbol{Y})$, the conditional mean squared prediction error is,

$$\begin{aligned} mspe(p_z|\boldsymbol{Y}) &= E\left[\{p_z(\boldsymbol{Y}) - Z\}^2 |\boldsymbol{Y}\right] \\ &= E\left([p_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\} + E\{Z|\boldsymbol{Y}\} - Z]^2 |\boldsymbol{Y}\right) \\ &= E\left([p_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}]^2 |\boldsymbol{Y}\right) + \\ &\quad 2E\left([p_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}]\ [E\{Z|\boldsymbol{Y}\} - Z]\ |\boldsymbol{Y}\right) + \\ &\quad \text{var}\{Z|\boldsymbol{Y}\}. \end{aligned} \tag{12.9}$$

Now, of the three terms that constitute the last equality in expression (12.9), the first is constant given $\boldsymbol{Y}$, and the second is easily shown to be equal to

zero (everything is again constant given $\boldsymbol{Y}$ except for the trailing Z which gives for the second factor $E\{Z|\boldsymbol{Y}\} - E\{Z|\boldsymbol{Y}\} = 0$). The result is that,

$$mspe(p_z|\boldsymbol{Y}) = [p_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}]^2 + var\{Z|\boldsymbol{Y}\}. \tag{12.10}$$

Because both terms in (12.10) are non-negative, the minimum is achieved by taking $p_z(\boldsymbol{Y}) = E\{Z|\boldsymbol{Y}\}$ which agrees with what we already know. Of more import is the fact that the conditional mean squared prediction error is then

$$mspe\{p_z|\boldsymbol{Y}\} = \text{var}\{Z|\boldsymbol{Y}\}, \tag{12.11}$$

and the unconditional or marginal mean squared prediction error is again the expected value of this conditional variance as in expression (12.8).

12.4 Untangling Prediction Errors

There are two issues involved in the interpretation of mean squared prediction errors that deserve consideration. The first involves the question of whether the marginal *mspe* of (12.8) or the conditional *mspe* of (12.11) is the more appropriate measure of prediction error in a given application. The second involves the fact that the use of plug-in estimators in either of these expressions leads to an under-estimation of the actual errors.

12.4.1 Conditional Versus Marginal Errors

In a given application the question arises as to whether the marginal mean squared prediction error (12.8) or the conditional mean squared prediction error (12.11) is the more appropriate quantification of uncertainty in predicted values. The decision rests on whether one wishes to quantify uncertainty in predicted values of Z given a particular set of observed values $\boldsymbol{y}$, or under

the procedure that consists of observing $\boldsymbol{y}$ and then predicting on the basis of these values. To compute the conditional version requires that observed values $\boldsymbol{y}$ are available. That is, the conditional mean squared prediction error of (12.11) can only be given a numerical value as,

$$mspe\{p_z|\boldsymbol{Y} = \boldsymbol{y}\} = \text{var}\{Z|\boldsymbol{Y} = \boldsymbol{y}\},$$

which may well be a function of both the parameter $\boldsymbol{\theta}_f$ and the observed values $\boldsymbol{y}$. The marginal mean squared prediction error of expression (12.8), on the other hand, typically will depend only on the parameter $\boldsymbol{\theta}_f$.

Example 12.5

Consider a problem of predicting the concentration of iron ore at an unobserved spatial location $\boldsymbol{s}_0$ in some given spatial domain. Observations are available at a specific set of locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ as the particular values $\boldsymbol{y} = \{y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$, which are assumed to be measured without error and to remain constant over the time frame of interest (e.g., 100 years). In this case the predictand is $Z = Y(\boldsymbol{s}_0)$; $\boldsymbol{s}_0 \notin \{\boldsymbol{s}_i : i = 1, \ldots, n\}$ and the most reasonble quantification of uncertainty in the predicted value $p_z(\boldsymbol{Y})$ would seem to be the conditional prediction mean squared error of (12.11), using the specific values $\boldsymbol{y}$.

Example 12.6

Now consider a problem of predicting ozone concentration in and around Chicago at an unobserved spatial location $\boldsymbol{s}_0$. On a given day, observations are available at n locations as $\boldsymbol{y}$ in the same way as for the problem of iron ore prediction, and the predictand is again $Z = Y(\boldsymbol{s}_0)$. But the observed values will change from day to day, which is finer division of time than any period in which we are likely to have interest. In this case one might well

defend the use of the marginal mean squared prediction error of expression (12.8) as the more appropriate quantification of uncertainty in predictions.

Many statisticians have a decided preference for the marginal mean squared prediction error as a quantification of uncertainty in predicted values. This might well have to do with the fact that a good deal of statistical activity involves the comparison of alternative models or approaches to various problems. It may then be rather natural for statisticians to focus on the expected mean squared prediction error rather than the realized mean squared prediction error in a particular situation. Nevertheless, it is advisable to keep in mind that in any given situation the conditional mean squared prediction error may be a more appropriate measure of uncertainty in predicted values than the marginal or expected value. A related issue involves the fact that the marginal mean squared prediction error is an expectation taken over the joint distribution of observed random variables $\boldsymbol{Y}$ and the unobserved variable Z. In most situations we will want predictions at a set of random field locations, not only one such location, that is, Z will actually be a vector $\boldsymbol{Z}$. In some special situations it is possible to use the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ to derive a set of predictors. This will be true, for example, in prediction of random effects in mixed linear models. In other cases, however, we typically derive predictors for a single random variable Z despite the fact that such a predictor will be applied repeatedly to a set of random variables. This is typically the case, for example, with Markov random field models in which we derive the necessary conditional expectations without attempting to explicity derive the distribution $F(\boldsymbol{y}, z|\boldsymbol{\theta}_f)$.

12.4.2 Total Error of Prediction

The expressions (12.8) for marginal and (12.11) for conditional mean squared prediction error only hold if $p_z(\boldsymbol{Y}) = E\{Z|\boldsymbol{Y}\}$. In most, if not all, applications we will use an estimate of this expectation so that what we have in reality is $\hat{p}_z(\boldsymbol{Y}) = \hat{E}\{Z|\boldsymbol{Y}\}$. If this estimator is used as a plug-in for $E\{Z|\boldsymbol{Y}\}$ in expressions for mean squared prediction error the result will under-estimate the actual uncertainty of prediction. Consider the conditional mean squared prediction error of expression (12.11), which results from (12.10) because $p_z(\boldsymbol{Y})$ was taken to be equal to $E\{Z|\boldsymbol{Y}\}$. With an estimate of this conditional expectation, (12.10) results not in (12.11) but rather,

$$\begin{aligned} mspe(\hat{p}_z|\boldsymbol{Y}) &= [\hat{p}_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}]^2 + \text{var}\{Z|\boldsymbol{Y}\} \\ &= \left[\hat{E}\{Z|\boldsymbol{Y}\} - E\{Z|\boldsymbol{Y}\}\right]^2 + \text{var}\{Z|\boldsymbol{Y}\}, \end{aligned} \tag{12.12}$$

and the marginal mean squared prediciton error becomes,

$$mspe(\hat{p}_z) = E\left(\left[\hat{E}\{Z|\boldsymbol{Y}\} - E\{Z|\boldsymbol{Y}\}\right]^2\right) + E\left[\text{var}\{Z|\boldsymbol{Y}\}\right]. \tag{12.13}$$

As expression (12.13) demonstrates, the total prediction error actually consists of the mean squared prediction error using the optimal predictor $E\{Z|\boldsymbol{Y}\}$ (which is the same as (12.8) that we had previously) plus a mean squared estimation error. Also note that there is nothing in the derivation of (12.13) that made use of the optimal predictor so that we could apply this expression to any predictor $\tilde{p}_z(\boldsymbol{Y})$ to arrive at,

$$mspe(\tilde{p}_z) = E\left([\tilde{p}_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}]^2\right) + E\left[\text{var}\{Z|\boldsymbol{Y}\}\right], \tag{12.14}$$

which shows that, under squared error loss, the total error of potential predictors are determined by their mean squared error in estimation of $E(Z|\boldsymbol{Y})$.

We can derive another form for the total mean squared error (12.14) of a predictor that results in an interpretation analogous to the usual result that mean squared errors are equal to variance plus squared bias. This derivation is simplified through the use of two general results, the proofs of which are left as an exercise.

Result 1.

For any two random variables X and Y,

$$\begin{aligned} E\{(X-Y)^2\} &= var\{X\} + var\{Y\} - 2cov\{X, Y\} + [E\{X\} - E\{Y\}]^2 \\ &= var\{X-Y\} + [E\{X-Y\}]^2 . \end{aligned}$$

Result 2.

For any two random variables X and Y and real-valued function ϕ,

$$cov\{\phi(X), Y\} = cov\{\phi(X), E(Y|X)\}.$$

Applying Result 1 to $\tilde{p}_z(\boldsymbol{Y})$ and Z immediately gives another form for the total error of prediction as follows.

$$\begin{aligned} pmse(\tilde{p}_z) &= E\left[\{\tilde{p}_z(\boldsymbol{Y}) - Z\}^2\right] \\ &= \text{var}\{\tilde{p}_z(\boldsymbol{Y}) - Z\} + [E\{\tilde{p}_z(\boldsymbol{Y}) - Z\}]^2 . \end{aligned} \tag{12.15}$$

Expression (12.15) has a familiar interpretation for mean squared errors as a variance, here of the raw prediction error $\tilde{p}_z(\boldsymbol{Y}) - Z$, plus the square of a bias term. A predictor $\tilde{p}_z(\boldsymbol{Y})$ of Z is called unbiased if the rightmost term in (12.15) is zero, that is, if the expected value of the predictor equals the expected value of the quantity to be predicted.

Result 2 allows us to demonstrate the equivalence of (12.14) and (12.15) as follows. Expression (12.15) is the same as,

$$pmse(\tilde{p}_z) = \text{var}\{\tilde{p}_z(\boldsymbol{Y})\} + var\{Z\} - 2\text{cov}\{\tilde{p}_z(\boldsymbol{Y}), Z\} + [E\{\tilde{p}_z(\boldsymbol{Y}) - Z\}]^2 . \tag{12.16}$$

Expression (12.14) is the same as,

$$mspe(\tilde{p}_z) = E\left([\tilde{p}_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}]^2\right) + \text{var}\{Z\} - \text{var}\left[E\{Z|\boldsymbol{Y}\}\right]. \tag{12.17}$$

Thus, if (12.16) and (12.17) are equivalent, so too are (12.14) and (12.15). Applying Result 1 (first line) to the first term of (12.17) gives

$$E\left([\tilde{p}_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}]^2\right) =$$

$$\text{var}\{\tilde{p}_z(\boldsymbol{Y})\} + \text{var}\left[E\{Z|\boldsymbol{Y}\}\right] - 2\text{cov}\left[\tilde{p}_z(\boldsymbol{Y}),\, E\{Z|\boldsymbol{Y}\}\right] + \left(E\left[\tilde{p}_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}\right]\right)^2,$$

which, when substituted back into (12.17) results in,

$$mspe(\tilde{p}_z) =$$

$$\text{var}\{Z\} + \text{var}\{\tilde{p}_z(\boldsymbol{Y})\} - 2\text{cov}\left[\tilde{p}_z(\boldsymbol{Y}),\, E\{Z|\boldsymbol{Y}\}\right] + \left(E\left[\tilde{p}_z(\boldsymbol{Y}) - E\{Z|\boldsymbol{Y}\}\right]\right)^2. \tag{12.18}$$

The application of Result 2 to (12.18) yields (12.16), showing the equivelance of (12.14) and (12.15).

12.5 Examples

Several examples are offered in this section to illustrate the uses and potential pitfalls of the various expressions developed previously in this chapter.

12.5.1 A Simple Normal Problem

Consider a model that takes $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$ to be independent and identically distributed with $N(\mu, \sigma^2)$ distributions. The minimum mean squared error predictor of a new random variable Y_0 from this same distribution is $p_0(\boldsymbol{Y}) = E(Y_0|\boldsymbol{Y}) = E(Y_0) = \mu$. The estimated predictor is then $\hat{p}_0(\boldsymbol{Y}) = \bar{Y}$. Using this as a plug-in in the mean squared prediction error of expression (12.8) gives

$$pmse(p_0) = \text{var}(Y_0) - \text{var}(\bar{Y}) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\left(1 - \frac{1}{n}\right), \tag{12.19}$$

which is clearly incorrect as it increases with increasing n. Using (12.15) with $p_z = \bar{Y}$ gives

$$pmse(\hat{p}_0) = \text{var}\{\bar{Y} - Y_0\} + \left[E\{\bar{Y} - Y_0\}\right]^2 = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right),$$

which is the correct expression for total mean squared error of prediction. If we recognize that the prediction mean squared error (12.8) should be calculated with the actual, rather than estimated, optimal predictor we would obtain $p_0(\boldsymbol{Y}) = E\{Y_0|\boldsymbol{Y}\} = E\{Y_0\} = \mu$, and (12.8) would then give $pmse(p_0) = \sigma^2$, which is an under-estimate of the actual value, but is not as serious an error as using the estimated predictor in the place of p_0 in (12.8) and ending up with (12.19). Finally, note that the independence in this example renders marginal and conditional prediction mean squared errors identical.

12.5.2 Mixed Linear Models

Mixed linear models present a situation in which it is beneficial to explicitly derive the joint distribution $F(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}_f)$, the optimal (under squared error loss) predictors of the random terms are unbiased, the marginal and conditional prediction mean squared errors are the same, and it is possible to compute the total $pmse$ (12.15) rather than estimate the version that ignores uncertainty due to estimation (12.8).

Consider first a linear random effects model presented for response variables $\{Y_{i,g} : i = 1, \ldots, n_g;\ g = 1, \ldots, G\}$ as,

$$Y_{i,g} = \boldsymbol{x}_{i,g}^T\boldsymbol{\beta} + \gamma_g + \epsilon_{i,g}, \tag{12.20}$$

where $\epsilon_{i,g} \sim iidN(0, \sigma^2)$ and $\gamma_g \sim iidN(0, \tau^2)$. In the general prediction problem $\boldsymbol{Z}$ would consist of the vector of random effects, $\boldsymbol{Z} = (\gamma_1, \ldots, \gamma_G)^T$.

We know for this model that,

$$\begin{aligned} E\{Y_{i,g}|\gamma_g\} &= \boldsymbol{x}_{g,i}^T\boldsymbol{\beta} + \gamma_g \qquad var\{Y_{i,g}|\gamma_g\} = \sigma^2 \\ E\{Y_{i,g}\} &= \boldsymbol{x}_{i,g}^T\boldsymbol{\beta} \qquad var\{Y_{i,g}\} = \sigma^2 + \tau^2 \\ cov\{Y_{i,g},\, Y_{k,g}\} &= \tau^2 \qquad cov\{Y_{i,g},\, \gamma_g\} = \tau^2 \end{aligned} \tag{12.21}$$

and all other covariances are zero.

For this model, the explicit form of the joint distribution of observable random variables $\boldsymbol{Y}$ and random effects to be predicted is available. Let $\eta_{i,g} \equiv \boldsymbol{x}_{i,g}^T\boldsymbol{\beta}$, and $n = \sum n_g$. The conditional distributions of the $Y_{i,g}$ given the γ_g are normal and the marginal distributions of the γ_g are normal, so then the joint distribution of $\boldsymbol{Y} = (Y_{1,1}, \ldots, Y_{n_1,1}, Y_{1,2}, \ldots, Y_{n_G,G})^T$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_G)^T$ is normal as,

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{Y} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{pmatrix} \Gamma & C \\ C^T & V \end{pmatrix} \right], \tag{12.22}$$

where $\boldsymbol{\mu}_y = (\eta_{1,1}, \ldots, \eta_{n_1,1}, \eta_{1,2}, \ldots, \eta_{nG,G})^T$, $\Gamma = \tau^2 I_{G\times G}$, V is an $n \times n$ block diagonal matrix with components V_g,

$$V_g = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \tau^2 & \ldots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \tau^2 & \ldots & \tau^2 \\ & & \vdots & & \\ \tau^2 & \tau^2 & \ldots & \tau^2 & \sigma^2+\tau^2 \end{pmatrix}_{n_g\times n_g}$$

and C is a $G \times n$ matrix with w, vth element $\tau^2 I(Y_{v,w} = Y_{v,g})$ (i.e., the first row has n_1 τ^2s followed by all zeros, the second row has n_1 zeros followed by n_2 τ^2s and then all zeros, etc.).

The minimum mean squared error predictors of the $\boldsymbol{\gamma}$ are the conditional expectations which are, from (12.22),

$$p_\gamma(\boldsymbol{Y}) = CV^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_y). \tag{12.23}$$

The mean squared prediction error of $p_\gamma(\boldsymbol{Y})$, ignoring error in estimation would be, from (12.8),

$$E\left[\text{var}\{p_\gamma(\boldsymbol{Y})\}\right] = \text{var}\{\boldsymbol{\gamma}\} - \text{var}\{p_\gamma(\boldsymbol{Y})\} = \Gamma - CV^{-1}C^T, \tag{12.24}$$

because, with V being a symmetric matrix, $\text{var}\{p_\gamma(\boldsymbol{Y})\} = CV^{-1}\text{var}(\boldsymbol{Y})V^{-1}C^T = CV^{-1}VV^{-1}C^T = CV^{-1}C^T$. Now, aside from variances contained in C and V, to estimate (12.23) requires estimation of $\boldsymbol{\mu}_y$ which depends only on the unknown parameter $\boldsymbol{\beta}$.

For this, and to give forms for estimated predictors of the random effects, it is helpful to write the model in matrix form as,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{12.25}$$

where now, X is an $n \times p$ matrix with ith row $\boldsymbol{x}_{i,g}^T$, Z is an $n \times G$ design matrix having ivth element $I(Y_{i,v} = Y_{i,g})$ (i.e., a matrix with 1s in the first column for all $Y_{i,1}$ and zeros elsewhere, 1s in the second column for all $Y_{i,2}$, etc.), and now we specify,

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \Gamma); \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_{n\times n})$$

Since this is simply an alternative way of writing the model, $\boldsymbol{Y}$ and $\boldsymbol{\gamma}$ still have the joint distribution (12.22), and now we might also write $C = \Gamma Z^T$. Assuming that an estimate $\hat{\boldsymbol{\beta}}$ is available through a least squares procedure, we will have,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(X^T V^{-1} X\right)^{-1} X^T \boldsymbol{Y} \\ \text{var}(\hat{\boldsymbol{\beta}}) &= X\left(X^T V^{-1} X\right)^{-1} X^T, \end{aligned}$$

and $\hat{\boldsymbol{\mu}} = X^T\hat{\boldsymbol{\beta}}$.

In this case, still ignoring the fact that we will need eventually to use estimates for V and $C = \Gamma Z^T$, we have the estimated predictors

$$\tilde{p}_\gamma(\boldsymbol{Y}) = CV^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}), \tag{12.26}$$

which are still unbiased for $\boldsymbol{\gamma}$. The total mean squared prediction error becomes, in contrast to (12.24),

$$E\left[\text{var}\{p_\gamma(\boldsymbol{Y})\}\right] = \Gamma - CPC^T, \tag{12.27}$$

where, (c.f., McCulloch and Searle, 2001, Chapter 6.5),

$$P = V^{-1} - V^{-1}\left(X^T V^{-1} X\right)^{-1} X^T V^{-1}.$$

There are then two options in a given application. One could estimate the variance matrices $C = \Gamma Z^T$ and V and use these in expression (12.24), which is the marginal (and conditional) mean squared prediction error of the optimal predictor (12.23). Alternatively , one could estimate the variance matrices and use these in expression (12.27), which is the total marginal (and conditional) mean squared prediction error of the estimated optimal predictor (12.26). You will see both in references on mixed linear models.

It should also be noted here that, while our focus has been on a random effects model as in expression (12.20), the forms of the predictor (12.23) and estimated predictor (12.26), as well as the forms of the mean squares in (12.24) and (12.27) would remain the same for any mixed linear model that is written in the form of expression (12.25). What changes for different models are the specific forms of the matrices Γ, C, and V in the joint distribution (12.22) but, given these changes, all else remains the same.

12.5.3 A Conditionally Specified Spatial Mixture Model

Kaiser, Daniels, Furakawa and Dixon (2002) develop a conditionally specified spatial mixture model for the analysis of air pollution, along with relevant predictors. This example illustrates a case in which predictors are developed for individual Z (as opposed to the simultaneous predictors allowed for mixed linear models), it is not necessary to explicitly determine the joint distribution $F(\boldsymbol{y}, z)$ in order to derive minimum mean squared error predictors, marginal and conditional mean squared prediction errors are not the same, and it is not possible (with what we currently know) to determine the total mean squared prediction errors.

The problem involved observations of particulate matter air pollution at a number of spatial locations $\boldsymbol{s}_i \equiv (u_i, v_i)$ where u_i denotes a horizontal coordinate and v_i denotes a vertical coordinate of a geographic projection from the globe to a Euclidean plane called a Universal Trans-Mecator projection (UTM); denote the number of stations as S. Observations were available at these locations on a daily basis, indexed as $t = 1, \ldots, T$, although not all locations reported a value on every day. In addition, for many but not all of the stations and days, replicate values were reported, indexed as $j = 1, \ldots, J_{i,t}$ for station $\boldsymbol{s}_i$ and day t. Response random variables are defined as $\boldsymbol{Y} \equiv \{Y(\boldsymbol{s}_i, t, j) : i = 1, \ldots, S;\ t = 1, \ldots, T;\ j = 1, \ldots, J_{i,t}\}$. The development of predictors also made use of the transformed daily averages for station and day combinations as $Z(\boldsymbol{s}_i, t) = (1/J_{i,t}) \sum_j \log\{Y(\boldsymbol{s}_i, t, j)\}$; $i = 1, \ldots, S;\ t = 1, \ldots, T$.

The data model consisted of specifying conditionally independent lognormal distributions for the $Y(\boldsymbol{s}_i, t, j)$ as,

$$f(y(\boldsymbol{s}_i,t,j)|\mu(\boldsymbol{s}_i,t),\,\sigma^2) =$$
$$\{2\pi\sigma^2\}^{-1/2}\{y(\boldsymbol{s}_i,t,j)\}^{-1}\exp\left(-\frac{1}{2\sigma^2}\left[\log\{y(\boldsymbol{s}_i,t,j)\}-\mu(\boldsymbol{s}_i,t)\right]^2\right). \quad (12.28)$$

In (12.28) the $\mu(\boldsymbol{s}_i,t)$ are the expected values of the $Z(\boldsymbol{s}_i,t)$ or the $\log\{Y(\boldsymbol{s}_i,t,j)\}$. Thus, these expectations are taken as the underlying air pollution process of primary interest. The process model consists of a conditionally specified Gaussian spatial model for the $\mu(\boldsymbol{s}_i,t)$. Specifically, for $i=1,\ldots,S$ and $t=1,\ldots,T$, let

$$g(\mu(\boldsymbol{s}_i,t)|\{\mu(\boldsymbol{s}_k,t):k\neq i\}) =$$
$$(2\pi\tau^2)^{-1/2}\exp\left[-\frac{1}{2\tau^2}\left\{\mu(\boldsymbol{s}_i,t)-A_i\left(\{\mu(\boldsymbol{s}_k,t):k\neq i\}\right)\right\}^2\right]. \quad (12.29)$$

Here, the functions A_i are the conditional expected values $E\left(\mu(\boldsymbol{s}_i,t)|\{\mu(\boldsymbol{s}_k,t):k\neq i\}\right)$, and are further modeled as,

$$A_i\left(\{\mu(\boldsymbol{s}_k,t):k\neq i\}\right)=\lambda(t)+\sum_{k\in R(t)}c_{i,k}(t)\{\mu(\boldsymbol{s}_k,t)-\lambda(t)\}, \quad (12.30)$$

where $R(t)$ is the set of locations reporting on day t.

Kaiser et al. (2002) examined a number of possible models for the $\lambda(t)$ and the $c_{i,k}(t)$ within the context of the problem. Our concern here is using this model to develop predictors for the process of interest $\mu(\boldsymbol{s}_0,t)$ and the observed process $Y(\boldsymbol{s}_0,t)$ at a spatial location $\boldsymbol{s}_0$ that is not in the set of observed locations on day t. Distributional results from the above model (see Kaiser et al. (2002) for details) that will be used to derive predictors are,

$$\begin{aligned}
\boldsymbol{\mu} &\sim Gau\left(\boldsymbol{\lambda},\,(I-C)^{-1}M\right)\\
\boldsymbol{Z}|\boldsymbol{\mu} &\sim Gau\left(\boldsymbol{\mu},\,V\right)\\
\boldsymbol{Z} &\sim Gau\left(\boldsymbol{\lambda},\,G+V\right)\\
\boldsymbol{\mu}|\boldsymbol{z} &\sim Gau\left(B,\,W\right). \quad (12.31)
\end{aligned}$$

In (12.31) $\boldsymbol{\mu} \equiv \{\mu(\boldsymbol{s}_i, t) : i = 1, \ldots, S;\ t = 1, \ldots, T\}$, $\boldsymbol{Z} \equiv \{Z(\boldsymbol{s}_i, t) : i = 1, \ldots, S;\ t = 1, \ldots, T\}$, and $\boldsymbol{\lambda} \equiv (\boldsymbol{\lambda}(1), \ldots, \boldsymbol{\lambda}(T))^T$ with $\boldsymbol{\lambda}(t)$ an $n(t)-$length vector of repeated values of $\lambda(t)$ where $n(t) = |R(t)|$, the number of locations reporting on day t. Also, with $N \equiv \sum_t n(t)$, I is the $N \times N$ identity matrix, M is an $N \times N$ diagonal matrix with diagonal entries τ^2, the conditional variance of the $\mu(\boldsymbol{s}_i, t)$, and C is an $N \times N$ matrix with entries $c_{i,k}(t)$ (see expression 12.29). Finally, V is an $N \times N$ matrix with entries $\sigma^2/J_{i,t}$, $G = (I - C)^{-1}M$, $B = (V^{-1} + G^{-1})^{-1}(V^{-1}\boldsymbol{Z} + G^{-1}\boldsymbol{\lambda})$, and $W = (V^{-1} + G^{-1})^{-1}$. The primary interest in this problem was the air pollution field represented by $\boldsymbol{\mu}$, and one objective of prediction was to predict the value $\mu(\boldsymbol{s}_0, t)$ at an unobserved spatial location $\boldsymbol{s}_0$ on a given day t. The minimum mean squared error predictor will be $E\{\mu(\boldsymbol{s}_0, t)|\boldsymbol{Y}(t)\}$. To derive this predictor, note that by transformation and sufficiency, $E\{\mu(\boldsymbol{s}_0, t)|\boldsymbol{Y}(t)\} = E\{\mu(\boldsymbol{s}_0, t)|\boldsymbol{Z}(t)\}$. Also note that the conditional distribution of $\mu(\boldsymbol{s}_0, t)$ given $\boldsymbol{\mu}(t)$ is the same as that of $\mu(\boldsymbol{s}_0, t)$ given $\boldsymbol{\mu}(t)$ and $\boldsymbol{Z}(t)$. The key step in the derivation of Kaiser et al. (2002) is that these two facts implies that

$$E\left\{\mu(\boldsymbol{s}_0, t)|\boldsymbol{Y}(t)\right\} = E\left\{\mu(\boldsymbol{s}_0, t)|\boldsymbol{Z}(t)\right\} = E\left[E\left\{\mu(\boldsymbol{s}_0, t)|\boldsymbol{\mu}(t)\right\}|\boldsymbol{Z}(t)\right] \quad (12.32)$$

This step may be presented in a general result as follows.

Result 3.

Let $p(\cdot)$ be generic notation for a density so that $p(x)$ denotes the density of a random variable X, $p(y|x)$ denotes the conditional density of Y given $X = x$, and so forth. For three continuous random variables X, Y, and Z, such that all necessary densities exist, if $p(x|y) = p(x|y, z)$, then

$$E\{X|Z\} = E\left[E\{X|Y\}\,|\,Z\right],$$

where the inner expectation is taken with respect to the distribution of X

given Y and the outer expectation is taken with respect to the distribution of Y given Z.

Proof

$$\begin{aligned}
E\left[E\{X|Y\}\,|\,Z\right] &= \int\int xp(x|y)\,dx\,p(y|z)\,dy \\
&= \int\int x\,p(x|y)\,p(y|z)\,dx\,dy \\
&= \int\int x\,p(x|y,\ z)\,p(y|z)\,dx\,dy \\
&= \int\int x\,\frac{p(x,y,z)}{p(y,z)}\,\frac{p(y,z)}{p(z)}\,dx\,dy \\
&= \int\int x\,p(x,y|z)\,dy\,dx \\
&= \int x\,p(x|z)\,dx \\
&= E\{X|Z\}
\end{aligned}$$

Once this result has been established, in the current model we then have,

$$\begin{aligned}
p_{\mu_0}(\boldsymbol{Y}) &= E\left\{\mu(\boldsymbol{s}_0,t)|\boldsymbol{Y}\right\} = E\left[\mu(\boldsymbol{s}_0,t)|\boldsymbol{Z}(t)\right] \\
&= E\left[E\{\mu(\boldsymbol{s}_0,t)|\boldsymbol{\mu}(t)\}|\boldsymbol{Z}(t)\right] \\
&= E\left[\lambda(t) + \sum_{k\in R(t)} c_{0,k}\{\mu(\boldsymbol{s}_k,t) - \lambda(t)\}\right] \\
&= \lambda(t) + \sum_{k\in R(t)} \left(c_{0,k}\left[E\{\mu(\boldsymbol{s}_k,t)\} - \lambda(t)\right]\right) \\
&= \lambda(t)\left(1 - \sum_{k\in R(t)} c_{0,k}\right) + \sum_{k\in R(t)} c_{0,k}b_k
\end{aligned} \tag{12.33}$$

where b_k is the kth element of the vector B.

While prediction of the underlying air pollution process $\boldsymbol{\mu}$ is of primary scientific interest, predictions using (12.33) provide no information about the aptness of the model used. From a statistical viewpoint, prediction of the observation model $\boldsymbol{Y}$ can be important for model assessment, such as via cross-validation. Result 3 again provides a crucial step in the development of a predictor for the observed process at an unobserved location, namely prediction of $Y(\boldsymbol{s}_0, t)$. In particular, note that the distribution of $Y(\boldsymbol{s}_0, t)$ given $\mu(\boldsymbol{s}_0, t)$ is the same as that of $Y(\boldsymbol{s}_0, t)$ given $\mu(\boldsymbol{s}_0, t)$ and $\boldsymbol{Y}$. Then Result 3 implies that

$$\begin{aligned} E\{Y(\boldsymbol{s}_0, t)|\boldsymbol{Y}\} &= E\left[E\{Y(\boldsymbol{s}_0, t)|\mu(\boldsymbol{s}_0, t)\} \,|\, \boldsymbol{Y}\right] \\ &= \exp\left\{\frac{1}{2}\sigma^2\right\} E\left[\exp\{\mu(\boldsymbol{s}_0, t)\}|\boldsymbol{Y}\right]. \end{aligned} \tag{12.34}$$

The second line of (12.34) follows from standard results for lognormal distributions. We again have that $E[\exp\{\mu(\boldsymbol{s}_0, t)\}|\boldsymbol{Y}] = E[\exp\{\mu(\boldsymbol{s}_0, t)|\boldsymbol{Z}]$ and

$$E\left[\exp\{\mu(\boldsymbol{s}_0, t)\}|\boldsymbol{Z}\right] = E\left(E\left[\exp\{\mu(\boldsymbol{s}_0, t)\}|\boldsymbol{\mu}\right]|\boldsymbol{Z}\right).$$

Now, from (12.31) the distribution of $\mu(\boldsymbol{s}_0, t)$ given $\boldsymbol{\mu}$ is Gaussian with variance τ^2 and mean,

$$\lambda(t) + \sum_{k \in R(t)} c_{0,k}\{\mu(\boldsymbol{s}_k, t) - \lambda(t)\},$$

so that,

$$E\left(\exp\{\mu(\boldsymbol{s}_0, t)\} \,|\, \boldsymbol{\mu}\right) = \exp\left[\lambda(t) + \sum_{k \in R(t)} c_{0,k}\{\mu(\boldsymbol{s}_k, t) - \lambda(t)\} + \frac{1}{2}\tau^2\right].$$

As a consequence,

$$
\begin{aligned}
& E\left[E\left(\exp\{\mu(\boldsymbol{s}_0,t)\}\,|\,\boldsymbol{\mu}\right)\,|\boldsymbol{Z}\right] \\
&= E\left(\exp\left[\lambda(t)+\sum_{k\in R(t)} c_{0,k}\{\mu(\boldsymbol{s}_k,t)-\lambda(t)\}+\frac{1}{2}\tau^2\right]\,|\boldsymbol{Z}\right) \\
&= \exp\left[\lambda(t)\left(1-\sum_{k\in R(t)} c_{0,k}\right)+\frac{1}{2}\tau^2\right]\,E\left[\exp\left\{\sum_{k\in R(t)} c_{0,k}\mu(\boldsymbol{s}_k,t)\right\}\,|\boldsymbol{Z}\right].
\end{aligned}
\tag{12.35}
$$

Now $\boldsymbol{\mu}|\boldsymbol{Z}\sim Gau(B,W)$ and we have that, given $\boldsymbol{Z}$,

$$
\sum_{k\in R(t)} c_{0,k}\mu(\boldsymbol{s}_k,t)\sim Gau\left(\sum_{k\in R(t)} c_{0,k}\,b_k,\;Q\right)
$$

where

$$
Q=\sum_{k\in R(t)} c_{0,k}^2\,\psi_k^2+\sum_{k,m\in R(t)}\sum_{k\neq m} c_{0,k}c_{0,m}\,\psi_{k,m},
$$

and ψ_k^2 is the kth diagonal element of W and $\psi_{k,m}$ is the km^{th} element of W. The expectation in (12.35) is then $\exp\{\sum_k c_{0,k}b_k+(1/2)Q\}$ and

$$
\begin{aligned}
E\left\{\mu(\boldsymbol{s}_0,t)|\boldsymbol{Y}\right\} &= E\left\{\mu(\boldsymbol{s}_0,t)|\boldsymbol{Z}\right\} \\
&= E\left[E\left(\exp\{\mu(\boldsymbol{s}_0,t)\}\,|\,\boldsymbol{\mu}\right)\,|\boldsymbol{Z}\right] \\
&= \exp\left\{\lambda(t)\left(1-\sum_{k\in R(t)} c_{0,k}\right)+\frac{1}{2}\tau^2+\sum_k c_{0,k}b_k+(1/2)Q\right\}
\end{aligned}
$$

and substitution into (12.34) results in,

$$
\begin{aligned}
p_{Y_0}(\boldsymbol{Y}) &= E\left[Y(\boldsymbol{s}_0,t)|\boldsymbol{Y}\right] \\
&= \exp\left\{\lambda(t)\left(1-\sum_{k\in R(t)} c_{0,k}\right)+\frac{1}{2}(\sigma^2+\tau^2)+\sum_{k\in R(t)} c_{0,k}b_k+\frac{1}{2}\,Q\right\} \\
&= \exp\left\{p_{\mu(\boldsymbol{s}_0,t)}(\boldsymbol{Y})+\frac{1}{2}(\sigma^2+\tau^2+Q)\right\}.
\end{aligned}
\tag{12.36}
$$

Expression (12.36) actually corrects an error in expression (18) of Kaiser et al. (2002) which dropped the $(1/2)\sigma^2$.

In this example we have not derived any forms for mean squared prediction errors, either conditional or marginal. Doing so would be a complex matter, which is not an uncommon occurrence in situations for which derivation of the predictors themselves is a rather involved matter. In such situations it is typical to rely on assessment of prediction error through the use of cross-validation, which was the approach taken by Kaiser et al. (2002).

Part V

Topics in Bayesian Analysis

Part VI

Topics in Bayesian Analysis

Chapter 13

Inference in Hierarchical Models

13.1 Viewpoints of Hierarchical Models

In Stat 520 we indicated that there are a number of ways to think about the manner in which hierarchical models represent the real world, and interpretation of such models is often less than a trivial manner. This is related to the issue of which components of a statistical model embody the scientific mechanism or phenomenon of interest. Recall that one of our goals in formulation of a parametric statistical model is to represent (stochastically, not literally) the mechanism in a small number of parameters, or functions of parameters. Since hierarchical models can contain a large number of parameters through the combination of data model and random parameter distributions, pinpointing which of those should be considered as representing the scientific mechanism can become an involved exercise.

13.1.1 Mixtures and Multi-stage Priors

Included in the Statistics 520 notes was a discussion about two viewpoints that we might take toward making inference based on hierarchical models. Recall that the basic issue was whether or not one is interested in making inference on the basis of the posterior distributions of data model parameters. The issue can be summarized by contrasting a viewpoint in which we consider having a mixture model on which a prior is specified, versus a data model with a multi-stage prior. We first recap the 520 discussion for completeness.

Hierarchical Models as Mixtures

Here, we would think of the model as,

$$\begin{array}{ll} \text{Data Model:} & f(\boldsymbol{y}|\boldsymbol{\theta}) \\ \text{Mixing Distribution:} & g(\boldsymbol{\theta}|\boldsymbol{\lambda}) \\ \text{Prior:} & \pi(\boldsymbol{\lambda}). \end{array}$$

which leads to

$$\begin{array}{ll} \text{Data Model:} & h(\boldsymbol{y}|\boldsymbol{\lambda}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\, g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta} \\ \text{Prior:} & \pi(\boldsymbol{\lambda}). \end{array}$$

Whether we perform the integration analytically or through simulation in an MCMC algorithm, the posterior of primary interest is $p(\boldsymbol{\lambda}|\boldsymbol{y})$. If we adopt this view of a hierarchical model we are considering our overall model to represent a scientific mechanism or phenomenon of interest as the mixing distribution $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ which is controlled by the parameter $\boldsymbol{\lambda}$. If the data model is constructed from conditionally independent pieces that represent different situations which have been observed,

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{k} f_i(\boldsymbol{y}_i|\boldsymbol{\theta}_i),$$

then the manner in which the mechanism has manifested itself in situation i is embodied in $\boldsymbol{\theta}_i$, and this is true for each situation observed, $i = 1, \ldots, k$.

Hierarchical Models with Multi-stage Priors

Under this viewpoint we would consider a model to consist of,

$$\begin{aligned} &\text{Data Model:} && f(\boldsymbol{y}|\boldsymbol{\theta}) \\ &\text{Stage 1 Prior:} && g(\boldsymbol{\theta}|\boldsymbol{\lambda}) \\ &\text{Stage 2 Prior:} && \pi(\boldsymbol{\lambda}). \end{aligned}$$

which leads naturally to

$$\begin{aligned} &\text{Data Model:} && f(\boldsymbol{y}|\boldsymbol{\theta}) \\ &\text{Prior:} && \pi(\boldsymbol{\theta}) = \textstyle\int g(\boldsymbol{\theta}|\boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, d\,\boldsymbol{\lambda}, \end{aligned}$$

Again, whether we perform this integration analytically or through application of an MCMC algorithm (which is usually the case), the focus of inference is likely to be $p(\boldsymbol{\theta}|\boldsymbol{y})$. If we adopt this viewpoint we are most likely considering the scientific mechanism or phenomenon of interest to be represented by the data model parameter $\boldsymbol{\theta}$. If the data model is constructed from conditionally independent pieces that represent problems with a similar structure,

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{k} f_i(\boldsymbol{y}_i|\boldsymbol{\theta}_i),$$

then the overall mechanism is considered to contain a collection of k similar or related mechanisms embodied in the $\boldsymbol{\theta}_i$.

13.1.2 Examining the Distinction

Is there really a distinction to be made between the two views of hierarchical models offered in the previous sub-sections? Certainly there is no mathematical distinction. All of the same distributions are available under either

viewpoint, in particular the posteriors $p(\boldsymbol{\lambda}|\boldsymbol{y})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$. The major distinction made was in the case that the data model can be constructed as a product of conditionally independent pieces $f(\boldsymbol{y}_i|\boldsymbol{\theta}_i)$; $i = 1, \ldots, k$. The language used previously was that if we are considering a hierarchical model to be a Bayesain formulation of a mixture model, we are thinking of the $\boldsymbol{\theta}_i$ as representing different manifestations of an overall scientific mechanism in different situations. In contrast, if we are viewing the problem as a data model with multi-stage prior then we are thinking of the $\boldsymbol{\theta}_i$ as representing different particular but related mechanisms for a set of related problems. For this to amount to something more than an exercise in semantic gymnastics, there must be some distinction between a set of "manifestations of a scientific mechanism in different situations" and a set of "particular scientific mechanisms for related problems". It is difficult to give a completely general description of the distinction, but we can say that it depends on the concepts of temporal extent and replication. What is being called a situation here is characterized by short temporal duration and lack of replication, relative to what is being considered a defined problem. The implication is that, although these characteristics operate along a continuum, a situation does not persist for a long enough period of time to allow repeated observation, and is thus not replicable. A defined problem, in contrast, results when a set of conditions (a situation) persists long enough to allow re-visitation or is characterized by physical entities that persist, in which case replicate observation of the situation (now termed a problem) is possible.

13.1.3 Determining Which View to Adopt

In the last portion of of the material devoted to this topic in the Stat 520 notes, several items were offered as providing some guidance for which viewpoint might be considered more appropriate in a given problem. The first of these was:

1. If all situations of interest have been observed then we are likely interested in the posterior of data model parameters $p(\boldsymbol{\theta}|\boldsymbol{y})$, where $\boldsymbol{\theta}$ is the vector of parameters over all situations.

2. If the observed situations constitute a random sample of the situations of interest, then we are likely interested in the posterior of mixing distributions parameters $p(\boldsymbol{\lambda}|\boldsymbol{y})$.

Because in many (if not most) problems neither of these two ideals will be met, we offered another way to gain some guidance about which distributions should be the primary focus of inferece:

1. If a situation covered by the data model could, at least in principle, lead to an additional (future) observation from the model with the same values of the data model parameters as in analysis, then we may well be interested in making inference based on the posterior distributions of data model parameters.

2. If a situation covered by the data model could not lead to an additional observation from the model with the same values of the data model parameters as in analysis, then we are unlikely to be interested in making inference based on the posterior distributions of data model parameters.

Soil Carbon

Soil organic carbon (SOC) is the carbon contained in the organic matter of soil, which itself makes up a fairly small percentage of total soil mass. However, SOC is an important determinant of the potential for two very different processes, both of which have implications for climate change. One process, *carbon sequestration* occurs when soil captures carbon dioxide (CO_2) from the atmosphere and stores it in the soil carbon pool. The other *soil respiration* releases CO_2 from soil, primarily through microbial activity. Also important in this issue is that total soil carbon can be divided into two (sometimes three) pools. Sufficient for our purposes is to consider two pools, the *stable* pool and the *labile* pool, although the characteristics that determine this division are actually a gradient, rather than a sharp division. The stable organic carbon pool refers to carbon that is bound to humus and soil minerals. The stable pool is persistent, and change occurs slowly over a period of multiple years and decades. Carbon sequestration in the stable pool has been recognized as a potential mitigating factor in the overall increase of CO_2 in the atmosphere. The labile organic carbon pool, in contrast, contains carbon from decaying plant material and carbon incorporated in soil organisms. The labile pool can change much more rapidly than the stable pool, over periods of hours or days, primarily through the process of soil respiration due largely to microbial activity, in which carbon is released into the atmosphere in the form of CO_2. Many factors may influence the rate of soil respiration, one of the primary of which is believed to be soil temperature (e.g., Raich, Kaiser, Dornbush and Martin, 2023). Consider, then a study to determine the mean levels of stable soil organic carbon and soil respiration in two of the eight phisiogeographic regions of the United States, Intermontane Plateaus and

Interior Plains. Physiogeographic regions are determined by several factors, one of the most important of which is the underlying geology which is, in turn, an important determinant of the SOC that is contained in the stable pool. The analysis for this study boils down to a basic two group mean comparison.

Suppose a study is conducted by selecting n_1 forested study sites of an appropriate size (maybe an ecological unit such as a watershed, or a political unit such as a county) in the Intermontane Plateaus region, and similarly n_2 grassland study sites in the Interior Plains region. Within each study site, m_i sampling locations are selected for $i = 1, \ldots, n_1$ or $i = 1, \ldots, n_2$, depending on the region involved. At each sampling location, we define two random variables, $Y_{i,j}$ corresponding to the SOC contained in the stable pool (g/kg), and $X_{i,j}$, corresponding to soil respiration rate ($g\, m_{-2}\, d^{-1}$). We assume that the sampling locations within a study site are spaced sufficiently so that the $Y_{i,j}$ ($X_{i,j}$) can be assumed to be conditionally independent across sampling locations $j = 1, \ldots, m_i$. A model for either SOC contained in the stable pool, or the rate of soil respiration in one physiogeographic region might be,

$$\begin{aligned} Y_{i,j}(\text{ or } X_{i,j}) &\sim \mathrm{N}(\mu_i, \sigma_i^2) \\ \mu_i &\sim \text{iid } \mathrm{N}(\mu_0, \tau_0^2) \\ \sigma_i \tilde{} \text{iid } \mathrm{IG}(\alpha, \beta) & \\ \mu_0 &\sim \text{iid } \mathrm{N}(\lambda, \psi) \\ \tau_0^2 &\sim \text{iid } \mathrm{Unif}(0, T) \\ \alpha &\sim \mathrm{Unif}(0, A) \\ \beta &\sim \mathrm{Ga}(\eta_1, \eta_2) \end{aligned} \tag{13.1}$$

This model has the structure of a two-level hierarchical model for one-sample

problems. Relative to the discussion of this chapter, if viewed as Bayesian formulations for analysis of mixture models, interest would focus on the distributions of the data model parameters, and the controlling parameters of those distributions. In particular, if our central interest is in mean values, μ_0 would be the focus of inference, which we would approach through the posteriors $p(\mu_0|\boldsymbol{y})$ or $p(\mu_0|\boldsymbol{x})$. While the posterior of μ_0 might also be of interest, a primary concern from the viewpoint of models with multiple priors would likely be the data model means, μ_i, and the posterior distributions $p(\mu_i|\boldsymbol{y})$; $i = 1, \ldots, n_1$ or $p(\mu_i|\boldsymbol{x})$; $i = 1, n_2$.

Given the science underlying this problem, it would seem clear that for SOC contained in the stable pool this model would lend itself to interpretation as a model with multiple prior distributions, while for soil respiration the model would lend itself to interpretation as a Bayesian formulation of a mixture model. The stable pool of SOC changes slowly over periods of multiple years or decades. One might very well want to have an estimate of the mean SOC in the stable pool for study sites based on $p(\mu_i|\boldsymbol{y})$; $i = 1, \ldots, n_1$ which could be used, inter alia, to help guide land use practices for those sites. We could, at least hypothetically, obtain more observations with distributions identical to the $Y_{i,j}$ with the same true value of μ_i because this quantity will not have changed much if our new observations are taken within the same decade. The posterior $p(\mu_0|\boldsymbol{y})$ could still be used to draw conclusions about differences between physiogeographic regions. In contrast, soil respiration can change rapidly, and knowing what $p(\mu_i|\boldsymbol{x})$ is for this study provides little information about what μ_i might be in another study period separated from our current time by even weeks. Thus, it is unreasonable to assume that we could obtain more observations from the same data model with the same values of μ_i for soil respiration. Here, our inference would necessarily focus

on the posterior $p(\mu_0|\boldsymbol{x})$ and a comparison of this posterior between regions.

Under the viewpoint that a hierarchical model is a Bayesian formulation of a general mixture model we have said that the scientific mechanism is embodied in the mixing distribution and its controlling parameters. Inference about this (conceptualization) of the mechanism can come from the posterior distributions of those controlling parameters or simple functions of those parameters, as in the example of soil respiration described previously. But the mixing distribution itself can be taken as a representation of the probabilities with which the mechanism manifests itself in different situations. We may very well want to make inferential statements on the basis of this distribution that go beyond simple functions of its parameters. Such statements often involve the concept of risk, such as the probability that a regulatory threshold is exceeded, or a failure of some system component occurs. We will address the formulation and computation of such statements directly.

13.2 Posterior Predictive Inference

Posterior predictive distributions are often discussed in the context of model assessment and the production of posterior predictive $p-$values. The posterior predictive distributions that are relevant to model assessment are distributions for additional observable effects $\boldsymbol{Y}^*$ assumed to arise from the same underlying process as the observed data $\boldsymbol{Y}$, namely $p(\boldsymbol{y}^*|\boldsymbol{y})$. Since both $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are integrated out of this distribution it does not matter which of the viewpoints described previously one is operating under. Let $p(\cdot)$ denote a

generic probability density or mass function. Then

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}^*, \boldsymbol{y})}{p(\boldsymbol{y})} \\ &= \frac{1}{p(\boldsymbol{y})} \int \int p(\boldsymbol{y}^*, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\lambda})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda} \end{aligned} \tag{13.2}$$

If our basic view is that of a hierarchical model as involving a Bayesian analysis of a mixture model we might write a continuation of (13.2) as

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{y}) &= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y}^*, \boldsymbol{y}, \boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\ &= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y}^*, \boldsymbol{y}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\ &= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y}^*|\boldsymbol{\lambda})\, p(\boldsymbol{y}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\ &= \int p(\boldsymbol{y}^*|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda}|\boldsymbol{y})\, d\boldsymbol{\lambda}. \end{aligned} \tag{13.3}$$

If our basic view is that of a hierarchical model involving a multi-level prior for $\boldsymbol{\theta}$ we might continue (13.2) as

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{y}) &= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y}^*, \boldsymbol{y}, \boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y}^*, \boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y}^*|\boldsymbol{\theta})\, p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int p(\boldsymbol{y}^*|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}. \end{aligned} \tag{13.4}$$

Notice that both (13.3) and (13.4) depend on conditional independence of $\boldsymbol{Y}^*$ and $\boldsymbol{Y}$ given either $\boldsymbol{\lambda}$ or $\boldsymbol{\theta}$, although (13.3) makes use of the posterior of $\boldsymbol{\lambda}$ while (13.4) makes use of the posterior of $\boldsymbol{\theta}$.

For general mixture models consisting of a data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ and a random parameter model $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$ we typically think of the scientific mechanism

of interest as represented by the mixing distribution or random parameter model, which governs the relative frequencies with which the mechanism manifests itself under different situations. Thus, a focus of inference is $g(\boldsymbol{\theta}|\boldsymbol{\lambda})$. In a frequentist analysis we considered the fitted mixture distribution $g(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$ and functionals of this distribution, for which we can often determine asymptotic standard errors through use of the delta method. In a Bayesian analysis of a mixture model we could use a similar strategy, with $\hat{\boldsymbol{\lambda}}$ in $g(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$ being taken as the posterior expected value or mode of $p(\boldsymbol{\lambda}|\boldsymbol{y})$. Doing so, however, would ignore uncertainty in our knowledge of $\boldsymbol{\lambda}$.

An alternative that is perhaps more justifiable, or at least more in line with Bayesian philosophy, is to make inference about the process represented in the model by the mixing distribution $g(\cdot)$ based on a posterior predictive distribution for a new value $\boldsymbol{\theta}^*$, assumed to follow the same model as $\boldsymbol{\theta}$. This posterior predictive distribution can be developed as

$$\begin{aligned} p(\boldsymbol{\theta}^*|\boldsymbol{y}) &= \frac{p(\boldsymbol{\theta}^*, \boldsymbol{y})}{p(\boldsymbol{y})} \\ &= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{\theta}^*, \boldsymbol{y}, \boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\ &= \frac{1}{p(\boldsymbol{y})} \int f(\boldsymbol{y}|\boldsymbol{\lambda})\, g(\boldsymbol{\theta}^*|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\ &= \frac{1}{p(\boldsymbol{y})} \int g(\boldsymbol{\theta}^*|\boldsymbol{\lambda})\, p(\boldsymbol{y}, \boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\ &= \int g(\boldsymbol{\theta}^*|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda}|\boldsymbol{y})\, d\boldsymbol{\lambda}. \end{aligned} \tag{13.5}$$

In (13.5), the posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y})$ has been integrated over and, thus, the uncertainty in our knowledge of $\boldsymbol{\lambda}$ has been accounted for. Using (13.5) to make inference about the scientific process of concern is very much in line with a view of the model as a mixture (on which we have put a prior for the purpose of conducting a Bayesian analysis). Consider, again,

the case that the data model is constructed for a set of situations that have been observed, situations that will never occur again, but are representative of the total collection of situations about which we desire inference. The distribution of "new" data model parameters $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*)$ tells us what we should anticipate in any additional collection of k situations drawn from the total collection of concern.

13.3 Case Study: Regional Analysis of Nitrogen Trials

In this section we will consider a case study that illustrates a number of issues connected with the use of hierarchical models. In particular, we will need to consider for this problem the question of what quantities or distributions are most appropriate for the purpose of making inferences. The related question of how one might go about assessing the model as a representation of the mechanism generating the data is important, particularly with respect to assessing the modeling of unobservable quantities.

13.3.1 Problem Background

The data to be analyzed come from a set of nitrogen trials conducted in Iowa in the years 2001, 2002, and 2003. Nitrogen trials are conducted as the primary source of information used to set fertilizer recommendations for corn by agronomists. Such recommendations can have broad impacts on at least two fronts, economic and environmental. To fully appreciate this, we first need to have a rough understanding of why there is (or is not) a need to fertilize corn, and the fate of nitrogen fertilizer applied to farm fields.

Nitrogen supply to corn

Minor sources of nitrogen to corn include rain and release from clay minerals, but major sources, aside from fertilizer, are soil organic matter and crop residue (plant material left in the field from the previous year). But the nitrogen in these sources is not readily available to corn, which needs what is called *plant available nitrogen*, which is primarily ammonium and nitrate, both inorganic forms. Organic matter nitrogen in soil and crop residue can be converted to plant available nitrogen by the action of microorganisms in the soil, a process called nitrogen mineralization. The amount of mineralization that occurs in a year depends on soil and weather conditions, with warm, moist, aerated, neutral-pH soil being the best. Because of this, it is impossible to predict with any degree of accuracy and precision prior to the growing season how much nitrogen will be available to corn. To make this uncertain situation even more complex, corn needs different amounts of nitrogen at different stages of plant growth which, although somewhat predictable in time, does vary from year to year as well.

The upshot of this discussion is that determining a precise amount of non-fertilizer nitrogen that will be available to corn in a given area in a given year at just the right growth stage is impossible. Historically, the solution has been to provide corn with more nitrogen in the form of fertilizer than it could possibly use. Nitrogen was cheap, potential adverse impacts of excessive fertilization were not well known or appreciated, and whatever was not used by the crop was assumed to be harmlessly washed away in runoff from fields.

Need for reassessment

Since the days of cheap nitrogen fertilizer and reduced environmental concern several things have changed that motivate a reassessment the way in which fertilizer recommendations are made. First, we are now aware that high levels of nitrogen fertilizer in streams and rivers is a type of pollution. High nitrate concentrations in groundwater have been associated with some diseases such as breast cancer. A major concern for agronomists in the United States is the contribution of nitrogen fertilizer runoff to the anoxic zone (or so-called dead zone) that forms in the Gulf of Mexico each summer. Anoxic zones in coastal waters are regions in which oxygen has been depleted from the bottom layer of near-shore waters. The anoxic zone in the Gulf of Mexico forms around the mouth of the Mississippi river in Louisiana and stretches along the coast east to Mississippi and west to Texas. Although the size of this zone varies from year to year, in the 1990s it was somewhat over $20,000$ km^2, which is about the size of the state of New Jersey. The anoxic zone forms because high levels of nutrients, primarily nitrogen fertilizer, that are carried down the Mississippi River from the midwest corn growing states result in algal blooms on the surface of the water. These algae then die, sink, and decay on the bottom, with the process of decay consuming most of the available oxygen in the lower water level. The result is that ground-dwelling marine life such as shrimp, crabs, and many species of fish, must leave or die. Not only is this of concern from a purely environmental viewpoint, shrimpers and other marine fishers are adversely impacted economically and may include the need for them to relocate. The idea of legislating the application of fertilizer for corn has been raised in Congress and elsewhere.

Over roughly the same period of time during which environmental con-

cerns were increasing, so too was the price of nitrogen fertilizer. If one greatly over-fertilizes, yield will be maximized. But, depending on the relative prices of fertilizer and corn, net profit may be reduced below what would have resulted from a certain level of under-fertilizing. Many agronomists in the midwest are currently making fertilizer recommendations based on the concept of an *economically optimal nitrogen rate*, defined as the nitrogen fertilization rate beyond which any additional fertilizer will fail to increase yield sufficiently to pay for the additional fertilizer. Of course, what maximum yields are, as well as the prices of nitrogen fertilizer and corn, vary from year to year and place to place just as does the amount of nitrogen mineralization that occurs from non-fertilizer sources.

Regional risk-based fertilizer recommendations

One approach to making fertilizer recommendations that has gained favor with agronomists is based on communicating risks to producers associated with various fertilization rates. Ideally, such risks could be communicated as probabilities, such as the probability of various levels of decreased profit from either over-fertilizing or under-fertilizing. Currently (to the best of my knowledge) what is reported in fertilizer recommendations is a range of rates deemed to be profitable. Adding probability statements would be a step forward. Replacing these ranges with probabilities of net profits and losses across a gradient of fertilization rates would be more of an eventual goal.

As indicated previously, the data used to determine fertilizer recommendations are regional collections of individual nitrogen trials. The structure of such trials will be expanded on in what follows, but the basic idea is to conduct field experiments with varying rates of nitrogen application and relate

yield to nitrogen rate through a nonlinear regression. Collections of estimates from such trial-specific statistical models are used in a non-formal manner to determine the range of profitable fertilization rates. The basic goal of a statistical analysis at this point is to provide the quantification necessary to replace this informal process with a more exactly defined procedure.

13.3.2 Nitrogen Trials and Statistical Models

A typical nitrogen trial is laid out as a randomized complete block design of small plots in a field. Treatments are usually 5 to 6 levels of nitrogen application; the data we examine here had 6 levels of 0, 45, 90, 135, 180 and 225 kg/ha. The response variables of interest are taken to be corn yield in kg/ha, and observations consist of average yield over 4 replicates (the blocks in the experimental design).

Representative results from four trials are presented in Figure 13.1. The fitted curves in this figure will be discussed shortly. Note that the response functions are all of the same general shape, increasing in a linear fashion to a plateau. This is what is expected in nitrogen trials, with nitrogen fertilization increasing yield up to a point beyond which no further increase occurs. Despite the similarity of general shape, however, also note that the four curves of Figure 13.1 differ in the level of the plateau, the steepness of the initial linear increase, and the nitrogen rate at which the plateau is reached.

Not all nitrogen trials result in such "picture perfect" data, however. Data from four additional trials are presented in Figure 13.2. For the data in the top two panels of Figure 13.2 it was possible to fit the same model as used for the data of Figure 13.1, although this was not easily accomplished and finding starting values was difficult. For the trials resulting in the data

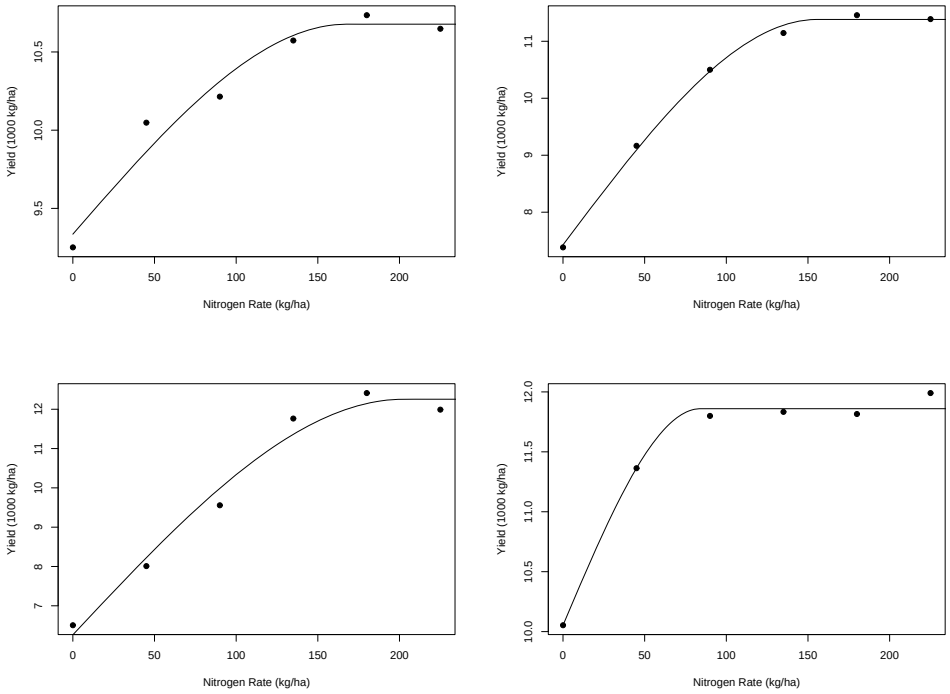

Figure 13.1: Examples of data from individual nitrogen trials. Fitted curves are from the spherical response model with estimation by generalized least squares.

of the lower two panels of Figure 13.2 I was unable to fit the same model as used for the other trials.

Analysis of Individual Trials

A first step in developing an overall analysis for the set of 43 nitrogen trials from Iowa is to fit a common model to data from individual trials separately. The approach I used was that of an additive error nonlinear regression model. An additive error model was chosen because (1) at least for nitrogen trials such as those of Figure 13.1, variability about a response function was small, and, perhaps more importantly, (2) with only six values of nitrogen rate (x) and yield (y) for a single trial, there is very little information in a data set that could be used to consider an appropriate random model component before determining an appropriate form of expectation function. This latter point also motivated the use of a constant variance error term. While an assumption of constant variance may or may not be truly appropriate for the problem, there is simply not enough information in the data from an individual trial to consider any more complex error structure.

Underlying scientific knowledge suggests that an expectation function should have the behavior of those shown in Figure 13.1. A convenient function with these properties is that of a theoretical spherical variogram model from spatial statistics. Here, however, we are not concerned with the use of this function as a model for variograms (which are structures from what is called geostatistical analysis), we simply want to use it as a function to describe the relation between expected yield and nitrogen rate. To formulate a model for an individual nitrogen trial then, let Y_j be a random variable

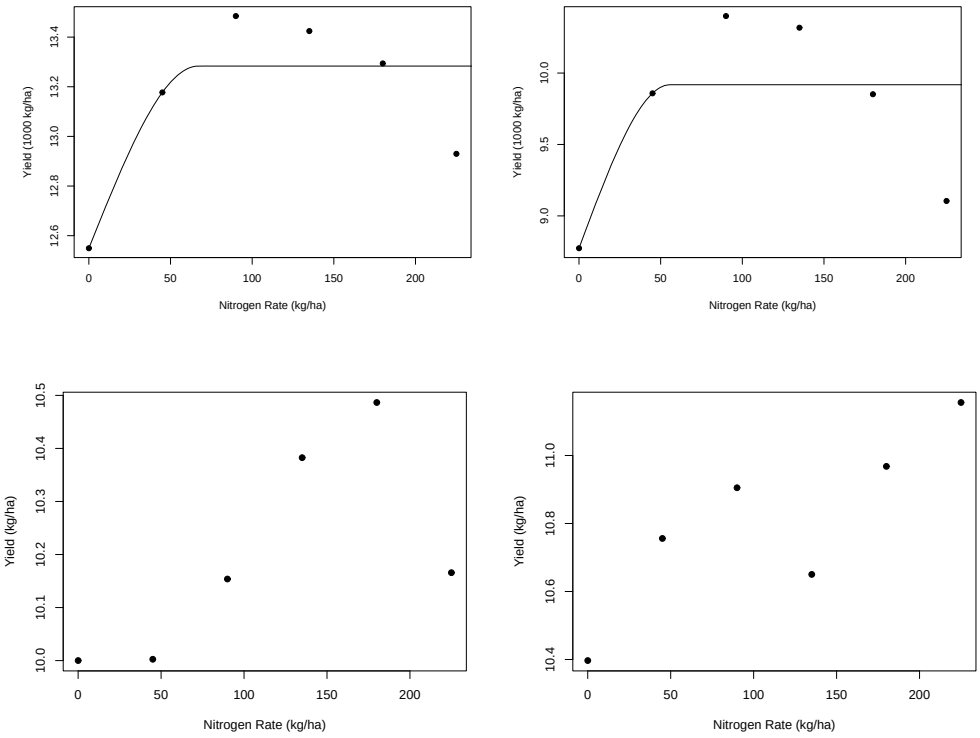

Figure 13.2: Examples of data from individual nitrogen trials. Fitted curves are from the spherical response model with estimation by generalized least squares.

associated with yield at nitrogen rate x_j and take, for $j = 1, \ldots, n$

$$Y_j = g(x_j, \boldsymbol{\theta}) + \sigma\epsilon_j, \tag{13.6}$$

where $\epsilon_j \sim iidN(0,1)$. In (1) we will let $\boldsymbol{\theta} = (c_0, c_1, a)^T$ and,

$$g(x_j, \boldsymbol{\theta}) = \begin{cases} c_0 + c_1 \left\{ \frac{3}{2}\left(\frac{x_j}{a}\right) - \frac{1}{2}\left(\frac{x_j}{a}\right)^3 \right\} & 0 < x_j \leq a \\ c_0 + c_1 & x_j > a \end{cases}$$

In (13.6), c_0 is an intercept term (yield with no fertilizer), $c_0 + c_1$ is the maximum or plateau value, and a is the minimum nitrogen rate at which the

yield reaches its maximum. Thus, the parameters of this model have ready interpretation in the context of the problem.

A next step would be to fit model (13.6) to data from each of the 43 nitrogen trials separately. To accomplish this, a function was written to automatically choose starting values for use in a Gauss-Newton algorithm. The starting value for c_0 was selected as the value of yield at nitrogen rate 0. The starting value for c_1 was selected as the value of yield at nitrogen rate 90 (the third value) minus the yield at nitrogen rate 0. Finally, the starting value of a was set at 90. Running the 43 individual trials through this function resulted in 28 trials for which estimates were found. The agronomist who provided the data had warned that there were some trials included in the data that agronomists call non-responsive, trials for which no discernible response to nitrogen fertilization was seen. There was a need to define a formal rule to declare a trial as non-responsive. Of the 15 trials for which automatically chosen starting values did not lead to convergence of the estimation algorithm, it was possible to produce estimates for 7 through more careful selection of starting values or deletion of one aberrant data value. For the other 8 trials convergence could not be achieved with any starting values tried. This resulted in dividing the entire data set into three groups, the "non-responsive" group consisting of the 8 trials for which no estimates could be produced, the "difficult" group consisting of the 7 trials for which estimates could be found but only through careful selection of starting values, and the "nice" group for which estimates could easily be found through the automated selection of starting values. Incidentally, the agronomist suggested that there might be 6 non-responsive trials included in the entire data set, based on visual inspection of the individual scatterplots. Using a more formal rule, we would suggest that there were 8 such trials.

A Regional Hierarchical Model

At this point, it may be beneficial to review the implications of the problem background section for formulation of an overall model for the entire set of nitrogen trials. There is expected to be large variability among responses of yield to fertilization across the set of trials. We currently do not have the ability to make use of auxiliary information (such as additional covariates) to predict what the responses to fertilizer will be in any one given trial. One of our goals is to quantify this uncertainty based on the set of nitrogen trial data available. This is a natural situation in which to contemplate the use of a hierarchical or mixture model. To formulate such a model we re-write the data model of (13.6) making the assumption that the parameters c_0, c_1 and a are random variables across trials, which we now index as $i = 1, \ldots, T$. Specifically, we now define $Y_{i,j}$ as the yield associated with nitrogen rate $x_{i,j}$ in trial i and take the data model to be

$$Y_{i,j} = g(x_{i,j}, \boldsymbol{\theta}_i) + \sigma\epsilon_{i,j} \tag{13.7}$$

where $\epsilon_{i,j} \sim iidN(0,1)$. In (2) we will let $\boldsymbol{\theta}_i = (c_{0i}, c_{1i}, a_i)^T$ and,

$$g(x_{i,j}, \boldsymbol{\theta}_i) = \begin{cases} c_{0i} + c_{1i}\left\{\frac{3}{2}\left(\frac{x_{i,j}}{a_i}\right) - \frac{1}{2}\left(\frac{x_{i,j}}{a_i}\right)^3\right\} & 0 < x_{i,j} \leq a \\ c_{0i} + c_{1i} & x_{i,j} > a_i \end{cases}$$

We now need to specify mixing distributions for the c_{0i}, c_{1i} and a_i; $i = 1, \ldots, T$. Each of these parameters must be positive for all i in model (13.7). One way to select reasonable distributions might be to consider the estimates from individual model fits as "surrogate observations" of these parameters, and formulate marginal distributions for each separately through inspection of histograms or stem-and-leaf plots of the estimates for individual trials. We will use only estimates from the group of nice trials as defined previously.

Since we have only 28 such values, stem-and-leaf plots might be a better option than histograms. Doing so results in the following plots.

The plot for values of c_{0i} is,

```
The decimal point is at the |

 5 | 9
 6 | 39
 7 | 34499
 8 | 0566888
 9 | 33468
10 | 129
11 | 1135
12 | 5
```

The plot for values of c_{1i} is,

```
The decimal point is at the |

0 | 77
1 | 133688
2 | 567778
3 | 122567
4 | 0378
5 | 55
6 | 03
```

and the plot for values of a_i is,

```
The decimal point is 1 digit(s) to the right of the |

  4 | 6
  6 | 5749
  8 | 4478
 10 | 979
 12 | 2359
 14 | 36566
 16 | 0787
 18 | 0
 20 | 3
 22 | 7
```

The plot for values of c_{0i} appears fairly symmetric. The location seems sufficiently above 0 to avoid having a normal distribution for these random variables place substantial probability on the negative line. The plot for values of c_{1i} may be skew to the right, and it is clear that a normal distribution here would result in unacceptable probability on the negative line, but a gamma distribution might be reasonable choice. Finally, the plot for values of a_i is highly spread out over large values (from about 46 to about 227) and not definitely skew in either direction. There is almost a hint of bimodal behavior in this plot, but with only 28 values we should resist the temptation to make too fine-grained a judgment. Lacking any strong guidance from this plot a normal distribution might be the default choice. So, at least as an

initial attempt, we model the random parameters c_{0i}, c_{1i} and a_i as follows.

$$\begin{aligned} c_{0i} &\sim iidN(\mu_0, \tau_0^2) \\ c_{1i} &\sim iidGa(\alpha_c, \beta_c) \\ a_i &\sim iidN(\mu_a, \tau_a^2) \end{aligned} \tag{13.8}$$

Although not impossible in principle to approach estimation and inference from a likelihood standpoint, which would require heavy doses of numerical integration to be successful, a Bayesian analysis would seem a natural choice with this model. To enact such an analysis we will need prior distributions on parameters of the distributions in (13.8) and on σ^2 in (13.7) as well. We have little to guide us in choice of these priors, and will use a combination of conditional conjugacy and naive uniform priors. If all of our priors are proper we do not need to face the daunting task of demonstrating posterior propriety.

$$\begin{aligned} \sigma^2 &\sim \text{Inverse Gamma}(\psi_1, \psi_2) \\ \mu_0 &\sim N(M_0, V_0) \\ \tau_0^2 &\sim \text{Uniform}(0, K_0) \\ \alpha_c &\sim \text{Uniform}(0, K_c) \\ \beta_c &\sim \text{Gamma}(\lambda, \gamma) \\ \mu_a &\sim N(M_a, V_a) \\ \tau_a^2 &\sim \text{Uniform}(0, K_a) \end{aligned} \tag{13.9}$$

A joint prior is then formed as a product form of the distributions listed immediately above.

Our model for a regional analysis then consists of the data structure in expression (13.7), the random parameter model (or mixing distributions) in expression (13.8) and the prior distributions in expression (13.9).

Mixture or Multi-level Prior

We will pause in the development of our analysis at this point to consider whether we should think of our model from the viewpoint of a Bayesian analysis of a mixture model, or analysis of a model with multiple priors. First, recall that, no matter what viewpoint we adopt, all of the mathematics remain the same and all of the same distributions will be available to us for inference. If there is a difference in these viewpoints it lies in which distributions we choose to make use of for inference, which in turn depends on how we are thinking about the various portions of the overall model. In particular, the issue of how (or even if we should) make inference about what I have called the random parameter model may be affected by our overall viewpoint as we approach the analysis.

The development and presentation of the model to this point in these notes has been conducted by viewing (13.7) as a mixture model. Other than σ^2, parameters of this model have been considered as random variables which were given distributions under an assumption of independence and identical distribution across nitrogen trials in expression (13.8). Prior distributions, which under this approach I assign to fixed but unknown (i.e., non-random) parameters to reflect belief or knowledge were used only in expression (13.9).

If instead, we would adopt the viewpoint of multi-level priors, we would have simply assumed that the c_{0i}, c_{1i} and a_i of the data model (13.7) are unknown exchangeable quantities and would have assigned what we believe or know about them the distributions in expression (13.8) skipping the random variable language of independence and identical distribution. These distributions could then be thought of as priors, often called level 1 priors, and the parameters of those distributions (which are also unknown quantities)

assigned additional level 2 priors as in expression (13.9). We might appeal to representation theorems as motivation or justification for formulating the priors c_{0i}, c_{1i} and a_i as mixture distributions.

Should we prefer one or the other of these viewpoints, or is this really simply much ado about nothing? I would suggest that the following aspects of the problem are relevant in consideration of this question.

1. Given the information about how corn obtains nitrogen for growth and development of grain, there exists a complex set of chemical, physical, and biological reactions and processes that determines how much usable nitrogen is available to the plants at various stages of growth and development. This, in turn, determines how much increase in yield is possible with the addition of nitrogen fertilizer. We can roll all of this complexity into what we think of as the scientific mechanism of interest – the phenomenon of how fertilization affects yield.

2. The mechanism we would like to model manifests itself in a unique way in each nitrogen trial conducted. No two trials will ever result in exactly the same expression of the mechanism.

3. If the way the mechanism is manifested in a given nitrogen trial is captured by the values of c_{0i}, c_{1i} and a_i in (13.7), this means that we will never have repeated values of the $Y_{i,j}$ from the same set of c_{0i}, c_{1i} and a_i.

The implication of these points is that the distributions in expression (13.8) represent not only what we might believe about c_{0i}, c_{1i} and a_i for an individual nitrogen trial, but they represent a statistical conceptualization (i.e., a model) of the scientific mechanism of concern. As such, they

should have interpretation under relative frequency probability – the relative frequencies with which the mechanism manifests itself in different ways across nitrogen trials. If the model is viewed as a *data generating mechanism*, these distributions are the central portion of our formulation of the problem. As the analysis progresses, then, we wish not so much to update our prior knowledge about c_{0i}, c_{1i} and a_i for the given trials in the data set (indexed by i), but rather to update our knowledge about the distributions of these quantities. That updating should take the form not only of computing posterior distributions for $\mu_0, \tau_0^2, \alpha_c, \beta_c, \mu_a$ and τ_a^2, but of updating knowledge about specification of the distributions in expression (13.8) in the first place, including distributional form and independence (or lack thereof). This all makes these distributions much more than simply level 1 priors.

Associated with this discussion is the question of whether or not it is fully appropriate to examine empirical distributions of surrogate observations of c_{0i}, c_{1i} and a_i to formulate distributions for them, as was done previously. Because we are considering these distributions part of the model rather than the prior we should be willing to examine whatever information is available in the data to help guide model selection. What statistician would eschew examination of a scatterplot in selecting an expectation function for regression analysis? The only difference here is that we am formulating a model for unobservable quantities. This makes examination of the data more difficult and makes the appropriateness of modeling choices less certain, leading directly back to the point of the previous paragraph that an important question will be what information is available from an analysis to provide feedback on the formulation of this critical portion of the model?

13.3.3 Analysis Through MCMC Methods

We now turn back to the more technical aspects of producing posterior distributions in this problem. The model would appear to lend itself to simulation of posteriors through an overall Gibbs sampling algorithm. To do this we need to derive full conditional distributions for all of the quantities involved in the model other than the observed values. In this section we will use the generic notation $p(x|z)$ to denote the density of a random quantity X given the value of another Z as $Z = z$, keeping in mind that if X and/or Z are fixed (non-random) quantities what we are really deriving are distributions of belief or knowledge about their values. We will also use the notation $p(x|\cdot)$ to mean the conditional density of X given all other quantities in the analysis. In addition, let $g_{i,j}$ denote the response function $g(x_{i,j}, \boldsymbol{\theta}_i)$ of expression (13.7) and note that these values contain the data model parameters c_{0i}, c_{1i} and a_i.

1. $p(c_{0i}|\cdot)$ for $i = 1, \ldots, T$

   The full conditional densities of the c_{0i}, for $i = 1, \ldots, T$ are,

$$\begin{aligned} p(c_{0i}|\cdot) &\propto p(c_{0i}|\mu_0, \tau_0^2)\, p(\boldsymbol{y}_i|c_{0i}, c_{1i}, a_i, \sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\tau_0^2}(c_{0i} - \mu_0)^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{i,j} - g_{i,j})^2\right\} \end{aligned} \tag{13.10}$$

   Now, with $g_{i,j}$ as in expression (13.7), define

$$t_{i,j} = \begin{cases} y_{i,j} - c_1\left\{\frac{3}{2}\left(\frac{x_j}{a}\right) - \frac{1}{2}\left(\frac{x_j}{a}\right)^3\right\} & 0 < x_j \leq a \\ y_{i,j} - c_1 & x_j > a \end{cases}$$

Then (13.10) may be written as,

$$p(c_{0i}|\cdot) \propto \exp\left\{-\frac{1}{2\tau_0^2}(c_{0i}-\mu_0)^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(t_{i,j}-c_{0i})^2\right\} \tag{13.11}$$

and the usual result for combining two normals gives that $p(c_{0i}|\cdot)$ is $N(A,B)$ with

$$\begin{aligned} A &= \frac{\sigma^2\mu_0 + \tau_0^2\sum_{j=1}^{n_i} t_{i,j}}{\sigma^2 + n_i\tau_0^2}, \\ B &= \frac{\sigma^2\tau_0^2}{\sigma^2 + n_i\tau_0^2} \end{aligned} \tag{13.12}$$

and simulation is straightforward using the `R` function `rnorm`.

2. $p(c_{1i}|\cdot)$ for $i = 1,\ldots,T$

$$\begin{aligned} p(c_{1i}|\cdot) &\propto p(c_{1i}|\alpha_c,\beta_c)\, p(\boldsymbol{y}_i|c_{0i},c_{1i},a_i,\sigma^2) \\ &\propto c_{1i}^{\alpha_c-1}\exp(-\beta_c c_{1i})\, \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{i,j}-g_{i,j})^2\right\} \end{aligned} \tag{13.13}$$

which does not simplify. We will simulate from $p(c_{1i}|\cdot)$ using an adaptive ratio of uniforms algorithm as described in the Stat 520 notes, although we could also use a Metropolis within Gibbs.

3. $p(a_i|\cdot)$ for $i = 1,\ldots,T$

$$\begin{aligned} p(a_i|\cdot) &\propto p(a_i|\mu_a, \tau_a^2)\, p(\boldsymbol{y}_i|c_{0i}, c_{1i}, a_i, \sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\tau_a^2}(a_i - \mu_a)^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{i,j} - g_{i,j})^2\right\}. \end{aligned} \tag{13.14}$$

Unlike c_{0i}, a_i does not factor out of $g_{i,j}$ and so we will also use an adaptive ratio of uniforms algorithm here.

4. $p(\sigma^2|\cdot)$

$$\begin{aligned} p(\sigma^2|\cdot) &\propto p(\sigma^2|\psi_1, \psi_2) \prod_{i=1}^{T} p(\boldsymbol{y}_i|c_{0i}, c_{1i}, a_i, \sigma^2) \\ &\propto \frac{\exp(-\psi_2/\sigma^2)}{(\sigma^2)^{\psi_1+1}} \frac{1}{(\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{T}\sum_{j=1}^{n_i}(y_{i,j} - g_{i,j})^2\right\}, \end{aligned} \tag{13.15}$$

where $N = \sum n_i$.

Here, let $S = \sum_{i=1}^{T}\sum_{j=1}^{n_i}(y_{i,j} - g_{i,j})^2$. Then (13.15) becomes,

$$p(\sigma^2|\cdot) \propto \frac{\exp\left\{-(\psi_2 + \frac{1}{2}S)\right\}/\sigma^2}{(\sigma^2)^{\psi_1+N/2+1}}, \tag{13.16}$$

which is Inverse Gamma $(\psi_1 + N/2,\ \psi_2 + S/2)$ and is easily simulated from by taking the reciprocal of simulated values from a gamma distribution with the same parameters.

5. $p(\mu_0|\cdot)$

Let $\boldsymbol{c}_0 = (c_{01}, c_{0,2} \ldots, c_{0,T})$.

$$\begin{aligned} p(\mu_0|\cdot) &\propto p(\mu_0|M_0, V_0)\, p(\boldsymbol{c}_0|\mu_0, \tau_0^2) \\ &\propto \exp\left\{-\frac{1}{2V_0}(\mu_0 - M_0)^2\right\} \exp\left\{-\frac{1}{2\tau_0^2}\sum_{i=1}^{T}(c_{0i} - \mu_0)^2\right\}, \end{aligned} \tag{13.17}$$

which, in the same manner as for the c_{0i} in item 1 is $N(A, B)$ where

$$\begin{aligned} A &= \frac{\tau_0^2 M_0 + V_0 \sum_{i=1}^{T} c_{0i}}{\tau_0^2 + TV_0}, \\ B &= \frac{\tau_0^2 V_0}{\tau_0^2 + TV_0} \end{aligned} \tag{13.18}$$

6. $p(\tau_0^2|\cdot)$

$$\begin{aligned} p(\tau_0^2|\cdot) &\propto p(\tau_0^2|K_0)\, p(\boldsymbol{c}_0|\mu_0, \tau_0^2) \\ &\propto I(\tau_0^2 < K_0)\, \frac{1}{(\tau_0^2)^{T/2}}\, \exp\left\{-\frac{1}{2\tau_0^2}\sum_{i=1}^{T}(c_{0i} - \mu_0)^2\right\}, \end{aligned} \tag{13.19}$$

which can be recognized as an Inverse Gamma $(T/2-1,\ (1/2)\sum(c_{0i} - \mu_0)^2)$ truncated at K_0.

7. $p(\alpha_c|\cdot)$

Let $\boldsymbol{c}_1 = (c_{11}, c_{12}, \ldots, c_{1T})$.

$$\begin{aligned} p(\alpha_c|\cdot) &\propto p(\alpha_c|K_c)\, p(\boldsymbol{c}_1|\alpha_c, \beta_c) \\ &\propto I(\alpha_c < K_c)\, \left\{\frac{\beta_c^{\alpha_c}}{\Gamma(\alpha_c)}\right\}^T \prod_{i=1}^{T} c_{1i}^{\alpha_c - 1}, \end{aligned} \tag{13.20}$$

which is another distribution we will simulate from using an adaptive ratio of uniforms algorithm.

8. $p(\beta_c|\cdot)$

$$\begin{aligned} p(\beta_c|\cdot) &\propto p(\beta_c|\lambda, \gamma)\, p(\boldsymbol{c}_1|\alpha_c, \beta_c) \\ &\propto \beta_c^{\lambda - 1} \exp(-\gamma\beta_c)\, \exp\left(-\beta_c \sum_{i=1}^{T} c_{1i}\right), \end{aligned} \tag{13.21}$$

which can be recognized as a Gamma $(\lambda + T\alpha_c,\ \gamma + \sum c_{1i})$ distribution.

9. $p(\mu_a|\cdot)$

   Let $\boldsymbol{a} = (a_1, a_2, \ldots, a_T)$.

$$\begin{aligned} p(\mu_a|\cdot) &\propto p(\mu_a|M_a, V_a)\, p(\boldsymbol{a}|\mu_a, \tau_a^2) \\ &\propto \exp\left\{-\frac{1}{2V_a}(\mu_a - M_a)^2\right\} \exp\left\{-\frac{1}{2\tau_a^2}\sum_{i=1}^{T}(a_i - \mu_a)^2\right\}, \end{aligned} \tag{13.22}$$

   which, in the same manner as for the c_{0i} in item 1 and μ_0 in item 5 is $N(A, B)$ where

$$\begin{aligned} A &= \frac{\tau_a^2 M_a + V_a \sum_{i=1}^{T} a_i}{\tau_a^2 + TV_a}, \\ B &= \frac{\tau_a^2 V_a}{\tau_a^2 + TV_a} \end{aligned} \tag{13.23}$$

10. $p(\tau_a^2|\cdot)$

    This is entirely parallel to $p(\tau_0^2|\cdot)$.

$$\begin{aligned} p(\tau_a^2|\cdot) &\propto p(\tau_a^2|K_a)\, p(\boldsymbol{a}|\mu_a, \tau_a^2) \\ &\propto I(\tau_a^2 < K_a)\, \frac{1}{(\tau_a^2)^{T/2}} \exp\left\{-\frac{1}{2\tau_a^2}\sum_{i=1}^{T}(a_i - \mu_a)^2\right\}, \end{aligned} \tag{13.24}$$

    which can be recognized as an Inverse Gamma $(T/2 - 1,\ (1/2)\sum(a_i - \mu_a)^2)$ truncated at K_a.

Two additional issues deserve brief discussion at this point. First, in order to use the full conditional distributions just derived in a Gibbs sampling

algorithm we need to select values of the parameters ψ_1 and ψ_2 (for σ^2), M_0 and V_0 (for μ_0), K_0 (for τ_0^2), K_c (for α_c), λ and γ (for β_c), M_a and V_a (for μ_a), and finally K_a (for τ_a^2). For the results that follow these values were chosen as $\psi_1 = 2$, $\psi_2 = 10$ (more on this later), $M_0 = 9$, $V_0 = 100$, $K_0 = 1000$, $K_c = 5000$, $\lambda = 5$, $\gamma = 1$, $M_a = 150$, $V_a = 100$ and $K_a = 10000$. These values were basically pulled from the sky with a rough idea of what to anticipate having been gained from previous dealings with the data. In contrast to using surrogate observations of the data model parameters to help formulate the random parameter model, there might be some legitimate concern if inspection of the data had been used to help select these prior parameters. The obvious way to deal with this concern is a sensitivity analysis, although that has not been done at the time these notes were written. Several different sets of values were used in a crude examination of how long a burn-in to choose for the Markov chain, but a systematic investigation remains to be completed. More interestingly, note that we must also have starting values for the c_{0i}, c_{1i} and a_i; $i = 1, \ldots, T$ in order to get the chain running. Initially, the generalized least squares estimates from the individual model fits were used, and an attempt was made to vary these by randomly permuting the values among data from different nitrogen trials. What occurred is that multiple Markov chains seemed to mix quite rapidly (by about 250 to 500 iterations depending on the parameter examined) when values of the prior parameters were varied but the data model parameters were left at their gls estimates. When values of data model parameters were shuffled, however, the chains never even got off the ground, in the sense that algorithms crashed due to the production of infinite values or the `R` NaN. The suggestion from this is that Gibbs algorithms that include simulation from full conditional distributions of data model parameters might be sensitive to starting values in

the same way that Newton-type algorithms for maximum likelihood or Gauss-Newton algorithms for generalized least squares can be. This is perhaps not so surprising for our current model, which involves a highly nonlinear response function. In the end, a burn-in of 2000 iterations was used and then the next 5000 values were collected for estimation of posterior distributions.

13.3.4 Posterior Distributions

In this section we will first discuss the issue of which distributions we would like to use for the purpose of inference, and will then examine those distributions based on output of the Gibbs algorithm described previously when it is applied to data from the 28 nitrogen trials.

Distributions for Inference

The Gibbs algorithm produces samples from the joint distribution,

$$p(c_{01}, \ldots, c_{0T}, c_{11}, \ldots, c_{1T}, a_1, \ldots, a_T, \sigma^2, \mu_0, \tau_0^2, \alpha_c, \beta_c, \mu_a, \tau_a^2 | \boldsymbol{y})$$

Typically, inferences about individual quantities are made from the marginal distributions which, of course, we also have samples from, although correlations are also easily examined. The marginal distributions that we could consider for the purpose of inference are:

1. For $i = 1, \ldots, T$
   - $p(c_{0i}|\boldsymbol{y})$ "posterior" of c_{0i} for nitrogen trial i
   - $p(c_{1i}|\boldsymbol{y})$ "posterior" of c_{1i} for nitrogen trial i
   - $p(a_i|\boldsymbol{y})$ "posterior" of a_i for nitrogen trial i
2. $p(\sigma^2|\boldsymbol{y})$, the posterior of the data model variance σ^2.

3. $p(\mu_0|\boldsymbol{y})$ and $p(\tau_0^2|\boldsymbol{y})$, the posteriors of μ_0 and τ_0^2 that control the model distribution of the c_{0i}.

4. $p(\alpha_c|\boldsymbol{y})$ and $p(\beta_c|\boldsymbol{y})$, the posteriors of α_c and β_c that control the model distribution of the c_{1i}.

5. $p(\mu_a|\boldsymbol{y})$ and $p(\tau_a^2|\boldsymbol{y})$, the posteriors of μ_a and τ_a^2 that control the model distribution of the a_i.

I put "posterior" in quotes for the distributions in the first item above because, although I don't really have a problem with calling these posterior distributions (which just means a conditional distribution of an unobservable given the observed data), we are not thinking of these distributions in the same way as the other posteriors. Inference about the parameters σ^2, μ_0, τ_0^2, α_c, β_c, μ_a, and τ_a^2 is not controversial. Of particular interest in this problem is the value of μ_a, the expected value of the lowest nitrogen rate that produces maximum yield. But what use is to be made, if any, of the distributions $p(c_{0i}|\boldsymbol{y})$, $p(c_{1i}|\boldsymbol{y})$ and $p(a_i|\boldsymbol{y})$?

It has been argued previously that the value of the data model parameters c_{0i}, c_{1i} and a_i for a particular nitrogen trial, $i = 5$ say, are of little interest. It is the distribution of these values across trials that should be the focus of attention in this problem. We do not have information about these distributions directly in the form of a posterior. We do have collections or ensembles of the conditional distributions of the data model parameters given the observed data (as in item 1 above). Can't we use these for the purpose of inference about distributions of the data model parameters? I believe the answer is no, at least not as they stand. How do we make inference on the basis of any collection of values, including collections of data values? We embed them in a larger structure, typically a distribution, and make inference

about that distribution. Similarly, we have simply a collection of values from each of $p(c_{0i}|\boldsymbol{y})$, $p(c_{1i}|\boldsymbol{y})$ and $p(a_i|\boldsymbol{y})$; $i = 1, \ldots, T$. In order to make these useful for inference they need to be embedded in a larger structure. But what that structure might be is not clear. I will suggest that inference about the distributions of data model parameters is better based on predictive distributions than on individual posteriors such as these. Nevertheless, these distributions can be used in model assessment. In model formulation, we made use of surrogate observations of the data model parameters c_{0i}, c_{1i} and a_i, which were taken to be generalized least squares estimates from fitting models to data from individual nitrogen trials. We could consider these as individual surrogate values because they were produced from data for each nitrogen trial without any consideration of the set of trials as a whole. We now also have (estimated) expected values of c_{0i}, c_{1i} and a_i for each trial indexed by i. We might then consider these as simultaneous surrogate values because they were produced using data from the entire set of nitrogen trials. The simultaneous surrogate values may have a number of uses, including a comparison with the individual surrogate values as an indication of the degree to which the hierarchical model analysis has borrowed strength across the entire data set.

The posteriors listed in items 2,3, 4, and 5 above give information about the parameters of model distributions for c_{0i}, c_{1i} and a_i under an assumption that the model is adequate, that is, that the specification of the distributions in expression (**??**) is adequate. But they don't necessarily allow us to directly answer the questions of primary interest, such as "in a new trial what is the probability that the maximum yield will be achieved with 120 kg/ha or less of nitrogen fertilizer?". What are the choices for ways to address this and similar questions?

1. Use the model distribution for a_i evaluated at posterior expectations of its parameters. In the model we took $a_i \sim iidN(\mu_a, \tau_a^2)$, so the probability desired would be taken from $N(\hat{\mu}_a, \hat{\tau}_a^2)$, where $\hat{\mu}_a$ is the mean of $p(\mu_a|\boldsymbol{y})$ and $\hat{\tau}_a^2$ is the mean of $p(\tau_a^2|\boldsymbol{y})$. This would be similar to using maximum likelihood estimates in a modeled distribution to estimate quantiles. When we use likelihood estimates in this way, however, we typically apply the delta method to obtain approximate standard errors for quantiles. The analogous procedure with a posterior would be to compute quantiles from the modeled distribution using each of a large number of samples from the posterior of parameter values which would result in essentially a posterior distribution for quantiles of a distribution in the model. I don't know that I have ever seen this done, but it might be interesting to attempt.

2. Simulate values from posterior predictive distributions of c_{0i}, c_{1i} and a_i. In a simple model consisting of $f(y|\theta)$ with prior $\pi(\theta)$ we determine the posterior $p(\theta|y)$ and define the posterior predictive distribution for a new value y^* say as $p(y^*|y) = \int f(y^*|\theta)p(\theta|y)d\theta$. Here, posterior predictive distributions for the data model parameters from an as yet unobserved trial indexed by k would be

$$\begin{aligned} p(c_{0k}|\boldsymbol{y}) &= \int_0^\infty p(c_{0k}|\mu_0, \tau_0^2)p(\mu_0, \tau_0^2|\boldsymbol{y})\, d\mu_0\, d\tau_0^2 \\ p(c_{1k}|\boldsymbol{y}) &= \int_0^\infty p(c_{1k}|\alpha_c, beta_c)p(\alpha_c, \beta_c|\boldsymbol{y})\, d\alpha_c\, d\beta_c \\ p(a_k|\boldsymbol{y}) &= \int_0^\infty p(a_k|\mu_a, \tau_a^2)p(\mu_a, \tau_a^2|\boldsymbol{y})\, d\mu_a\, d\tau_a^2 \end{aligned} \tag{13.25}$$

   To simulate values from these distributions we take a large number of samples from the posteriors (which we have during a run of the Gibbs algorithm) and further simulate an additional value of c_{0i}, c_{1i} and a_i

for each from their model distributions with parameters given by the posterior sample.

It can be argued that this is the most straightforward and clearly justifiable method to produce distributions from which probability statements can be made about what we believe will occur in future trials (or, by extension, on entire farms). In addition, it lends itself readily to the production of posterior predictive distributions for other quantities of interest. Recall that some current fertilization recommendations make use of the concept of economic optimum nitrogen rate, the nitrogen rate above which additional fertilization will fail to pay for itself in increased yield. For a spherical response function as used in our hierarchical model, this quantity can be computed (according to information obtained from the agronomists) as,

$$W = \begin{cases} \left[\frac{1.5c_{1i}a_i^2 - a_i^3/R}{1.5c_{1i}}\right]^{1/2} & R > \frac{a_i}{1.5c_{1i}} \\ 0 & R \leq \frac{a_i}{1.5c_{1i}} \end{cases}, \tag{13.26}$$

where R is a ratio of corn price to nitrogen price. This ratio is often reported in terms of pounds of nitrogen to bushels of corn, and I have seen a value of 10 used in examples. Using 10 in expression (13.26) supposes that the model has been fit to data recorded in pounds of nitrogen per acre and bushels of corn per acre. Our data is in kg/ha for both nitrogen and yield, but we divided yield by 1000 prior to analysis. A little investigation suggests there should be about 70lbs of corn in a bushel. So, for (13.26) to work with our estimates we should multiply 10 by 1000/70 or $R \approx 140$. To produce a posterior predictive distribution for W we simply compute its value for each sample of c_{0i}, c_{1i} and a_i from their respective posterior predictive distributions.

The upshot of this subsection is that, while basic posterior distributions for σ^2, μ_0, τ_0^2, α_c, β_c, μ_a and τ_a^2 are certainly useful for inference about values of those parameters, it is predictive statements about quantities such as economic optimal nitrogen rate and lowest nitrogen rate to achieve maximum yield that are of the greatest interest.

Basic Output from Gibbs Sampling

Summary quantities for the posterior distributions of the fixed parameters in our model, σ^2, μ_0, τ_0^2, α_c, β_c, μ_a and τ_a^2 are presented in Table 15.1. Histograms of the full posterior distributions are presented in Figures 13.3 through 13.6.

Parameter	Mean	Std. Dev.	95% Interval
σ^2	0.345	0.0501	$(0.260, 0.456)$
μ_0	9.110	0.3200	$(8.484, 9.734)$
τ_0^2	2.649	0.9085	$(1.378, 4.905)$
α_c	7.524	2.3227	$(3.966, 13.354)$
β_c	2.460	0.7449	$(1.313, 4.264)$
μ_a	144.980	7.5302	$(130.546, 159.898)$
τ_a^2	890.963	842.1100	$(62.484, 3094.812)$

Table 13.1: Summary values from posterior distributions – based on 5000 samples

Examination of these figures indicates that the marginal posteriors of μ_0 and μ_a would be well approximated by normal distributions, the marginal posteriors of τ_0^2 and τ_a^2 by inverse gamma distributions, and the marginal posteriors of both α_c and β_c might be approximated by gamma distributions.

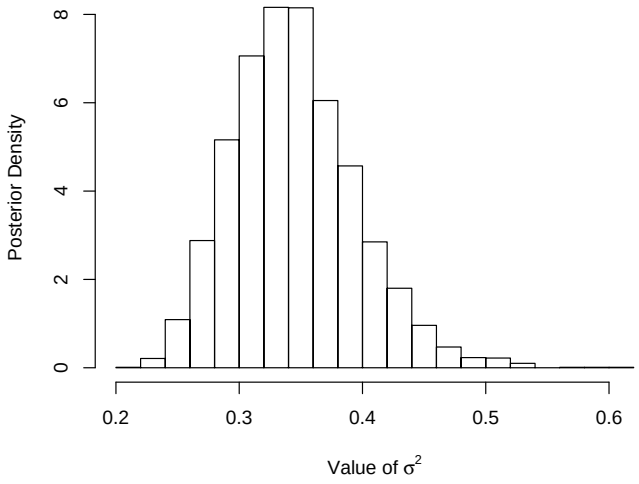

Figure 13.3: Posterior distribution of data model variance σ^2 based on 5000 samples.

This suggests that our choice of priors was pretty much in concert with indications from the data, although we have gained a good deal of knowledge about the parameters that had been assigned uniform priors (τ_0^2, τ_a^2 and α_c). If we were now faced with data from a new set of nitrogen trials, we would most likely feel pretty comfortable with normal priors for μ_0 and μ_a with means given by Table 13.1 and variances a modest multiple (2 to 5) of the values in Table 13.1. We might well assign all three variances, σ^2, τ_0^2 and τ_a^2 inverse gamma priors with parameters that would give densities to approximate the histograms of Figures 13.3, 13.4 and 13.6, and gamma priors for both α_c and β_c chosen in a similar manner. The most difficult of these approximations to accomplish would be for τ_a^2 which has a large posterior expectation and variance; developing a new prior for τ_a^2 might be a bit more

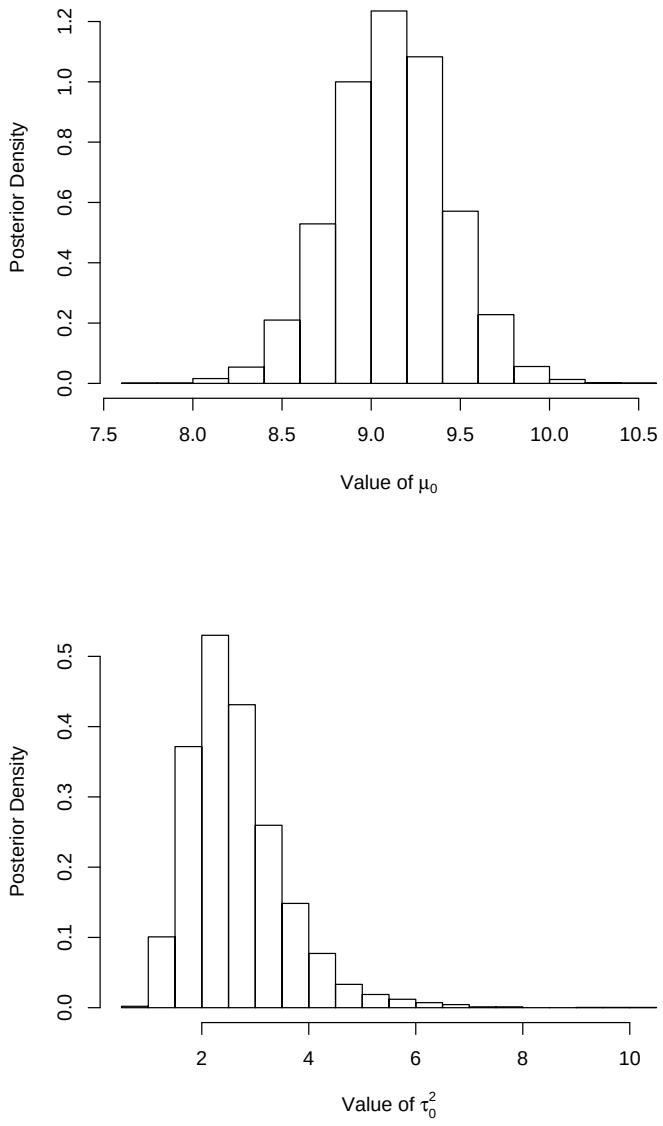

Figure 13.4: Posterior distributions of μ_0 and τ_0^2 – parameters of the distribution of data model c_{0i}

.

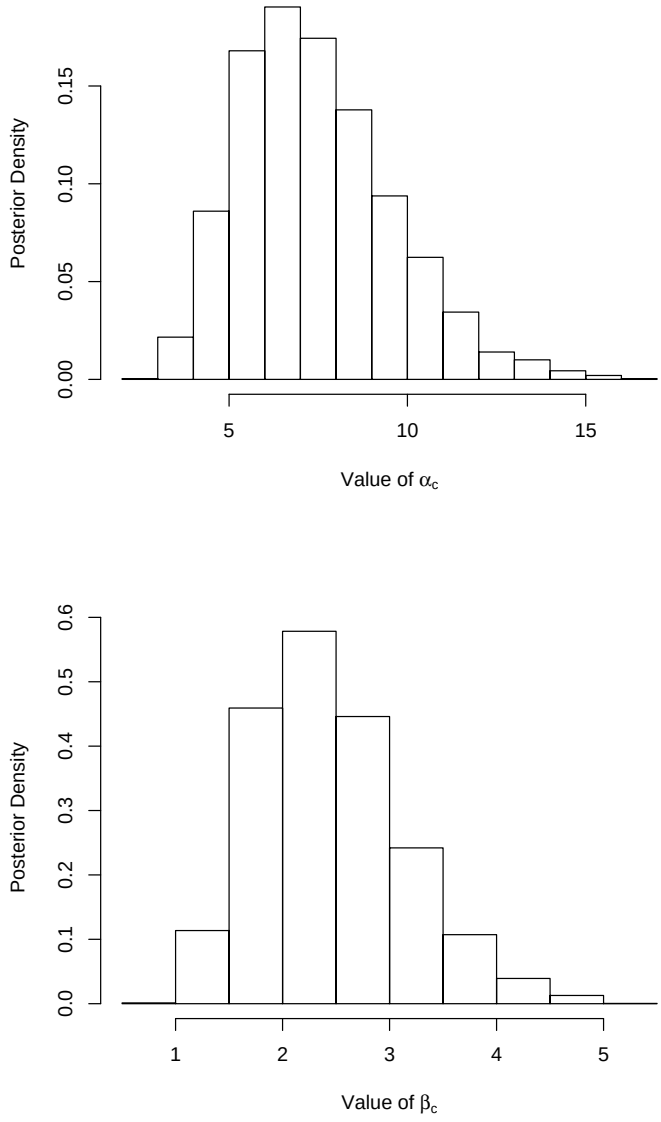

Figure 13.5: Posterior distributions of α_c and β_c – parameters of the distribution of data model c_{1i}

.

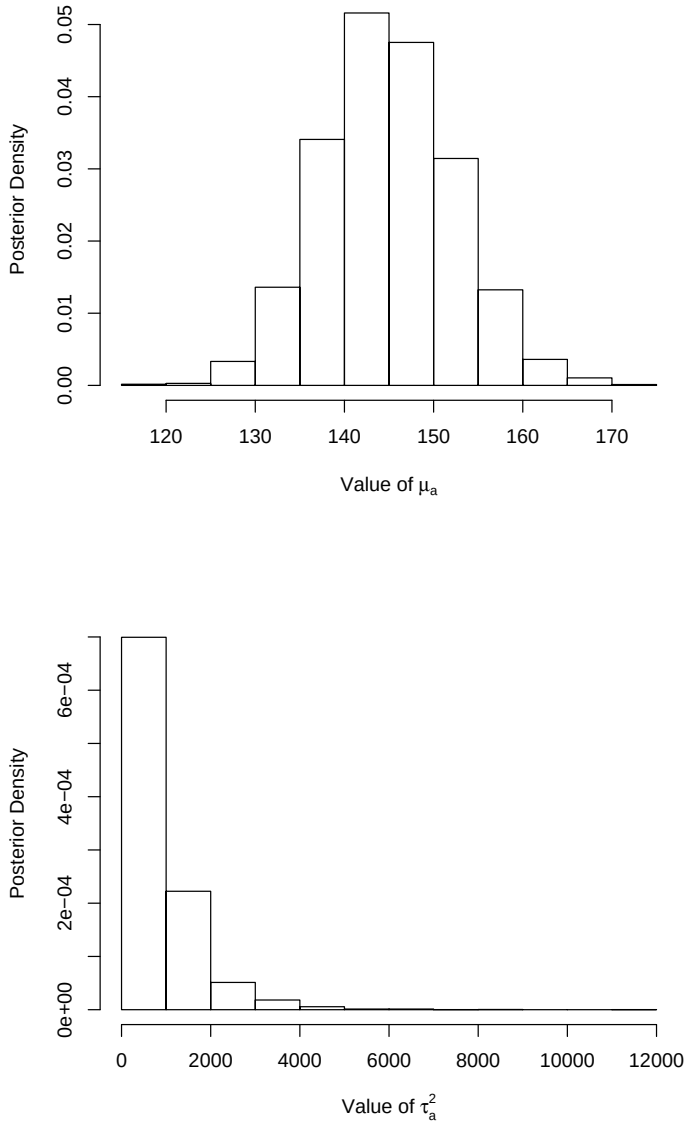

Figure 13.6: Posterior distributions of μ_a and τ_a^2 – parameters of the distribution of data model a_i

.

of a challenge than the for the other parameters.

Posterior Predictive Distributions

Now we can examine several posterior predictive distributions for the purpose of predictive inference. In particular, agronomists would like to be able to predict the lowest level of fertilizer that will produce the maximum yield, as well as the economic optimal nitrogen rate as defined in (13.26). The former is simply the posterior predictive distribution $p(a_k|\boldsymbol{y})$, while the latter is the posterior predictive distribution of W. Fixing the ratio of corn to nitrogen price at $R = 140$, samples of size 5000 were taken from these posterior predictive distributions. The distribution for lowest nitrogen fertilizer rate that results in maximum yield is presented in Figure 13.7, while that for economic optimal nitrogen rate is presented in Figure 13.8.

Probabilities can be computed directly from these posterior predictive distributions. For example, the probability that maximum yield occurs at 150 kg/ha of nitrogen or less is predicted as 0.61, while the probability that the economic optimal nitrogen rate is 150 kg/ha or less is 0.89. In fact, we can easily produce distribution functions for both lowest nitrogen at maximum yield and economic optimal nitrogen, which is displayed in Figure 13.9.

The nitrogen rate at which the greatest absolute difference between the distribution functions occurs is 136.4 kg/ha and this is shown as the vertical line in Figure 15.9. The probability that lowest nitrogen to achieve maximum yield is less than 136.4 is 0.358, or about 36%, while the probability that the economic optimal nitrogen rate is less than 136.4 is 0.745 or about 75%, a difference of 39%.

Although this analysis is an example, based on only a subset of nitrogen

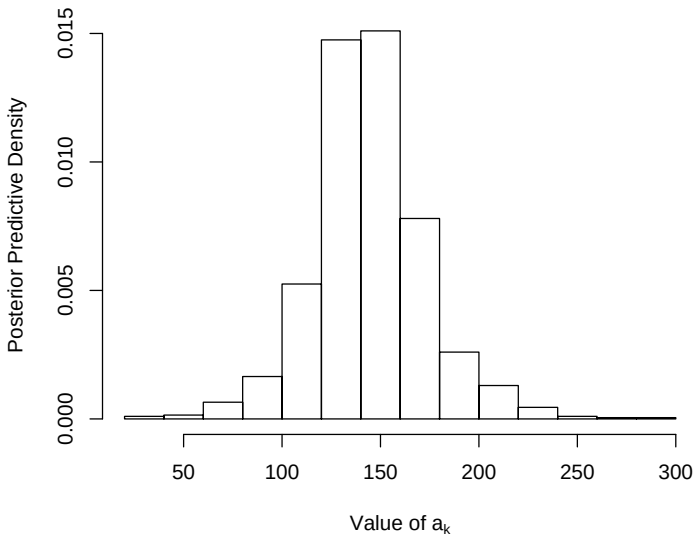

Figure 13.7: Posterior predictive distribution of lowest nitrogen rate that results in maximum yield, based on 5000 samples.

trials from Iowa alone, the results would seem to have some implications for communicating risks connected with fertilization. Achieving maximum yield has long been a goal of producers. According to the results summarized in Figure 13.9, 192.6 kg/ha of nitrogen fertilizer would result in a 95% probability of reaching this goal. But 192.6 kg/ha of fertilizer would be over 30 kg/ha more than is needed to reach a 95% probability of achieving the economic optimal rate, and makes the probability of having exceeded the optimal rate 99.5%.

Other summaries are available from these samples of the posterior predictive distributions. If one applies a fixed nitrogen rate, what are the probabilities it exceeds the economic optimal rate by various amounts such as 20

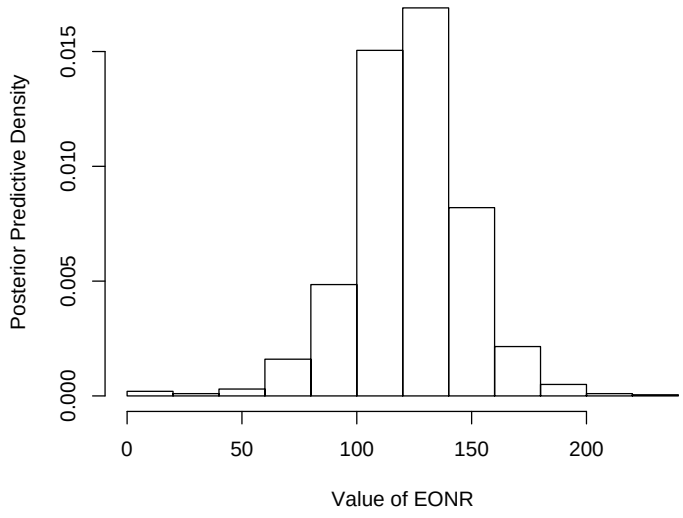

Figure 13.8: Posterior predictive distribution of economic optimal nitrogen rate, based on 5000 samples.

or 30 kg/ha? What are the probabilities it is less than the amount needed to first reach maximum yield by these same values? Graphs are presented in Figure 15.10 that answer these questions for four levels of nitrogen fertilization, which correspond to the 25−percentile, 50−percentile, 75−percentile and 95−percentile of the predictive distribution of nitrogen to reach maximum yield as shown in Figure 13.7.

Is there a nitrogen rate at which the probabilities of using less than needed to reach maximum yield (by any amount) and using more than the economic optimal rate (again, by any amount) are the same? This question is answered by the plot of Figure 13.11. Numerically, this amount is between 133 and 134 kg/ha and the two probabilities are both about 70%. A word of caution

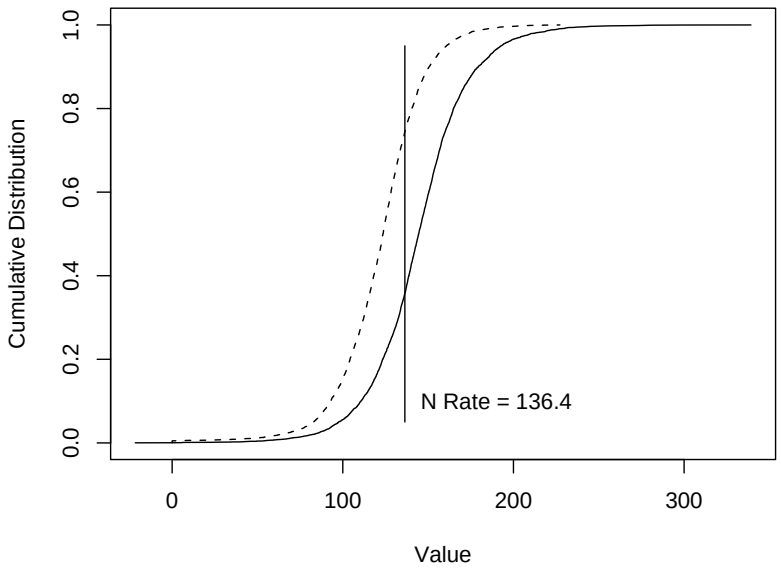

Figure 13.9: Distribution functions for nitrogen at maximum yield (solid curve) and economic optimal nitrogen rate (dashed curve), based 5000 samples.

is in order here. I am not suggesting that this plot be interpreted in the same way as those supply-and-demand graphs I remember from my introductory economics course. The economic loss of fertilizing more than the economic optimal rate is not the same as the economic loss of fertilizing less than just enough to produce maximum yield. But I am suggesting that if agronomists and/or economists have enough knowledge to adjust the probabilities of Figure 13.10 to represent probabilities of equal economic risk then we would be able to produce such a plot using samples from the posterior predictive distribution.

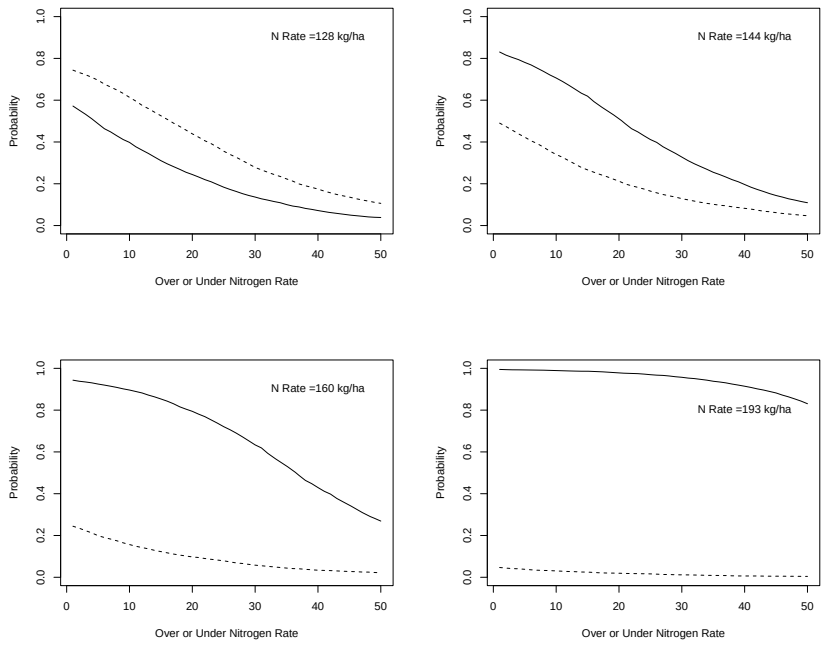

Figure 13.10: Probabilities that nitrogen at four rates are various amounts given on the horizontal axis over the economic optimal rate (as solid curve) and under rate at maximum yield (as dashed curve). Nitrogen rates shown are 25, 50, 75, and 95 percentiles of posterior predictive distribution of nitrogen needed for maximum yield.

13.3.5 Assessing the Model

There are any number of procedures that we might use to assess the overall model. Here, we will focus on an attempt to assess the random parameter model of expression (**??**) since this is arguably the most important portion of the model for the problem. This assessment is difficult because we do not observe the random data model parameters c_{0i}, c_{1i} and a_i. Thus, a straightforward use of posterior predictive model assessment is not possible. We

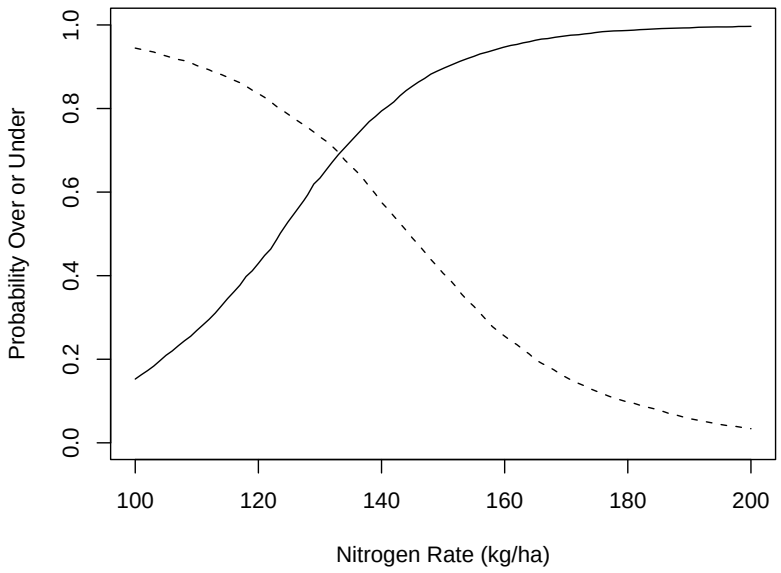

Figure 13.11: Probabilities of exceeding economic optimal (solid curve) or failing to reach maximum yield (dashed curve) for various nitrogen rates.

do, however, have quantities we have already used as surrogate observations for the data model parameters, those being generalized least squares estimates for models fit to data from individual nitrogen trials. It might also be possible to examine what I have suggested could be used as simultaneous surrogate values, the means of the posterior distributions for the individual c_{0i}, c_{1i} and a_i. In this section we will make use of the individual surrogate values we examined in the stem-and-leaf plots previously. We have $T = 28$ surrogate values of each of the c_{0i}, c_{1i}, and a_i. We can compare these values to values simulated from the posterior predictive distributions to address two questions that go a long way toward determining whether our random pa-

rameter model is adequate. First we will examine the marginal distributions of each of these three parameters. While having the marginals right does not ensure having the joint right, certainly we cannot have a correct joint with inadequate marginals. Secondly, we can examine correlation in the simulated values with the "observed" correlation to see if sufficient dependence is being produced in the joint posterior predictive.

To examine marginal distributions we will make use of Kolmogorov-Smirnov statistics for comparison of simulated and surrogate "data sets" in the following way. Let c_{0n}, c_{1n} and a_n represent new values of the three data model parameters, and let $\theta = \{(c_{0n}, c_{1n}, a_n) : n = 1, \ldots, T\}$ denote a set of these values of the same size as the surrogate values. Also, let $\hat{\theta} = \{(\hat{c}_{0i}, \hat{c}_{1i}, \hat{a}_i) : i = 1, \ldots, T\}$ denote the set of surrogate values. The following steps are then conducted.

1. Simulate M sets of T values each from $p(c_{0n}, c_{1n}, a_n|\boldsymbol{y})$,

$$\theta^*_m = \{(c_{0n}, c_{1n}, a_n) : n = 1, \ldots, T\}; \quad m = 1, \ldots, M$$

2. Simulate an additional R sets of T values each from $p(c_{0n}, c_{1n}, a_n|\boldsymbol{y})$,

$$\theta^*_r = \{(c_{0n}, c_{1n}, a_n) : n = 1, \ldots, T\}; \quad r = 1, \ldots, R \tag{13.27}$$

3. Compute Kolmogorov-Smirnov statistics $K(\hat{\theta}, \theta^*_m)$ from empirical distribution functions of any element ($c_0, c_1,$ or a) of $\hat{\theta}$ and θ^*_m for $m = 1, \ldots, M$

4. Compute average

$$K^{act} = \frac{1}{M} \sum_{m=1}^{M} K(\hat{\theta}, \theta^*_m)$$

5. Compute Kolmogorov-Smirnov statistics $K(\theta_m^*, \theta_r^*)$ from empirical distribution functions of any element of θ_m^* and θ_r^* for $m = 1, \ldots, M$ and $r = 1, \ldots, R$

6. Compute averages for $r = 1, \ldots, R$

$$K_r^{sim} = \frac{1}{M} \sum_{m=1}^{M} K(\theta_m^*, \theta_r^*)$$

7. Compute what is basically a posterior predictive p-value as,

$$P^* = \sum_{r=1}^{R} I(K^{act} \leq K_r^{sim})$$

A similar process can be used to compare correlations as follows.

1. Compute correlations C^{act} among elements of $\hat{\theta}$.

2. Compute correlations C_r^{sim} among elements of θ_r^*; $r = 1, \ldots, R$, where these sets are the same as those in step two of the previous procedure.

3. Compute p-value for correlation between any two elements (c_0, c_1, a) as,

$$P^* = \sum_{r=1}^{R} I\left(C^{act} \leq C_r^{sim}\right)$$

The two procedures outlined above were applied to the nitrogen trials using $M = 1000$ and $R = 1000$, with the following results.

These values would imply that we are, indeed, modeling the marginal distributions of the three data model parameters in a reasonable fashion. We may not be doing quite as well with the actual joint distribution, however. Our model could most likely be improved by including dependence terms in the random parameter model of expression (**??**), particularly between

Comparison	Case	p-value
Marginal Distribution	c_0	0.942
	c_1	0.777
	a	0.271
Correlation	c_0 and c_1	0.029
	c_0 and a	0.314
	c_1 and a	0.20

Table 13.2: Posterior predictive p-values for assessment of random parameter model.

c_{0i} and c_{1i}. The "observed" correlation between surrogate values of these quantities was -0.5169. The average correlation between these quantities in the $R = 1000$ sets of values simulated in step two of the procedure for assessing marginal distributions (and used in the procedure for assessing correlations) was -0.0268.

Chapter 14

Bayes Factors

Bayes factors are a general structure that can be used to assist in testing hypotheses about data model parameter values and in model selection. They appear in several forms, which contributes to both their flexible nature and also to some controversies regarding their use. Our goal in this chapter is to untangle some of the issues connected with Bayes factors and their uses. We will consider Bayes Factors within the setting of a standard Bayesian problem where the data model is $f(\boldsymbol{y}|\boldsymbol{\theta})$ with support $y \in \Omega$ and parameter space $\boldsymbol{\theta} \in \Theta$, the prior is $\pi(\boldsymbol{\theta})$ with support $\boldsymbol{\theta} \in \Theta$, and these lead to a proper posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ also with support $\boldsymbol{\theta} \in \Theta$.

Before considering various situations in which a Bayes factor might prove useful, we offer a general definition. Central to the use of Bayes factors is the notion of a *model*, and a Bayes factor typically is used to compare support in the data for two models, M_1 and M_2. Models may correspond to different portions of the parameter space for the data model parameter θ, different prior specifications, or even different data models, depending on the specific situation under consideration. Leaving the definition of a model somewhat

vague for the time being, the Bayes factor in favor of model M_1 over model M_2 can be defined as follows.

Definition:

Given two models, M_1 and M_2 the Bayes factor (BF) in favor of model M_1 over model M_2 is,

$$BF(M_1, M_2) = \frac{Pr(M_1|\boldsymbol{y})}{Pr(M_2|\boldsymbol{y})} / \frac{Pr(M_1)}{Pr(M_2)}. \tag{14.1}$$

That is, a BF is the mathematical factor that relates the prior odds in favor of model M_1, $Pr(M_1)/Pr(M_2)$ to the posterior odds in favor of model M_1, $Pr(M_1|\boldsymbol{y})/Pr(M_2|\boldsymbol{y})$. Notice that,

$$\frac{Pr(M_1|\boldsymbol{y})}{Pr(M_2|\boldsymbol{y})} = \frac{Pr(\boldsymbol{y}|M_1)\, Pr(M_1)}{Pr(\boldsymbol{y}|M_2)\, Pr(M_2)},$$

which indicates that we may also take the Bayes factor in favor of M_1 to be,

$$BF(M_1, M_2) = \frac{Pr(\boldsymbol{y}|M_1)}{Pr(\boldsymbol{y}|M_2)}, \tag{14.2}$$

and, if the ratio $Pr(\boldsymbol{y}|M_1)/Pr(\boldsymbol{y}|M_2)$ is equal to a ratio of density functions (more on this in what follows), then we can compute a Bayes factor as,

$$BF(M_1, M_2) = \frac{h_1(\boldsymbol{y}|M_1)}{h_2(\boldsymbol{y}|M_2)} = \frac{\int f_1(\boldsymbol{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\nu(\boldsymbol{\theta}_1)}{\int f_2(\boldsymbol{y}|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)d\nu(\boldsymbol{\theta}_2)}, \tag{14.3}$$

where ν is either Lebesgue or counting measure. When it is possible to compute $Pr(M_i|\boldsymbol{y})$ directly from the posterior distribution, then (14.1) usually provides the easiest way to compute a Bayes Factor. In other cases, (14.3) is more convenient, namely when it is possible to obtain the marginal data densities more easily than to determine the posterior model probabilities. But note that given a Bayes Factor and prior model probabilities, it is always possible to determine the posterior odds for model M_1 over M_2.

A note on what are called odds is in order here. What are called the *odds* of an event E are generally taken to be $O(E) = Pr(E)/Pr(E^c)$, the

probability that the event occurs divided by the probability that it does not occur. This will be true for us in cases for which the two models M_1 and M_2 constitute the entire universe of possibilities under consideration. We may also, however, call the ratio of probabilities of two events E_1 and E_2 an odds even if $E_2 \neq E_1^c$, if we are careful to identify the position of both events, in which case $O(E_1, E_2) = Pr(E_1)/Pr(E_2)$ is the odds of event E_1 relative to event E_2, or the odds in favor of event E_1 over event E_2. Note, however, that this use of the word odds changes the usual definition used by many statisticians.

14.1 A Binomial Example

We will use the following example to illustrate many of the various situations in which a Bayes factor may prove useful. The problem deals with the sex ratio at birth in a South American ungulate. Guanacos, along with Llamas, Vicunas, and Alpacas, are members of the camel family that live in the Andean mountains of South America, from Peru and Bolivia south to Patagonia and Tierra del Fuego in Chile and Argentina. We will be concerned with one population of guanacos that live in Torres del Paine National Park located in the eastern portion of Chili.

Guanacos live in family groups with one dominant male and several females and their offspring. Breeding female guanacos give birth to one offspring per season. There is a hypothesis that the probability of a male birth is greater than the probability of a female birth and this might be related to a higher mortality rate in the first year of life for males than for females. If this is true, it has nothing to do with somehow *balancing the population* or being for the *good of the species*, it would have to result from an adult

female guanaco being able to pass along more of her genes if her offspring is male than female because males are in shorter supply, but counter-balanced with the possibility that a male offspring also has a lower chance of surviving until it can breed and pass along any genes at all. And, there must also be a genetic link to the physiological processes involved in implantation, development, and successful birth that operates in a differential manner for male and female embryos. This latter is a separate topic in evolutionary biology that is not related to the analysis of our data, although we note that in many systems such physiological mechanisms are believed to exist.

Relative to the guanacos of Torres del Paine National Park, as part of a larger study on mortality in juvenal guanacos, newborn guanacos were radio-collared and monitored to assess survival and causes of mortality (Sarno *et al.*, 1999). The study covered five years, but sampling effort was substantially reduced in the final two years. We will use data from the first three years of the study, 1991, 1992, and 1993 in this example. In 1991 the researchers recorded 54 male and 45 female newborn. In 1992 there were 50 male and 43 female newborns and in 1993 there were 45 males and 49 females. While the exact sampling plan by which newborns were obtained is not detailed in Sarno *et al.*, we will presume it yields values representative of the population. If we sum values across years, we have 149 males and 137 females, giving an observed proportion of male birth of 0.521. Behl (1992) reported that survival of juvenal male guanacos was consistently 10% lower than females in this population, although that could not be verified by Sarno *et al.* (1999). If, based on the results of Behl (1992), we would take the probability of a male birth to be 1.1 times the probability of a female birth, the proportion of male births would be 0.524.

The model for our data is a single binomial distribution with parameter

θ representing the probability of a male birth. Without giving the matter any thought, we might be tempted to put a uniform prior on θ, but that would give equal prior probability to the event that the chances of a male birth are between 1% and 11% and the event that the chances are between 45% and 55%. This would be so biologically unrealistic as to be laughable. A more scientifically serious prior would be a beta distribution with parameters $\alpha_0 = 35$ and $\beta_0 = 35$, which has expected value 0.50 and puts roughly 90% probability on the interval $(0.40, 0.60)$. The resultant posterior in our problem is then a beta distribution with parameters $\alpha = 184$ and $\beta = 172$.

14.2 Hypothesis Tests about Parameter Values

Consider testing hypotheses about the data model parameter $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \in \Theta_1$. A distinction is often drawn between *simple* and *composite* hypotheses. The hypothesis H_0 is a simple hypothesis if Θ_0 consists of a single point in Θ and is a composite hypothesis if Θ_0 consists of multiple points in Θ, usually an interval in $\mathbb{R}^1$ for a scalar parameter or a region in $\mathbb{R}^p$ for a vector-valued parameter. Similar definitions of simple and composite apply to H_1 and Θ_1. While the distinction between simple and composite hypotheses are most likely familiar, an important distinction that is not always explicitly considered is between sets of hypotheses (H_0 and H_1) that form a *logical disjunction* and sets of hypotheses that do not. Two hypotheses H_0 and H_1 form a logical disjunction if Θ_0 and Θ_1 partition the parameter space Θ. That is, Θ_0 and Θ_1 are disjoint (or mutually exclusive) and $\Theta_0 \cup \Theta_1 = \Theta$. The distinction between pairs of hypotheses that form a logical disjunction

and pairs of hypotheses that do not will be important to us in considering situations in which Bayes factors are uncontroversial and situations in which there may be some controversy concerning their use.

To make use of Bayes factors in testing hypotheses about possible values of data model parameters we identify models directly in terms of the portions of the parameter space dictated by the hypotheses under consideration. In the Neyman-Pearson framework of hypothesis testing we generally identify H_1 with a hypothesis that we are hoping can be accepted over the hypothesis H_0. In (14.1) we have defined the Bayes factor in favor of model M_1 over model M_2. For convenience, then, let M_1 be equivalent to H_1 and let M_2 be equivalent to H_0. Scales against which to judge the value of a Bayes factor have been suggested by several authors. Jeffreys (1961) suggested a Bayes factor less than about 3.2 provides poor evidence in favor of M_1 over M_2, a value between 3.2 and 10 provides substantial evidence in favor of M_1, a value between 10 and 100 provides strong evidence and a value greater than 100 provides decisive evidence in favor of M_1. Kass and Raftery (1995) suggest categories of 3 to 20, 20 to 150 and greater than 150 corresponding to some, strong, and decisive evidence in favor of M_1. We must recognize that these scales are arbitrary in the same sense that the traditional significance levels of 0.10, 0.05 and 0.01 are arbitrary in the assessment of frequentist $p-$values.

14.2.1 Two Disjuctive Composite Hypotheses

In the example of sex ratio at birth in guanacos, we might start by considering the models $M_1 : \theta > 0.5$ versus $M_2 : \theta < 0.5$. Given a beta prior with parameters 35 and 35, prior probabilities for these models are both 0.50 so the prior odds in favor of model M_1 are $Pr(M_1)/Pr(M_2) = 1.0$. From

(14.1), the Bayes Factor for this example is then equal to the posterior odds $Pr(M_1|\boldsymbol{y})/Pr(M_2|\boldsymbol{y})$ which is 2.815, providing perhaps a bit, but certainly not much, evidence in favor of M_1 that $\theta > 0.50$.

14.2.2 Unequal Prior Probabilities

In our test with the guanaco data, if we would change the hypotheses to be $M_1 : \theta > 0.524$ versus $M_2 : \theta < 0.524$, we would no longer have prior odds in favor of M_1 equal to 1.0. But we may still obtain the Bayes Factor in favor of model M_1 using expression (14.1). The prior probability of model M_1 is obtained from the beta (35, 35) prior as $Pr(M_1) = 0.3444$ and that for model M_2 as $Pr(M_2) = 0.6556$. The posterior probability of the models are obtained from the integrals,

$$\begin{aligned} Pr(M_1|y) &= \frac{\Gamma(\alpha_0+\beta_0+n)}{\Gamma(\alpha_0+y)\,\Gamma(\beta_0+n-y)} \int_{0.524}^{1} \theta^{\alpha_0+y-1}(1-\theta)^{\beta_0+n-y-1}\,d\theta \\ Pr(M_2|y) &= \frac{\Gamma(\alpha_0+\beta_0+n)}{\Gamma(\alpha_0+y)\,\Gamma(\beta_0+n-y)} \int_{0}^{0.524} \theta^{\alpha_0+y-1}(1-\theta)^{\beta_0+n-y-1}\,d\theta \end{aligned} \tag{14.4}$$

With $\alpha_0 = \beta_0 = 35$, $n = 286$ and $y = 149$ these become

$$Pr(M_1|y) = \frac{\Gamma(356)}{\Gamma(184)\,\Gamma(172)} \int_{0.524}^{1} \theta^{183}\,(1-\theta)^{171}\,d\theta = 0.3941$$

$$Pr(M_2|y) = 1 - Pr(M_1|y) = 0.6058.$$

The Bayes Factor is then computed from (14.1) as $BF(M_1, M_2) = 1.2384$ which provides essentially no evidence in favor of model M_1 over model M_2 and no evidence in favor of model $M2$ over model $M1$ either (the BF in favor of model $M2$ is 0.8075).

This example allows us to make an additional point about the distinction between Bayes factors and posterior (or prior) model probabilities. Notice

that in this example both the prior and posterior odds in favor of model M_1 are less than 1.0. In fact, we have that *a posteriori* model M_2 is 1.537 times as likely as is model M_1. Yet, the BF in favor of M_1 is greater than 1.0, because *a priori* model M_2 was 1.904 times as likely as model M_1. Thus, although neither prior to observation of data nor after observation could we claim it is more likely that $\theta > 0.524$ than $\theta < 0.524$, but the data have increased the odds of M_1 over what they had been based on our prior belief.

14.2.3 Two Computational Forms

The intention of this example is not to introduce a different situation but, rather, to simply demonstrate that computing a Bayes Factor on the basis of expression (14.1) as in the previous example is equivalent to computing that factor on the basis of expression (14.3). The key to use of (14.3) is to determine what the prior distributions are under models M_1 and M_2. For model $M_1 : \theta > 0.524$, the original beta (35, 35) prior gives $\pi_1(\theta)$ as a beta (35, 35) distribution truncated below at 0.524, while $pi_2(\theta)$ becomes a beta (35, 35) distribution truncated above at 0.524,

$$\pi_1(\theta) = \frac{\theta^{34}(1-\theta)^{34}}{\int_{0.524}^{1} u^{34}(1-u)^{34}\,du},$$
$$\pi_2(\theta) = \frac{\theta^{34}(1-\theta)^{34}}{\int_{0}^{0.524} u^{34}(1-u)^{34}\,du}.$$

The marginal probability mass functions for these models are then given by the integrals

$$h_1(y|M_1) = \frac{n!}{y!\,(n-y)!}\int_{0.524}^{1} \theta^y(1-\theta)^{n-y}\,\pi_1(\theta)\,d\theta$$
$$h_2(y|M_2) = \frac{n!}{y!\,(n-y)!}\int_{0}^{0.524} \theta^y(1-\theta)^{n-y}\,\pi_2(\theta)\,d\theta. \qquad (14.5)$$

With $\pi_1(\theta)$ and $\pi_2(\theta)$ as given previously, $y = 149$ and $n = 286$, the numerical values for this example are,

$$h_1(149|M_1) = 4.9848 \times 10^{-87}$$
$$h_2(149|M_2) = 4.0258 \times 10^{-87},$$

so the Bayes factor in favor of model M_1 is $BF = h_1(149|M_1)/h_2(149|M_2) = 1.2382$ which, within rounding error is equal to the Bayes factor computed in the previous subsection.

14.2.4 Two Disjunctive Simple Hypotheses

Testing $\theta > 0.50$ against $\theta < 0.50$ or $\theta > 0.524$ against $\theta < 0.524$ allows the result to be influenced by the posterior distribution of probability in regions that are clearly irrelevant to the question at hand, although that should be mitigated to a large extent by choosing a scientifically realistic prior. Consider testing the two point hypotheses $H_0 : \theta = 0.50$ and $H_1 : \theta = 0.524$. In order for these hypotheses to constitute a logical disjunction, we must discard our model with a beta (35, 35) prior for θ and arbitrarily assign point probability masses to the two values of θ. Continuing to associate model M_1 with H_1 and model M_2 with H_0, a typical approach would be to take the prior probabilities for these models to be $Pr(M_1) = Pr(M_2) = 0.50$. This is the approach taken, for example, by Lesaffre and Lawson (2012, Chapter 3.8.2). Posterior model probabilities may then be computed from Bayes Theorem. For model M_1 we would have,

$$\begin{aligned}\Pr(M_1|\boldsymbol{y}) &= \frac{Pr(y|M_1)Pr(M_1)}{Pr(y|M_1)Pr(M_1) + Pr(y|M_2)Pr(M_2)} \\ &= \frac{f(y|\theta = 0.524)}{f(y|\theta = 0.524) + f(y|\theta = 0.50)}.\end{aligned}$$

With $y = 149$ and $n = 286$ this results in $Pr(M_1|y) = 0.0469$. Similarly, $Pr(M_2|y) = 0.0367$, and the Bayes factor in favor of M_1 is,

$$BF(M_1, M_2) = \frac{Pr(M_1|y)}{Pr(M_2|y)} = 1.2796,$$

again providing no evidence in favor of M_1 over M_2. Notice that we will clearly get the same result if we use (14.3) to compute the Bayes factor.

An objection to this procedure is that in order to achieve a disjunctive set of hypotheses, we have abandoned our specification of prior knowledge about θ. Thus, it could be claimed that we are no longer testing values of θ within the context of our chosen model, but rather under some other model. It would not be a coherent analysis to include both a posterior summary of our belief about θ (e.g., a credible interval) based on a beta prior, and a test of $\theta = 0.524$ versus $\theta = 0.500$ based on prior odds of 1.0. We will attempt to circumvent this criticism when we consider tests for hypotheses that do not constitute logical disjunctions, but will encounter other potential complications in doing so.

14.2.5 Disjunctive Simple and Composite Hypotheses

A familiar framework in hypothesis testing is to combine a simple hypothesis (usually as the null H_0) and a composite hypothesis (usually as the alternative H_1). This situation presents something of a challenge in a situation involving a continuous prior distribution. A solution suggested by Robert (2007) is to use a prior that specifies a point mass for the simple hypothesis and a continuous distribution for the composite hypothesis, with the overall prior then taken as the mixture of a discrete and a continuous distribution. Continuing with our guanaco example, associate model M_2 with $H_0 : \theta = 0.50$ and model M_1 with $H_1 : \theta \neq 0.50$. The prior for model M_2 is

$\pi_2(\theta) = I(\theta = 0.50)$ and the prior for model M_1, $\pi_1(\theta)$ remains a beta (35, 35) distribution. The overall prior is then, for some choice of $0 < \rho < 1$,

$$\pi(\theta) = \rho I(\theta = 0.50) + (1 - \rho)\pi_1(\theta).$$

The prior probabilities for the models are just $Pr(M_1) = 1-\rho$ and $Pr(M_2) = \rho$. The marginal probabilities of the data are

$$\begin{aligned} Pr(y|M_1) &= \int_{\Theta_1} f(y|\theta)\pi_1(\theta)\, d\theta, \\ Pr(y|M_2) &= f(y|\theta_0). \end{aligned}$$

We can then compute the posterior probabilities of the models using Bayes Theorem,

$$\begin{aligned} Pr(M_1|y) &= \frac{Pr(y|M_1)Pr(M_1)}{Pr(y|M_1)Pr(M_1) + Pr(y|M_2)Pr(M_2)} \\ Pr(M_2|y) &= \frac{Pr(y|M_2)Pr(M_2)}{Pr(y|M_1)Pr(M_1) + Pr(y|M_2)Pr(M_2)} \\ &= 1 - Pr(M_1|y). \end{aligned}$$

For our guanaco example,

$$\begin{aligned} Pr(y|M_1) &= \frac{n!}{y!(n-y)!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta \\ &= \frac{n!}{y!(n-y)!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)} \\ Pr(y|M_2) &= \frac{n!}{y!(n-y)!}\theta_0^y(1-\theta_0)^{n-y}. \end{aligned}$$

If we choose $\rho = 0.5$,

$$\begin{aligned} Pr(M_1|y) &= \frac{Pr(y|M_1)}{Pr(y|M_1) + Pr(y|M_2)}, \\ Pr(M_2|y) &= \frac{Pr(y|M_2)}{Pr(y|M_1) + Pr(y|M_2)}. \end{aligned}$$

With $y = 149$, $n = 286$, $\alpha = \beta = 35$, and $\theta_0 = 0.50$, we have

$$Pr(y|M_1) = \frac{n!}{y!(n-y)!} 4.356 \times 10^{-87}$$
$$Pr(y|M_2) = \frac{n!}{y!(n-y)!} 8.043 \times 10^{-87},$$

so that $Pr(M_1|y) = 0.3513$ and $Pr(M_2|y) = 0.6487$, with a Bayes factor in favor of M_1 over M_2 of $BF(M_1, M_2) = Pr(M_1|y)/Pr(M_2|y) = 0.5415$, giving no evidence in favor of $M_1 : \theta \neq 0.50$ over $M_2 : \theta = 0.50$. In fact, since the Bayes factor is less than 1, we can say that the Bayes factor in favor of M_2 over M_1 is $BF(M_2, M_1) = 1.8467$.

According to typical scales against which to judge Bayes factors, in this example there is no evidence for M_1 over M_2, nor is there any evidence for M_2 over M_1, not a terribly pleasing result. To reach a conclusion we could take one of two approaches. First, we could fall back on the frequentist notion that in order to control Type I errors at a sufficiently low level, a null hypothesis (here $\theta = 0.50$) needs to be rejected in order to accept an alternative (here $\theta \neq 0.50$). Although the Neyman-Pearson framework for testing hypotheses does not involve evidential measures, this approach would translate into a conclusion that, although we do not have positive evidence in favor of model M_2 over M_1, neither do we have sufficient evidence to reject it.

We can repeat the exercise of this subsection using model M_2 as representing the hypothesis $H_0 : \theta = 0.524$ and model M_1 the hypothesis $H_1 : \theta \neq 0.524$. In this situation, the only change is to $Pr(y|M_2)$ which now becomes $Pr(y|M_2) = f(y|0.524) = n!/[(n-y)!y!]10.292 \times 10^{-87}$. With this change we now have $Pr(M_1|y) = 0.2974$ and $Pr(M_2|y) = 0.7026$ or a Bayes factor in favor of model M_1 of $BF(M_1, M_2) = 0.4233$ and a Bayes factor in favor of model M_2 of $BF(M_2, M_1) = 2.362$. Based on the typical reference scales for

Bayes factors we would say there is no evidence in favor of either M_1 or M_2 as opposed to the other. We could, however, also assert that the evidence in favor of $\theta = 0.524$ over $\theta \neq 0.524$ is greater than the evidence in favor of $\theta = 0.50$ over $\theta \neq 0.50$.

14.2.6 Two Non-Disjunctive Simple Hypotheses

In a great deal of the literature on testing, both frequentist and Bayesian, it is assumed that one of the two models must hold so that prior model probabilities are $Pr(M_1) + Pr(M_2) = 1.0$ as was done in a previous subsection. As noted there, with simple hypotheses this can meet with some objection if the prior distribution for the overall problem gives positive probability to more than the discrete choice of M_1 or M_2. In this section, we want to assess the relative evidence for models that correspond to two simple hypotheses under the original data model and prior used in the problem. For our guanaco problem that means a binomial data model with a beta $(35, 35)$ prior. Under this continuous prior, and for models $M_1 : \theta = \theta_1$, $M_2 : \theta = \theta_0$, neither the prior probabilities $Pr(M_1)$ and $Pr(M_2)$ nor the posterior probabilities $P(M_1|y)$ and $Pr(M_2|y)$ exist. But both prior and posterior densities exist and are positive at θ_1 and θ_0. Consider models (hypotheses) constructed so that θ is specified as lying in some small intervals containing θ_1 and θ_0, $M_1 : \theta_1 - \epsilon_1 < \theta < \theta_1 + \delta_1$ and $M_2 : \theta_0 - \epsilon_0 < \theta < \theta_0 + \delta_0$ for some small positive values ϵ_1, δ_1, ϵ_0 and δ_0. There is no assumption that these values are equal or that θ_1 and θ_0 are necessarily the centers of the intervals. Now, the mean value theorem of integral calculus gives that for a function h continuous on the interval (a, b), there exists a value $a < x < b$ such that $\int_a^b h(t)\, dt = (b - a)h(x)$. Notice that this result takes a and b as fixed and

gives the existence of some x in the interval (a, b) but not its exact value. But based on this theorem, if π is a continuous prior density for θ, then there exist values T_1 and T_0 such that $Pr(\theta_1 - \epsilon_1 < \theta < \theta_1 + \delta_1) = (\delta_1 + \epsilon_1)\pi(T_1)$ and $Pr(\theta_0 - \epsilon_0 < \theta < \theta_0 + \delta_0) = (\delta_0 + \epsilon_0)\pi(T_0)$. Rather than the existence of T_1 and T_0 such that these probability statements hold for given ϵ and δ values, we want existence of the ϵ and δ for given and fixed values of T_0 and T_1, which will be the hypothesized values θ_1 and θ_0. And what we need then is for $(\delta_1 + \epsilon_1) = (\delta_0 + \epsilon_0)$. If this is so, then we have that,

$$\begin{aligned}\frac{Pr(M_1)}{Pr(M_2)} &= \frac{Pr(\theta_1 - \epsilon_1 < \theta < \theta_1 + \delta_1)}{Pr(\theta_0 - \epsilon_0 < \theta < \theta_0 + \delta_0)} \\ &= \frac{(\delta_1 + \epsilon_1)\pi(\theta_1)}{(\delta_0 + \epsilon_0)\pi(\theta_0)} \\ &= \frac{\pi(\theta_1)}{\pi(\theta_0)}.\end{aligned}$$

If there exist ϵ and δ values that satisfy the equality constraint and are also quite small, then we have that M_1 and M_2 correspond, for all practical purposes, to $\theta = \theta_1$ and $\theta = \theta_0$, respectively. A parallel development gives

$$\frac{Pr(M_1|y)}{Pr(M_2|y)} = \frac{p(\theta_1|y)}{p(\theta_0|y)}$$

and $BF(M_1, M_2)$ can be computed from (14.1).

For our guanaco example, $\pi(\theta)$ is a beta$(35, 35)$ distribution and $p(\theta|y)$ is a beta$(184, 172)$ distribution. For testing $M_1 : \theta = 0.524$ against $M_2 : \theta = 0.50$ we obtain $BF(M_1, M_2) = 1.280$ which indicates insufficient evidence to favor either M_1 or M_2 over the other.

Small Interval Hypotheses

A potential criticism of the previous procedure for two simple hypotheses is that technically we changed the problem from $M_1 : \theta = \theta_1$ to $M_1 :$

$\theta \in$ a small interval includingθ_1 and similarly for M_2. Why not, then, simply specify small intervals around θ_1 and θ_2 explicitly and avoid need to argue about the relation between probabilities and densities. In the guanaco problem we might, for example specify $M_1 : 0.523 < \theta < 0.525$ and $M_2 : 0.499 < \theta < 0.501$. This might go a long way toward being both scientifically meaningful and statistically pleasing.

Because the hypothesized intervals are equal in length, the prior odds under the original uniform prior is $Pr(M_1)/Pr(M_2) = 1$, and both of these probabilities actually exist. The BF is then again equal to the posterior odds,

$$\frac{Pr(M_1|y)}{Pr(M_2|y)} = \frac{\int_{0.523}^{0.525} p(\theta|y)}{\int_{0.499}^{0.501} p(\theta|y)} = 1.2795,$$

a nearly identical value to that from the previous subsection with two point hypotheses.

While this type of procedure seems quite pleasing, it does involve an arbitrary choice of intervals, but for many problems this is not a great concern. For example, if our choice of intervals had been $M_1 : 0.490 < \theta < 0.510$ and $M_2 : 0.515 < \theta < 0.535$ we would arrive at $BF = 1.2629$.

14.2.7 Comparing Groups

Consider now a two group comparison. Suppose that we wish to test whether the binomial parameter θ can be taken as the same for the first two and the last year of observation in the guanaco example. Here, let Y_1 be a random variable associated with the number of male births in 1991 and 1992, and let Y_2 be associated with the number of male births in 1993. The data model

consists of two independent (or at least exchangeable) binomial distributions,

$$\begin{aligned} f_1(y_1|\theta_1) &= \frac{n_1!}{y_1!\,(n_1-y_1)!}\theta_1^{y_1}\,(1-\theta_1)^{n_1-y_1}; \quad y_1 = 0, 1, \ldots, n_1; \\ f_2(y_2|\theta_2) &= \frac{n_2!}{y_2!\,(n_2-y_2)!}\theta_2^{y_2}\,(1-\theta_2)^{n_2-y_2}; \quad y_2 = 0, 1, \ldots, n_1; \\ &\textit{and} \\ f(y_1, y_2|\theta_1, \theta_2) &= f_1(y_1|\theta_1)\, f_2(y_2|\theta_2) \end{aligned}$$

Let model M_1 correspond to the hypothesis that $\theta_1 = \theta_2 = \theta$ and model M_2 correspond to the hypothesis that $\theta_1 \neq \theta_2$. While perhaps a bit redundant, we will rewrite the data model under model M_2 to make the difference in these models explicit as,

$$\begin{aligned} f_1(y_1|\theta) &= \frac{n_1!}{y_1!\,(n_1-y_1)!}\theta^{y_1}\,(1-\theta)^{n_1-y_1}; \quad y_1 = 0, 1, \ldots, n_1; \\ f_2(y_2|\theta) &= \frac{n_2!}{y_2!\,(n_2-y_2)!}\theta^{y_2}\,(1-\theta)^{n_2-y_2}; \quad y_2 = 0, 1, \ldots, n_1; \\ &\textit{and} \\ f(y_1, y_2|\theta) &= f_1(y_1|\theta)\, f_2(y_2|\theta) \end{aligned}$$

Now, under model M_2 we need two prior distributions $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$, both of which we will take to be beta distributions with parameters $\alpha = \beta = 35$). Under model M_1 we need only one prior $\pi(\theta)$ which will also be beta$(35, 35)$.

Posterior odds are difficult to calculate directly in this example, but the Bayes factor is not difficult to compute on the basis of expression (14.3) as

$$\begin{aligned} BF(M_1,\, M_2) &= \frac{\displaystyle\int_0^1 f_1(y_1|\theta)\, f_2(y_2|\theta)\, \pi(\theta)\, d\theta}{\displaystyle\int_0^1 \int_0^1 f_1(y_1|\theta_1)\, f_2(y_2|\theta_2)\, \pi_1(\theta_1)\, \pi_2(\theta_2)\, d\theta_1\, d\theta_2} \\ &= \frac{\displaystyle\int_0^1 f_1(y_1|\theta)\, f_2(y_2|\theta)\, \pi(\theta)\, d\theta}{\displaystyle\int_0^1 f_1(y_1|\theta_1)\, \pi_1(\theta_1)\, d\theta_1 \int_0^1 f_2(y_2|\theta_2)\, \pi_2(\theta_2)\, d\theta_2} \end{aligned} \tag{14.6}$$

The leading combinatorial constants in f_1 and f_2 cancel in this ratio, and the constant from the priors is $\Gamma(\alpha+\beta)/\{\Gamma(\alpha)\Gamma(\beta)\}$ in the numerator and the square of that value in the denominator. Then,

$$BF(M_1, M_2) = \frac{\Gamma(\alpha)\Gamma(\beta)\int_0^1 \theta^{\alpha+y_1+y_2-1}(1-\theta)^{\beta+n_1+n_2-y_1-y_2-1}\,d\theta}{\Gamma(\alpha+\beta)\int_0^1 \theta_1^{\alpha+y_1-1}(1-\theta_1)^{\beta+n_1-y_1-1}\,d\theta_1 \int_0^1 \theta_2^{\alpha+y_2-1}(1-\theta_2)^{\beta+n_2-y_2-1}\,d\theta_2}$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)\dfrac{\Gamma(\alpha+y_1+y_2)\Gamma(\beta+n_1+n_2-y_1-y_2)}{\Gamma(\alpha+\beta+n_1+n_2)}}{\Gamma(\alpha+\beta)\dfrac{\Gamma(\alpha+y_1)\Gamma(\beta+n_1-y_1)\Gamma(\alpha+y_2)\Gamma(\beta+n_2-y_2)}{\Gamma(\alpha+\beta+n_1)\Gamma(\alpha+\beta+n_2)}}$$

With $y_1 = 104$, $y_2 = 45$, $n_1 = 192$, and $n_2 = 94$, the Bayes Factor becomes

$$BF(M_1,\, M_2) = \frac{\Gamma(35)\Gamma(35)\dfrac{\Gamma(184)\,\Gamma(172)}{\Gamma(356)}}{\Gamma(70)\dfrac{\Gamma(139)\Gamma(123)\Gamma(80)\Gamma(84)}{\Gamma(262)\,\Gamma(164)}}$$

$$= 0.9392$$

We would interpret this Bayes Factor as providing no evidence in favor of model M_1 over model M_2, but also no evidence in favor of model M_2 over M_1; $BF(M_2, M_1) = 1.0648$.

In this case, prior odds cannot be produced as a consequence of an overall prior distribution as they were in previous examples. We are left to assign prior odds arbitrarily, which will then determine the posterior odds. If we assign prior odds of 1.0 then the posterior odds in favor of M_1 are equal to $BF(M_1, M_2)$, but if we would assign prior odds of 4, say, then the posterior odds would be $4(0.9392) = 3.757$. It seems a bit unsettling that the posterior odds of the models should depend on an arbitrary choice of prior odds.

14.2.8 Need for a Hierarchical Model

We can extend the two group comparison of the previous example to investigate whether data support the use of a mixture or hierarchical model. Consider simulated data in which two data sets consisting of five binomial observations each were generated, one from a model with constant binomial parameter $\theta = 0.55$ and one from a model with binomial parameters first simulated from a beta distribution. The beta distribution had parameters $\alpha = 4$ and $\beta = 3.2727$ which gives $E(\theta) = 0.55$ and $var(\theta) = 0.02999$. These simulated data are presented in Table 14.1. Let the data model be given as

Index i	n_{1i}	y_{1i}	n_{2i}	y_{2i}
1	30	14	30	11
2	30	19	30	16
3	30	18	30	18
4	30	16	30	25
5	30	14	30	20

Table 14.1: Data simulated from a binomial (n_{1i} and y_{1i}) and a beta binomial (n_{2i} and y_{2i}) with equal expected values of 0.55.

a set of 5 independent binomial distributions where, for $i = 1, \ldots, 5$,

$$f_i(y_i|\theta_i) = \frac{n_i!}{y_i!\,(n_i - y_i)!}\,\theta_i^{y_i}\,(1-\theta_i)^{n_i - y_i}; \quad y_1 = 0, 1, \ldots, n_i.$$

Model M_1 will assume that the binomial parameters are distinct, and assign each a beta prior with parameters $\alpha_0 > 0$ and $\beta_0 > 0$ so that, for $i = 1, \ldots, 5$,

$$\pi(\theta_i) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\,\theta_i^{\alpha_0 - 1}\,(1-\theta_i)^{\beta_0 - 1}; \quad 0 < \theta_i < 1.$$

Model M_2 will assume that $\theta_1 = \theta_2 = \ldots = \theta_5 = \theta$, and assign a beta prior with parameters α_0 and β_0 to the common θ. Let $S_y = \sum_{i=1}^5 y_i$ and $S_n = \sum_{i=1}^5 n_i$. The Bayes Factor to assess evidence in favor of model M_1 over M_2 may then be written as,

$$\begin{aligned} BF(M_1,\, M_2) &= \frac{\displaystyle\prod_{i=1}^{5} \int_0^1 f_i(y_i|\theta_i)\, \pi(\theta_i)\, d\theta_i}{\displaystyle\int_0^1 \left[\prod_{i=1}^{5} f_i(y_i|\theta)\right] \pi(\theta)\, d\theta} \\ &= \frac{\displaystyle\prod_{i=1}^{5} \left[\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\, \frac{\Gamma(\alpha_0+y_i)\Gamma(\beta_0+n_i-y_i)}{\Gamma(\alpha_0+\beta_0+n_i)}\right]}{\left[\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\, \frac{\Gamma(\alpha_0+S_y)\Gamma(\beta_0+S_n-S_y)}{\Gamma(\alpha_0+\beta_0+S_n)}\right]} \\ &= \frac{\left[\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\right]^4 \displaystyle\prod_{i=1}^{5} \left[\frac{\Gamma(\alpha_0+y_i)\Gamma(\beta_0+n_i-y_i)}{\Gamma(\alpha_0+\beta_0+n_i)}\right]}{\frac{\Gamma(\alpha_0+S_y)\Gamma(\beta_0+S_n-S_y)}{\Gamma(\alpha_0+\beta_0+S_n)}} \end{aligned} \tag{14.7}$$

Applying the Bayes Factor of (14.7) with $y_i = y_{1i}$ and $n_i = n_{1i}$ from Table **??** results in $BF(M_1,\, M_2) = 0.0209$ or $BF(M_2,\, M_1) = 1/BF(M_1,\, M_2) = 47.835$ so that for these data we would conclude there is strong evidence in favor of a constant binomial parameter (model M_2). Repeating with $y_i = y_{2i}$ and $N_i = n_{2i}$ gives $BF(M_1,\, M_2) = 8.286$ and we would conclude that there is some evidence that the binomial parameter should not be considered constant for these data. Use of Bayes Factors in this example has led to the conclusions we would expect, given that data in the first set were simulated from a model with constant binomial parameter while data in the second set had differing binomial parameters.

14.3 Sensitivity of Bayes Factors to Prior Specification

Bayes Factors do not exist for models with improper prior distributions. Despite the fact that they can often be computed on the basis of marginal likelihoods as in (14.3) they are defined by (14.1) which requires $0 < Pr(M_i) < 1$ for $i = 1, 2$. Bayes Factors are also known to be sensitive to the prior specification used in a model or models, even for proper priors. It is tempting to examine the form of a Bayes Factor as a ratio of marginal likelihoods $h_1(\boldsymbol{y}|M_1)/h_2(\boldsymbol{y}|M_2)$ and conclude that a Bayes Factor is driven entirely by the observed data. But the forms of the $h_i(\boldsymbol{y}|M_i)$ involve integrating over prior distributions of data model parameters, and the priors thus have an important influence on a Bayes Factor. A simple example will serve to illustrate that Bayes factors tend to be quite sensitive to the specification of prior distributions for model parameters. Consider a two sample problem with equal sample sizes such that $\{Y_{i,j} : j = 1, \ldots, n\}$ are independent with normal distributions having expected value μ_i and variance 1. We wish to compare model $M_1 : \mu_1 \neq \mu_2$ with model $M_2 : \mu_1 = \mu_2 = \mu$. Let the two prior distributions for model M_1 be normal with mean 0 and variance τ^2 and, similarly, let the single prior for model M_2 also be normal with mean 0 and variance τ^2. Because the data model variance is known in this example, we may consider only the distributions of sample means $\bar{Y}_1$ and $\bar{Y}_2$ which, under model M_1, are normal with expected values μ_1 and μ_2, respectively, and common variance $1/n$. Under model M_2 these distributions are both normal with common expected value μ and common variance $1/n$. Then we

have, for $i = 1, 2$,

$$\begin{aligned} f_i(\bar{y}_i|\mu_i) &= \frac{n}{(2\pi)^{1/2}} \exp\left\{-\frac{n}{2}(\bar{y}_i - \mu_i)^2\right\}, \\ \pi_i(\mu_i) &= \frac{1}{(2\pi\tau^2)^{1/2}} \exp\left\{-\frac{1}{2\tau^2}\,\mu_i^2\right\} \\ \pi(\mu) &= \frac{1}{(2\pi\tau^2)^{1/2}} \exp\left\{-\frac{1}{2\tau^2}\,\mu^2\right\}. \end{aligned}$$

Under model M_1, $f(\bar{y}_1, \bar{y}_2|\mu_1, \mu_2) = f_1(\bar{y}_1|\mu_1)\, f_2(\bar{y}_2|\mu_2)$ and

$$\begin{aligned} h(\bar{y}_1, \bar{y}_2|\tau^2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\bar{y}_1, \bar{y}_2|\mu_1, \mu_2)\, \pi_1(\mu_1)\, \pi_2(\mu_2)\, d\mu_1\, d\mu_2 \\ &= \int_{-\infty}^{\infty} f_1(\bar{y}_1|\mu_1)\, \pi_1(\mu_1)\, d\mu_1 \int_{-\infty}^{\infty} f_2(\bar{y}_2|\mu_2)\, \pi_2(\mu_2)\, d\mu_2 \end{aligned}$$

Using explicit forms for f_1, f_2, π_1, and π_2, completing the square in the exponent, and noting that the kernel of a normal distribution with expected value μ and variance σ^2 integrates to $(2\pi\sigma^2)^{1/2}$, we arrive at,

$$\int_{-\infty}^{\infty} f_i(\bar{y}_i|\mu_i)\, \pi_i\, d\mu_i = \left(\frac{n}{2\pi}\right)^{1/2} \frac{1}{(2\pi\tau^2)^{1/2}} \left(\frac{2\pi\tau^2}{n\tau^2+1}\right)^{1/2} \exp\left\{\frac{n^2\tau^2\bar{y}_i^2}{2(n\tau^2+1)} - \frac{n\bar{y}_i^2}{2}\right\}.$$

The marginal density for model M_1 then becomes

$$h(\bar{y}_1, \bar{y}_2|\tau^2) = \frac{n}{2\pi}\,\frac{1}{2\pi\tau^2}\,\frac{2\pi\tau^2}{n\tau^2+1} \exp\left\{\frac{n^2\tau^2}{2(n\tau^2+1)}(\bar{y}_1^2 + \bar{y}_2^2) - \frac{n}{2}(\bar{y}_1^2 + \bar{y}_2^2)\right\}. \tag{14.8}$$

Under model M_2, $f(\bar{y}_1, \bar{y}_2|\mu) = f_1(\bar{y}_1|\mu)\, f_2(\bar{y}_2|\mu)$ and

$$h(\bar{y}_1, \bar{y}_2|\tau^2) = \int_{-\infty}^{\infty} f(\bar{y}_1, \bar{y}_2|\mu)\, \pi(\mu)\, d\,\mu.$$

With the same progression used with model M_1, explicit forms of f_1, f_2 and π result in,

$$h(\bar{y}_1, \bar{y}_2|\tau^2) = \frac{n}{2\pi}\,\frac{1}{(2\pi\tau^2)^{1/2}} \left(\frac{2\pi\tau^2}{n\tau^2+1}\right)^{1/2} \exp\left\{\frac{n^2\tau^2}{2(n\tau^2+1)}(\bar{y}_1 + \bar{y}_2)^2 - \frac{n}{2}(\bar{y}_1^2 + \bar{y}_2^2)\right\}. \tag{14.9}$$

The ratio of (14.8) to (14.9) then gives the Bayes Factor in favor of M_1 as,

$$\begin{aligned} BF(M_1,\, M_2) &= \left(\frac{n}{2\pi}\right)^{1/2} \frac{1}{(2\pi\tau^2)^{1/2}} \left(\frac{2\pi\tau^2}{n\tau^2+1}\right)^{1/2} \\ &\quad \times \exp\left[\frac{n^2\tau^2}{2(n\tau^2+1)}\{\bar{y}_1^2 + \bar{y}_2^2 - (\bar{y}_1+\bar{y}_2)^2\}\right] \\ &= \frac{1}{(n\tau^2+1)^{1/2}} \exp\left[\frac{n^2\tau^2}{2(n\tau^2+1)}\{\bar{y}_1^2 + \bar{y}_2^2 - (\bar{y}_1+\bar{y}_2)^2\}\right]. \end{aligned} \tag{14.10}$$

Now, for given values of n, $\bar{y}_1$ and $\bar{y}_2$, the Bayes Factor in (14.10) is monotone decreasing in τ^2 with a limit of 0 as $\tau^2 \to \infty$. Thus, as prior variance τ^2 increases, the Bayes Factor places more and more weight on the model with a single mean (model M_2), regardless of the data. Contrast this behavior with that of the posterior mean. As prior variance increases the posterior expectation becomes more and more influenced by the data rather than the prior. The implication is that the more diffuse a prior distribution, the more results are driven solely by data. In fact, if the variance of a normal prior is allowed to go to infinity the result is an improper prior in which case the posterior mean is the maximum likelihood estimate. The behavior of the Bayes factor in this example is in direct conflict with that notion. As the prior variance is allowed to go to infinity, the data are completely excluded from what we intend as a measure of the relative evidence in favor of model M_1 over M_2. Of course, if the prior actually becomes improper then the Bayes Factor does not exist, as already discussed. Nevertheless, the message is that Bayes Factors can be sensitive to prior specification, regardless of the data.

14.4 Bayes Factors and Posterior Odds

When prior model odds are $Pr(M_1)/Pr(M_2) = 1$ the Bayes Factor is equal to posterior odds $BF(M_1, M_2) = Pr(M_1|y)/Pr(M_2|y)$ and interpretation is straightforward. But Bayes Factors are, fundamentally, an odds ratio that represents the proportional shift between prior and posterior odds (see 14.1)). This makes Bayes Factors subject to the same potential difficulties in interpretation that occur for all odds ratios. Specifically, for a given value of posterior odds, the smaller the prior odds in favor of model M_1 the greater the value of the Bayes Factor. Fortunately, prior odds are often dictated by the overall structure of a model. This is true for most cases in which Bayes Factors are used to test hypotheses about parameter values. In cases of model comparison, for which prior odds are subject to judgment, it is generally natural to assign equal prior probabilities to the models being compared. Nevertheless, it is wise to ascertain whether a given Bayes Factor is representative of posterior model probabilities (i.e., equal to the posterior odds) or representative of the change from prior to posterior probabilities. The posterior odds can nearly always be computed. It seems advisable to always report posterior odds as well as a Bayes Factor when the two are not equal.

As a final note, the situations considered in this chapter have dealt only with simple models consisting of a data model and a prior. For a discussion of how Bayes Factors can be obtained from the output of MCMC samplers see Carlin and Louis (2000, *Bayes and Empirical Bayes Methods for Data Analysis,* 2^{nd} *ed., Chapman and Hall/CRC*, especially Chapter 6).

Chapter 15

Exchangeability and Representation Theorems

This chapter covers two reasonably related topics, exchangeability and representation theorems. These topics have been used in various ways to motivate certain priors and data models, to give mathematical meaning to the concept of prior distributions, and even to claim that everyone is a Bayesian, whether they know it or not.

15.1 Exchangeability

A common assumption in Bayesian analyses is that the observable random variables $Y_1, \ldots, Y_n$ are *exchangeable*. The meaning of exchangeable is given in the following definition.

Definition:

1. $Y_1, \ldots, Y_n$ are marginally exchangeable if, for a probability density or

mass function $m(\cdot)$ and permutation operator $\mathcal{P}$,

$$m(y_1, \ldots, y_n) = m\left(\mathcal{P}(y_1, \ldots, y_n)\right).$$

2. $Y_1, \ldots, Y_n$ are conditionally exchangeable given z if, for a probability density or mass function $m(\cdot)$ and permutation operator $\mathcal{P}$,

$$m(y_1, \ldots, y_n|z) = m\left(\mathcal{P}(y_1, \ldots, y_n)|z\right).$$

The interpretation of these definitions needs clarification. For any valid joint distribution it is always true that the indices of variables may be permuted. That is, for random variables X_1 and X_2 is is always true that

$$Pr(X_1 = x_1,\, X_2 = x_2) = Pr(X_2 = x_2,\, X_1 = x_1).$$

This is trivial, not exchangeability. What exchangeability implies is that

$$Pr(X_1 = x_1,\, X_2 = x_2) = Pr(X_1 = x_2,\, X_2 = x_1),$$

which is a quite different condition. The implication of exchangeability is that the probability with which random variables assume various values does not depend on the "identity" of the random variables involved; this is essentially a symmetry condition.

It is true that independent and identically distributed random variables are exchangeable, and we often assume the condition of independent and identical distribution, but we should realize that exchangeability is not the same, as shown by the following example.

Example 15.1

1. Exchangeable but not Independent Random Variables. Let the pair of random variables (X, Y) be bivariate with possible values

$$(X, Y) \in \{(0, 1),\, (0, -1),\, (1, 0),\, (-1, 0)\},$$

such that each possible value has probability 0.25.

Exchangeability:

Clearly, $Pr(X = x, Y = y) = Pr(X = y, Y = x)$, since each possible value has the same probability.

Lack of Independence:

$Pr(X = 1) = 0.25$ and $Pr(Y = 0) = 0.5$, but $Pr(X = 1, Y = 0) = 0.25 \neq 0.25(0.5)$

2. Independent but not Exchangeable Random Variables.

Let X and Y be any two independent random variables with X being discrete with probability mass function $f_X(x);\ x \in \Omega_X$ and Y being continuous with probability density function $f_Y(y);\ y \in \Omega_Y$.

Independence: Independence is by assumption, so that the joint (mixed) density (and mass function) is

$$m(x, y) = f_X(x)f_Y(y);\ (x, y) \in \Omega_X \times \Omega_Y.$$

Lack of Exchangeability:

For any $y \notin \Omega_X$ we would have

$$f_X(y)f_Y(x) = 0 \neq f_X(x)f_Y(y),$$

so that X and Y cannot be exchangeable.

In the first portion of this example, what "messes up" independence is the lack of what has previously been called the positivity condition. That is, while $\Omega_X \equiv \{-1, 0, 1\}$ and $\Omega_Y \equiv \{-1, 0, 1\}$, and the probability distributions of X and Y on these sets is the same (i.e., X and Y are identically distributed), it is not true that $\Omega_{X,Y} = \Omega_X \times \Omega_Y$. In the second portion of

the example, what "messes up" exchangeability is that the sets of possible values Ω_X and Ω_Y are not the same, although the positivity condition does hold in this case.

In general, random variables that are not identically distributed cannot be exchangeable, random variables that are independent and identically distributed are exchangeable, but exchangeability is not the same property as independence.

15.2 Representation Theorems

What are known as representation theorems have a long history, dating back at least to 1930 when deFinneti published such a result for $0-1$ "random quantities" (see Bernardo and Smith, 1994, Chapter 4). There have since been a large number of representation results; for example, a 1987 publication by Diaconis and Freedman is titled "A dozen de Finetti-style results in search of a theory (*Ann. Inst. H. Poincare* **23**, 397-423). Such results have been claimed as justification for a wide range of statistical actions, from choice of a data model to justification of constructing a hierarchical as a mixture (e.g., hierarchical models with multi-stage priors), and even claims that, due to these theorems, one *must* be a Bayesian. Our goal here is not to present all of the mathematical intricacies of representation results, but rather to gain a sense for the underlying concept involved (of which there is really only one, despite the plethora of individual theorems) and an understanding of why so many procedures have been claimed as justified on the basis of these results.

In most of what is contained in this section we will adopt language along the lines used by *subjective* Bayesians and avoid to the degree possible referring to random variables. The purpose of this is to make clear one of the uses

of representation theorems, which is to justify the use of random variables and theoretical probability distributions (and density functions) in an analysis of *belief* about observable quantities. Even strident subjective Bayesians do not claim that all problems can be resolved by such justification. For example, Bernardo and Smith (1994, p. 167) state that,

> In some cases, we shall find that we are able to identify general types of belief structure which "pin down", in some sense, the mathematical representation strategy to be adopted. In other cases, this "formal" approach will not take us very far toward solving the representation problem and we shall have to fall back on rather more pragmatic modelling strategies.

The focus of Bernardo and Smith is justifying parametric data models for particular situations, as well as the use of a prior for the data model parameters. The assertion connected to the above quote is that sometimes we may be able to justify the choice of a specific distributional form (e.g., normal) for a data model, but other times we just choose models as we always have without a formal argument that the form chosen is *the correct* form.

15.2.1 de Finetti's Original Result

We present de Finetti's original theorem in much the style of Bernardo and Smith (1994, Proposition 4.1).

Theorem

Suppose that $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of random quantities that can assume values 0 or 1 each, and that $\mathcal{P}$ is a probability measure for these quantities. Then there exists a distribution function Q

such that, for any finite sequence of size n a probability mass function for $x_1, \ldots, x_n$ exists and has the form

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} dQ(\theta),$$

where,

$$\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i,$$

and,

$$Q(\theta) = \lim_{n\to\infty} \mathcal{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq \theta\right].$$

What are the important implications of this result? First, note that the result has defined a "parameter" θ as a function of the observable x_is. There is an existence portion of the result that then says this parameter has a distribution function Q that can be interpreted as a belief about the limiting relative frequency of 1s in sequences of the x_is. Finally, the statement of the result then also implies that the probability mass function of any finite sequence has the form of an integrated (over θ) Bernoulli distribution with parameter θ. From a subjectivist Bayesian perspective, then, this representation theorem says we may *behave* as if,

1. a sequence $x_1, \ldots, x_n$ is a random sample from a Bernoulli distribution with parameter θ

2. θ has a distribution (the prior Q)

So, in this case, an assumption of infinite exchangeablity for quantities that can assume only one of two values has justified the choice of data model and the existence of a distribution for the parameter of that data model. In

addition, the result implies an interpretation for the prior in terms of limiting relative frequencies. Also important, however, is what the theorem does not provide. The result does not indicate a form for Q.

Now, de Finetti's theorem has justified a particular data model for a particular situation and it is not hard to see that similar results for other situations are likely to be specific to the settings considered. This goes a long way toward explaining why there are a large number of representation results. There are also more general versions of representation theorems, one of which is given in the next subsection.

15.2.2 General Representation Theorem

Suppose the $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of real-valued random quantities that have a probability measure $\mathcal{P}$. Then there exists a probability measure Q over the space of all distribution functions $\mathcal{F}$ such that any finite sequence $x_1, \ldots, x_n$, has a joint distribution function P of the form,

$$P(x_1, \ldots, x_n) = \int_{\mathcal{F}} \prod_{i=1}^{n} F(x_i)\, dQ(F),$$

where,

$$Q(F) = \lim_{n \to \infty} \mathcal{P}(F_n),$$

for $F_n(x) = (1/n) \sum_{i=1}^{n} I(x_i \leq x)$, the empirical distribution function of $x_1, \ldots, x_n$.

In the general representation theorem F plays the role of a data model "parameter", much like θ in de Finetti's original theorem.

$$h(x_1, \ldots, x_n) = \int_{\Theta} \prod_{i=1}^{n} f(x_i|\theta)\, dQ(\theta),$$

where now, for some function $\phi(\cdot)$,

$$\theta = \lim_{n\to\infty} \phi(x_1, \ldots, x_n)$$
$$Q(\theta) = \lim_{n\to\infty} \mathcal{P}\left[\phi(x_1, \ldots, x_n) \leq \theta\right].$$

A subtle aspect of this is the form for $Q(\theta)$ as a limit whose argument involves a limit (as θ). Here, $\mathcal{P}$ is a probability measure for $x_1, \ldots, x_n$ but $\lim \phi(x_1, \ldots, x_n) = \theta$, so Q becomes a probability measure for θ. Finally, if Q corresponds to a distribution function G, then

$$h(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} f(x_i|\theta)\, dG(\theta),$$

and if G is absolutely continuous with density g then,

$$h(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} f(x_i|\theta)\, g(\theta)\, d\theta.$$

One aspect of all of this that should be clear is that the general representation theorem functions primarily as an existence result. We have moved quite a ways away from having it provide any practical guidance for model formulation in even generic classes of problems. Additional restrictions, such as what Bernardo and Smith (1994) call "centered spherical symmetry" can sometimes be used to infer particular distributional forms for data models; this spherical symmetry condition for example seems to imply a normal data model. Overall, the additional restrictions that can lead to particular models seem to pertain to invariance of presumed distributions with respect to geometric transformations (as with the spherical symmetry case), or with the existence of sufficient statistics. A bit of caution is needed in this latter case, because we usually attach sufficiency directly to parameters, and that is not what is desired here. One can also consider a statistic to be sufficient

based on conditional distributions of sets of random quantities. For example, if $x_{r1}, \ldots, x_{rm}$ and $x_{s1}, \ldots, x_{sq}$ are non-overlapping sets of quantities, one can define a statistic $t(x_{s1}, \ldots, x_{sq})$ as sufficient for the sequence $x_{r1}, \ldots, x_{rm}$ if $p(x_{r1}, \ldots, x_{rm}|x_{s1}, \ldots, x_{sq}) = p(x_{r1}, \ldots, x_{rm}|t)$. (Note: technically, this is not enough as we wish to consider all subsets of a single sequence $x_1, \ldots, x_n$ rather than just one particular division into two parts, but it serves to communicate the idea).

One might suggest that additional restrictions (beyond exchangeability) such as those given above "take some of the wind from the sail" of representation theorems. For example, one of the geometric invariance restrictions used by Bernardo and Smith (1994, Section 4.3) seems to be essentially a geometric version of a memoryless property (with respect to the origin of Euclidean space). Under this restriction one obtains an exponential data model, which would seem hardly surprising given that the memoryless property is characteristic of exponential distributions. The question that arises, then, is whether one has truly justified a data model based on representation theorems, or simply shifted the assumption of a particular parametric data model to a contrived geometric assumption that amounts to the same thing.

15.3 Uses of Representation Results

We conclude this brief discussion of representation theorems with an indication of some of the claims put forward about what is justified through these results.

1. Justification of Data Model Selection.

   Much of the discussion in the previous subsection centered on the use

of representation theorems to justify the choice of parameterized data models within a *belief framework* that seeks to use, but not depend on, the formulation of mathematical distributions as models. As might be inferred from the discussion presented, this seems most secure in the case de Finetti's original result.

2. Checks on Concordance of Mixture Distributions and Priors.
   Another use that has been made of representation theorems is as a check of agreement between data model, prior, and the marginal distribution for data. Assuming identifiability conditions hold, let $h(x_1, \ldots, x_n) = \int \prod f(x_i|\theta)\, g(\theta)\, d\theta$. Press (2003) makes use of representation results to check that constraints on h and g are in concert. That is, constraints on g (h) imply constraints on h (g), and these can be checked against each other in some situations.

3. Motivation for Hierarchical Models.
   Carlin and Louis (1996), among others, appeal to representation results (i.e., "de Finetti's theorem) to provide motivation for formulation of prior distributions as mixtures, leading to hierarchical models. That is, suppose we have a reasonable parameterized data model, without worrying about representation results to motivate or justify it, but for which we are willing to assume that priors assigned to data model parameters should reflect exchangeability. Then, these authors assert, de Finetti's theorem justifies using a mixture model for prior specification, that is, a two-stage prior.

4. Argue for Obligate Bayesian Analysis.
   The most extreme assertions made on the basis of representation results

seem to be those that claim these results *require* a Bayesian analysis (e.g., Bernardo, 1996). This is based on the existence portion of the results relative to a prior distribution Q. For example, in de Finetti's original result the claim is that if one is willing to interpret limiting relative frequency of 1s as meaningful, the existence of $Q(\theta)$ *mandates* that we deal with it through use of a prior distribution. This is notably stronger than an interpretation that the existence of Q indicates we should be willing to *accept* the use of a prior on θ. Needless to say, Bernardo's stance is not universally accepted even for this fundamental problem, and the picture becomes more clouded in situations that involve extensions and generalizations of the original representation theorem.

I will argue in conclusion of this topic that representation theorems are interesting probability results, but there is a desire on the part of some to infer more from them than is necessarily warranted. Other than the special case of binary quantities, there seems little gained from representation results to guide data model selection (note here that sets of *iid* binary trials also mandate a Bernoulli model although not necessarily a prior). The argument that hierarchical models are justified *in general* by representation theorems seems to over-reach the actual results. Certainly, many distributions *may* be formulated in terms of mixture models, and representation theorems provide additional justification that such structures are useful and mathematically coherent (i.e., not self-contradictory). I would claim, however, that appeal to exchangeability and representation results is not entirely sufficient to justify a particular hierarchical model in any particular application. I would again put forward the scientific argument of mechanisms that are manifested in

different ways under different circumstances as a more pleasing motivation for the use of hierarchical models, when that argument appears to apply to a problem. One end implication is that the use of a hierarchical model (with or without a Bayesian analysis) does not remove the need for model assessment, including the number of levels chosen, the probabilistic forms assigned to those levels, and possibly sensitivity to the choice of priors. A good many Bayesians would not disagree (e.g., Gelman, Carlin, Stern and Rubin 1995, Chapter 6).

Part VII

Topics in the Foundations of Statistics

Part VIII

Topics in the Foundations of Statistics

Chapter 16

Inferential Procedures in Statistical Analysis

In this chapter we offer some discussion on what are known as foundational issues in statistics. Foundational issues concern the philosophical and scientific bases on which the practice of statistics depends. Most of us are familiar with at least the existence of arguments between frequentist and Bayesian approaches to statistical analysis, if not the actual arguments themselves. But there is much more involved in attempting to organize and understand various frameworks that have been proposed within which to logically and philosophically justify the process of making inference on the basis of statistical analysis. There have been multiple such frameworks proposed, from both frequentist viewpoints and Bayesian viewpoints. That is, there are no such things as *the* frequentist approach to statistical analysis and *the* Bayesian approach to statistical analysis, there are multiple approaches and schools of thought in both of these major camps. While we will not attempt to provide any type of complete coverage of all the ideas that have been proposed, we

will highlight a few of the major issues that arise from a consideration of the foundations of statistics.

16.1 A Primer on Statistical Inference

The word inference means reaching conclusions on the basis of observations or information. If an individual eats only food from a particular sandwich shop chain for a year and looses a great amount of weight, we might infer that the food sold by that food chain is not fattening. If we wake in the morning to find it has snowed 14 inches overnight, we might infer that many activities and events scheduled for the day are going to be canceled. If uncollected newspapers accumulate on the porch of our neighbor we might infer that he or she is on a trip (avoiding more morbid possibilities). Depending on our belief system, if Punxsutawney Phil sees his shadow on groundhog day (2 February each year) we might infer there will be six more weeks of winter. Clearly, some of these inferences are more reasonable than others. A question is whether there are there formal rules for making inferences that justify one system more than another.

In making *statistical inference*, the observations or information available are the data that result from a particular study. We wish to draw conclusions on the basis of that information. Sometimes, that conclusion might apply to a real (i.e., physically existing) population, such as when public opinion polls are used to infer the preference of voters for a given candidate or proposition. The population of voters is a physical reality, and we do not need a theoretical model to make inference about the proportion of concern. But often, as has been the theme emphasized in this book, we make statistical inference about the possible values of a parameter in a theoretical model that is applied

to a given situation. Given a normal one sample model postulated for the beginning salaries of university graduates in major G and fitted to (estimated from) a sample of n individuals having that major, we may want to infer what the mean salary will be for individuals graduating this year. In this situation we would make inference *directly* about the parameter that gives the expected value in the normal model, usually μ. We make inference about the mean salary for individuals then only *indirectly*, and the degree to which our inference about μ is meaningful for what we care about in the real world (actually mean salary) depends on how well the normal model formulated and the data collected represent the distribution of salaries among real people in the real world.

We need to clarify our use of the words *directly* and *indirectly* in the previous paragraph because our use of these (in particular the word directly) does not align perfectly with what is by convention termed *direct inference* in much of the literature concerned with the philosophy of statistics (e.g., Seidenfeld, 1978). In much of the literature, direct probabilities are probabilities assigned to individual or singular cases on the basis of knowledge about relative frequencies in a larger reference class. Direct inference is at the heart of Kyburg's epistemological probability (Kyburg, 1971) in which he attempts to solve problems of inverse inference by turning them into problems of direct inference (see Seidenfeld, 1978). Kyburg's episemological probabiity and Fisher's fiducial probability share this objective, although that is about all they share. We will touch on fiducial probability in the discussion of statistical intervals in Chapter 18.

In a good deal of our education in statistics we are primarily concerned with the methodological issues involved, but it is impossible (or at least unwise) to not also give some consideration to the interpretations we attach

to the inferences that result. As a starting point, it is helpful to understand the distinction between *deductive* and *inductive* arguments. A deductive argument involves a conclusion that must be true if the propositions on which it is based are true. If we have propositions that (1) no cows can fly, and (2) Bessy is a cow, then under the assumption that those propositions are true the conclusion that Bessy cannot fly must be true. Note that the validity of a deductive argument does not depend on the actual truth or falsehood of the propositions. Given that (1) all Freshman are stupid and (2) Harry is a Freshman, then the conclusion that Harry is stupid is a valid deductive argument, even though it may not be at all true. In contrast to deductive arguments are *inductive* arguments, in which the truth of the conclusion reached is not a certainly even if the propositions on which it is based are true. Consider the propositions (1) the greatest 12-hour temperature swing in recorded history was 84° F (29° C) in Fairfield, Montana, which went from 63° F at noon to −21° F at midnight on December 14, 1924, and (2) the temperature at noon today was 70° F. An inference that the lowest temperature in the next 12 hours will not be less than −14° F would be a good bet, but its truth does not follow from the truth of the propositions. In 2018, the largest deficit that had ever been overcome by a team to win an NCAA college basketball game was 30 points (by Duke vs. Tulane in 1950 and by Kentucky vs. Louisiana State in 1994). On 22 February 2018, Drexel trailed Delaware by 34 points. A reasonable inference would have been that Drexel would loose that game, but it would not have been certain. In fact, Drexel ended up winning that game to post a new record for comebacks. If I draw 4 pennies in a row from a bag of coins, I might conclude that the bag contains only pennies, another example of an inductive inference – a conclusion supported by observation but about which there is uncertainty.

Here, the information or observation used as a basis for inference is in the form of a sample, bringing us closer to a statistical inference.

Return to our deductive syllogisms. Notice that at least the initial proposition applies to a larger class of entities than the conclusion. The opening proposition No cows can fly is a rather broad statement relative to the conclusion Bessy cannot fly. Similarly, the proposition All Freshman are stupid applies to a much larger collective than the conclusion Harry is stupid. That this is generally true of valid deductive syllogisms leads to the characterization of deductive argument as proceeding from the general to the specific. Consider what happens if we reverse the leading proposition and the conclusion, which leads to 1) Bessy cannot fly. 2) Bessy is a cow. 3) Therefore, cows cannot fly. The conclusion is clearly an inductive inference, and it is illustrations such as this that lead to the characterization of inductive arguments as proceeding from the specific to the general.

The injection of probability into the whole picture can throw a monkey-wrench into the issue of deductive and inductive arguments. Specifically, a valid deductive syllogism can be turned into an argument having an inductive statement as the conclusion. Suppose that we are given only that Bessy is a farm animal. Given what we know about farm animals and the names given them, we might believe the probability is high that Bessy is a cow, but she (or even he) might also be a chicken or a duck. Our deductive syllogism concerning Bessy now has the form 1) No cows can fly. 2) Pr(Bessy is a cow) is large. 3) Bessy cannot fly, This retains the structure of a valid deductive syllogism, and yet the conclusion must have some uncertainty attached to it. That is, we still have the form of a deductive argument that proceeds from the more general to the more specific, but the conclusion is no longer a deductive statement. We consider this to be an example of a deductive

argument that produces an inductive statement as a conclusion.

If you read a sufficient number of articles and/or texts on the foundations of statistics you will find some disagreement about the roles of induction and deduction. Statistical inferences are by nature inductive statements, because there is always uncertainty attached to the conclusion. This is not controversial and is not the source of the disagreements mentioned. The question is whether there are *rules* or a *logic* that renders some conclusion the result of inductive *reasoning*. There certainly do exist such rules connected to deductive reasoning, and those are the concern of formal logic, or formal deductive logic. But whether there can be or cannot be similar rules connected with arguments that proceed from the specific to the general and result in statements or conclusions that are inductive in nature has not been definitively resolved, although there have been many attempts to propose such rules or systems. We return to this issue in Chapter 17.9.3.

One point that is not controversial among statisticians is that, given that statements of statistical inference involve uncertainty, we desire a manner to quantify that uncertainty. It is also accepted that the form of such quantification should be based on, and obey, the rules of probability, although the quantification of uncertainty may not be a probability itself. This latter fact, that quantification of uncertainty, while based in probability, may not be probabilities themselves, is largely what leads to controversies about how to interpret them. It can be argued that this is also the root cause of the development of numerous concepts of probability such as Laplacian (Laplace, 1814, translation by F.W. Truscott and F.L. Emory 1951) relative frequency (Ellis, 1844; Venn, 1866, 1889), hypothetical limiting relative frequency (Reichenbach, 1949; von Mises, 1957), propensity (Kyburg, 1974), logical (Keynes, 1921; Carnap, 1950), nomic (Pollock, 1990), fiducial (Fisher,

1930), measure-theoretic (Kolmogorov, 1933, translation 1950), epistemic (Ramsey, 1931; Koopman, 1940), and pure subjectivist (Jeffreys, 1939). The references given are an attempt to provide some early expositions of these concepts, which are not necessarily completely disjoint, but many of the ideas we now label with these categories existed before they had been given names and evolved over time into the concepts of probability we recognize today.

16.2 Concepts of Probability in Statistical Methods

Of the concepts of probability offered as alternatives in understanding and developing a basis for the practice of statistics, four are easily recognizable in the form of statistical methodologies, Laplacean, relative frequency, hypothetical limiting relative frequency, and epistemic probability. For a review of the fundamental ideas associated with each of these concepts of probability see Chapter 1 of *Intermediate Statistical Methods*. We give here a very brief overview of how these distinct concepts of probability influence different statistical methodologies.

16.2.1 Laplacean Probability and Randomization Procedures

Many readers will have been introduced to the comparison of groups through the use of *permutation* or *randomization* tests in an experimental setting. Consider a simple experimental situation in which we wish to compare responses observed for experimental units under two different sets of conditions, or treatments. It is vital to the testing method described here that which

experimental unit is exposed to which treatment is under the control of the experimenter. The number of arrangements or assignments of experimental units to treatment groups can be determined through the use of combinatorics. For example, in an experiment with two groups, each of which is to have 10 experimental units, there are $K = 20!/(10!\,10!) = 184,756$ unique ways to assign 20 experimental units to the two treatment groups. The actual arrangement used in the experiment is one of those possibilities chosen at random. Under the hypothesis that the treatments are not associated with responses observed for the experimental units, we can compute the difference in responses between the two treatment groups (e.g., difference in sample means) that would have occurred under any of the possible treatment group assignments. Let T_A denote the difference between groups for the treatment assignment actually used in the experiment, and let $\{T_k : k = 1, \ldots, K\}$ denote the difference that would have resulted in treatment arrangement k had been used; note that assuming no ties in the T_k, for exactly one value of k we will have $T_A = T_k$. In a permutation test, a $p-$value associated with the hypothesis that treatments are not related to responses is,

$$p = \frac{1}{K} \sum_{k=1}^{K} I(T_k \leq T_A), \tag{16.1}$$

where $I(A)$ is the indicator function that assumes a value of 1 if A is true and a value of 0 otherwise. A randomization test replaces $\{T_k : k = 1, \ldots, K\}$ in (16.1) with a random sample without replacement of size $K' < K$, which is useful if K is large, which happens easily; $K = 184,756$ for even our small example of an experiment with two treatment groups of size 10 each. In a permutation procedure, the minimum $p-$value possible is $1/K$, in a randomization procedure the minimum $p-$value could be 0 (if T_A is not among the sampled values of T_k).

These randomization procedures involve no random variables, no theoretical probability distributions, and no statistical models. Probability enters the analysis only through the random assignment of experimental units to treatment groups, and the $p-$values are computed, under the hypothesis of no association between treatments and responses, based on a set of equally likely outcomes $\{T_k : k = 1, \ldots, K\}$, which is Laplacian probability.

16.2.2 Relative Frequency and Sampling Methods

By relative frequency probability here, we mean what is sometimes called *finite* relative frequency, which depends on a finite collection of objects or entities, such as a physically existing population of people, animals, cities, etc. The probability of an event E is then simply the relative frequency of those objects or entities that fit the definition of E. This concept of probability is used directly in survey sampling methods, in which we assume a finite population of N population units, from which we will draw a sample of size n based on some probabilistic rule. A common goal is to estimate the total τ (or mean $\mu = \tau/N$) of some attribute of the population units. Let U_i denote the population units and let y_i denote their associated attribute value, both for $i = 1, \ldots, N$. Let $\mathcal{S}$ denote the set of population units selected for the sample. A classic method for unbiased estimation of the population total of the attribute of concern is called a Horvitz-Thompson estimator,

$$\hat{\tau} = \sum_{U_i \in \mathcal{S}} \frac{y_i}{\pi_i}, \tag{16.2}$$

where π_i is the probability that population unit U_i was selected for the sample, $Pr(U_i \in \mathcal{S})$ and is called the *inclusion* probability for population unit U_i.

Calculation of the inclusion probabilities is typically accomplished through the use of finite relative frequency probability. In a simple random sample, this also reduces to Laplacian probability but more involved sampling designs require probabilities for non-equally likely events. As an example, suppose that in a human population of size $N = 10,000$, there are $N_f = 6,000$ females and $N_m = 4,000$ males. We would like an unbiased estimator of the population total of some attribute but we are concerned that values of the attribute might be somewhat different for males and females. In a random sample of size $n = 100$, the probability that a male is selected for the sample is $\pi_m = 4000/10000 = 0.40$ and the probability that a female is selected is $pi_f = 6000/10000 = 0.60$, where these probabilities are computed directly on the basis of finite relative frequency. Letting $\mathcal{S}_m$ and $\mathcal{S}_f$ denote the sets of male and female population units selected for the sample, respectively, The Horvitz-Thompson estimator of τ is,

$$\hat{\tau} = \sum_{U_i \in \mathcal{S}_f} \frac{y_i}{0.60} + \sum_{U_i \in \mathcal{S}_m} \frac{y_i}{0.40}.$$

16.2.3 Hypothetical Limiting Relative Frequency and Theoretical Probability Distributions

Hypothetical limiting relative frequency is the dominant concept of probability used in what we typically mean when we refer to frequentist methodology. Most theoretical probability distributions are interpreted according to this concept of probability. Probability distributions for finite populations are essentially empirical distributions that are unknown and give probabilities as finite relative frequencies. Theoretical probability distributions reflect relative frequencies that would occur if it were possible to take a sample of size n and then let n grow large without bound. Thus, theoretical distributions

give probabilities for values (discrete distributions) or intervals of values (continuous distributions) as the limits of relative frequencies as a sample size tends to infinity, which can only occur hypothetically. Although the device of taking relative frequencies in a population and then letting the population size grow large wihout bound is a fine way to envision the origin of limiting relative frequency probability, the use of this concept of probability is not tied to situations in which some physically existing population can be identified. Once we are into the realm of hypothetical populations, theoretical probability distributions have an existence of their own. In fact, one can even consider a theoretical distribution as defining a hypothetical population.

16.2.4 Epistemic Probability and Prior/Posterior Distributions

As was discussed in greater length in Chapter 7.1 of *Intermediate Statisitcal Methods*, an epistemic concept of probability is used to interpret prior distributions and posterior distributions for fixed data model parameters in Bayesian methods. Contrary to what is sometimes asserted, data model parameters in a Bayesian approach are not random variables unless we are using a hierarchical model, in which case they may or may not be considered random. Prior and Posterior distributions describe the way our belief about possible values of these parameters are spread across the parameter space, the set of all possible values.

In as much as Bayesian inference consists of making probability statements on the basis of posterior distributions, such inference depends on epistemic probability to obtain meaning. A $1-\alpha$ highest posterior density credible interval, for example, reflects an interval that contains $(1-\alpha)100\%$ of our

belief about what the value of the parameter of concern might be based on a combination of our belief about that parameter before (prior to) seeing any data, and information about the parameters that is then provided by observed data.

16.3 Decisions Versus Evidence

An issue that permeates any discussion of the philosophical etiologies of approaches to statistical inference is that of pre-data as opposed to post-data procedures or precision. Pre-data precision refers to rules for making conclusions that can be completely determined without having observed any data. Given the rule, the only role of data and quantities derived from them is to determine which branch of the rule our current problem belongs to. That is, data are involved in determining what the conclusion is, but not the *rule* for drawing conclusions. Procedures with pre-data precision typically take conclusions to be in the form of decisions – decisions for which we want to control certain aspects of the decision making process, such as how often the decision will be incorrect. As we will see, the Neyman-Pearson theory of hypothesis testing is a classic example of a procedure developed for pre-data precision.

Procedures developed around the concept of post-data precision cannot be used to determine an exact rule for drawing conclusions without input from the data. The same value of an inferential quantity might lead to different conclusions, depending on other aspects of the problem, such as what other conclusions might be possible. In procedures developed to attain post-data precision, the rules for drawing conclusions are often relative in nature, although this is not always true. The conclusions themselves may be in

the form of a decision, but that is also not necessarily the case. Rather than decision rules, procedures developed to have high post-data precision are based on evidential measures of support the data may have for various conclusions. Many statistical procedures, such as the Fisherian theory of significance testing, tests based on likelihood support, and most Bayesian inferential procedures at least purport to attain post-data precision and be based on measures of evidence the data provide for inferential statements.

16.3.1 An Illustration

A simple example based on prediction of group membership will serve to illustrate the difference between procedures developed to attain pre-data versus post-data precision. Suppose we have a problem that involves a specific group or sub-population of sampling units, and our focus is on values of a particular attribute. We have a sample of size n_1 selected from units known to belong to the group of interest, and have observed or measured the attribute of interest for each unit in the sample. Suppose, further, that it is reasonable to assume that we can represent the attribute of interest for these sampling units as a set of independent and identically distributed random variables having normal distributions with expected value μ_1 and variance σ_1^2, $Y_{1,i} \sim$ iid $\mathrm{N}(\mu_1, \sigma_1^2)$. We now have an additional sample of size n_2 sampling units with observed attributes, but do not know whether those units belong to the group of interest or not. Our goal is to determine the group membership of the sampling units from our second sample, based on the observed values that correspond to them.

A prediction procedure developed to control pre-data precision might be developed as follows. Using the group of n_1 observations from the sampling

units known to belong to the group of interest, estimate μ_1 and σ_1^2 using the usual sample mean and variance $\bar{y}_1$ and s_1^2. Let the observed attributes for the second sample with unknown group membership be denoted as $\{y_{2,j};\, j = 1, \ldots, n_2\}$. Compute the inferential quantities $z_j = (y_{2,j} - \bar{y}_1)/\sqrt{\sigma_1^2}$. Declare the sampling unit corresponding to observation $y_{2,j}$ to be a member of the group of interest if $|z_j| \leq t_{1-\alpha/2,n_1-1}$, where $t_{1-\alpha/2,n_1-1}$ is the $1-\alpha/2$ quantile of a $t-$distribution with $n-1$ degrees of freedom. Otherwise, declare the sampling unit corresponding to observation $y_{2,j}$ to not belong to the group of interest. Here, the conclusion to be made is in the form of a decision: the sampling unit corresponding to observation $y_{2,j}$ does or does not belong to the group of interest. The rule is completely specified without any data as long as the planned sample size of n_1 is given. For example, if n_1 is to be 75, then the rule is to declare the sampling unit corresponding to observation $y_{2,j}$ to be a member of the group of interest if $|z_j| \leq 1.9415$. All that remains is to determine values for the appropriate elements of the rule, namely the z_j, based on the values $y_{2,j}$ for the n_2 units of unknown group membership. One procedure developed to attain a high level of post-data precision would be to use the data model of $Y_{1,i} \sim$ iid $\mathrm{N}(\mu_1, \sigma_1^2)$, and assign μ_1 and σ_1^2 prior distributions $\pi_1(\mu_1)$ and $\pi_2(\sigma_1^2)$. We might take $\pi_1(\mu_1)$ to be normal and $\pi_2(\sigma_1^2)$ to be inverse gamma to exploit conditional conjugacy in derivation of a joint posterior $p(\mu_1, \sigma_1^2|\boldsymbol{y}_1)$. From the posterior distribution we will compute a posterior predictive distribution for a new observation y^0,

$$p(y^0|\boldsymbol{y}_1) = \int_0^\infty \int_{-\infty}^\infty f(y^0|\mu, \sigma^2)\, p(\mu, \sigma^2|\boldsymbol{y}_1)\, d\mu\, d\sigma^2.$$

If, for $j = 1, \ldots, n_2$, $y_{2,j}$ falls between the 0.025 and 0.095 quantiles of $p(y^0|\boldsymbol{y}_1)$ we will conclude that membership of the sampling unit corresponding to $y_{2,j}$ in the group of interest cannot be rejected as a plausible reality. Here,

the elements of the rule for drawing conclusions, namely the quantiles of the posterior predictive distribution, cannot be determined without the data $\boldsymbol{y}_1 = \{y_{1,i} : i = 1, \ldots, n_1\}$.

16.4 Decision-Theoretic Procedures

Since we have introduced the concept of evidential measures of support that data provide for potential inferential statements and contrasted that with the concept of making decisions, a clarification is necessary regarding what is known in statistics as *decision theory*. This is, fundamentally, a different use of the word decision than in our connotation of decision processes versus evidential measures in statements of statistical inference. In what is known as statistical decision theory, every part of an analysis is cast as a decision, such as point estimation of parameters. The computation of an estimator for a data model parameter is considered a decision that the parameter is equal to the estimated value. The reason for considering nearly every aspect of an analysis to be a decision is so that the concepts of loss and risk can be formulated quantitatively and used to determine procedures that minimize risk or possibly minimize the maximum possible loss. The incorrectness of decisions is quantified by the loss function, and risk is expected loss. Decision theory can be used to develop entire approaches to statistical analysis, such as Bayesian methods or minimum variance unbiased estimation. The main point is that the phrase *decision theoretic* provides no information as to whether a given procedure has been developed to provide pre-data or post-data precision.

16.5 Logical and Mathematical Bases

There is a theme that runs through comparisons between several schools of thought on both frequentist and Bayesian sides of the coin. That theme concerns the degree of importance attached to the grounding of methodological procedures in mathematics versus logic and reasoning. While the two are not necessarily in conflict, neither does one imply the other.

16.5.1 The Frequentist World

When we discuss frequentist testing procedures we will draw a contrast between the approaches of Neyman and Pearson with that of Fisher. Neyman, in particular, was primarily concerned that statistics develop directly as a mathematical system. He expressed the opinion that what was needed was to "construct a theory of mathematical statistics . . . based solely upon the theory of probability". He then asserted that this could be connected with "the conception of frequency of errors in judgment" (Neyman, 1935, pp. 74-75). The use of data was not to reason, but rather to guide one onto the correct path of a decision tree developed to control or minimize errors in a decision process. With respect to confidence intervals Neyman opined that "The reasoning ended when the functions [used to give endpoints] were calculated" (Neyman, 1941, p.134). Statistical procedures were to be developed from a completely mathematical basis and designed to control the rates of errors in decisions. Fisher, on the other hand, was much more concerned with procedures grounded in logic and reflective of reasoning. Fisher remarked that "there is something horrifying in the ideological movement represented by the doctrine that reasoning, properly speaking, cannot be applied to empirical data to lead to inferences valid in the real world" (Fisher,

1973, p. 7). This is reflected in the presentation of likelihood as a "measure of rational belief when we are reasoning from the sample to the population" (quote taken from Lehmann 1995). Neyman rejected this interpretation of likelihood, while Edwards (1972) embraced it wholeheartedly.

Several topics that we have discussed in this chapter should be evident in the above quotes. First, by attaching a purely mathematical development of statistical procedures to rates of errors in decisions, Neyman was clearly focused on developing procedures with pre-data precision. Also, in talking about reasoning "from the sample to the population", Fisher is promoting inductive reasoning. He had already stated this quite plainly (Fisher, 935b, p. 39),

> . . . the difficult task of making sense of figures is, in fact, essaying a logical process of the kind we call inductive, in that he is attempting to draw inferences from the particular to the general. Such inferences we recognize to be uncertain inferences.

Here, Fisher directly states that his view of making statistical inferences involves inductive argument and inductive statements of conclusion. In fact, in the same paper, Fisher contrasted this need for inductive argument sharply with inferences from "the classical theory of probability" which "are all deductive in character. They are statements about the behavior of individuals, or samples . . . drawn from populations which are fully known". In his later book on statistical methods, having, at least in his own mind, totally rejected the use of inverse probability (something that may have shifted a bit by the end of his career) he states (Fisher, 1973, p. 9-10)

> The rejection of the theory of inverse probability was for a time wrongly taken to imply that we cannot draw, from knowledge

> of a sample, inferences respecting the corresponding population. Such a view would entirely deny validity to all experimental science. What has now appeared is that the mathematical concept of probability is, in most cases, inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability I have used the term "Likelihood".

In this passage, Fisher rejects inverse probability as a method for producing inductive statements of belief (mental confidence) about populations from samples and offers likelihood as a suitable alternative. So as not to give the wrong impression that Fisher had by this point given up on fiducial probability, we point out that he does mention fiducial probability immediately before the above quotation.

Neyman rejected any argument that was not deductive in nature out of hand, and did not believe (proper) inductive arguments existed. Neyman coined the phrase *inductive behavior* to describe the conceptual content of his philosophy, asserting that what is "frequently described as induction" consists of three components, (1) visualization of possible hypotheses, (2) deductions made from these hypotheses, and (3) an action or decision to assume a particular attitude about the hypotheses in (1). Upon presenting these steps, Neyman (1957, p. 11) comments that "These processes are certainly not any sort of "reasoning", at least not in the sense in which this word is used in other instances; they are acts of will". After an extensive argument that examples previously presented by Fisher actually fall within

this framework rather than involving any type of induction he concluded (Neyman, 1957, p. 17)

> It must be obvious that, with the above essential contents on the inductive reasoning approach, its use as a basic principle underlying research is unsatisfactory. The beliefs of particular scientists are a very personal matter and it is useless to attempt to norm them by any dogmatic formula. Furthermore, our actions, while influenced by beliefs, are also motivated by considerations of consequences, that is to say, by considerations of what is desired to be achieved.

And, a paragraph later,

> The content of the concept of inductive behavior is the recognition that the purpose of every piece of serious research is to provide grounds for the selection of one of several contemplated courses of action. Also, the recognition that the desirability of this or that course of action depends on the circumstances and, of course, on the subjective preferences and beliefs of the individual concerned.

One has to wonder if the last sentence of this quote was not basically a preemptive defense maneuver. It would seem that philosophically Neyman and Fisher were worlds apart, and it all began with differential focus on the mathematical versus logical or epistemological underpinnings of statistical procedures.

16.5.2 The Bayesian World

As seemingly with a number of topics, the difference between mathematical and logical bases for statistical inference is less pronounced and less divisive in the Bayesian universe than in the frequentist one. This may be because the posterior probability basis for inferential statements is a unifying force. There are differences in the degree to which prior and posterior probabilities are considered to be totally subjective and personal as opposed to something on which experts should be able to reach a consensus. But these matters seem to involve varying degrees rather than completely different conceptual bases. Nevertheless, there are several schools of thought relative to the topic concerning at least the justification for generally accepted procedures.

Pragmatic Bayes

This school of thought might be said to involve statisticians comfortable with a certain degree of mathematical untidiness in the origins of procedures that seem to work well in practice. These statisticians tend to view inference as consisting entirely of statements of probability based on joint posterior distributions (e.g., Gelman *et al.* 1995). The emphasis is on the analysis of data, rather than justification of Bayesian procedures based on certain optimality criteria. The logical basis is straightforward, $PriorBelief + ObservedData = PosteriorBelief$, which is essentially beyond reproach as long as one accepts the legitimacy of an epistemic concept of probability. In this school of thought, there tends to be less emphasis on testing as a procedure, and Bayes factors in particular. Point and interval estimation, as procedures, are largely replaced by posterior summaries, although means and variances may certainly be a portion of those summaries. Given the rapid development,

both in use and in techniques, of Markov Chain Monte Carlo, this approach to Bayesian inference seems to have largely overwhelmed other approaches. A potential drawback to this pragmatic approach to the justification of inferential procedures, also tied to the methods of MCMC, is that one can get into a bit of a wild-wild-west scenario, meaning that to justify some technique "because it works" presupposes that we understand how to tell "it works". Experts in the area of Markov chain convergence have commented that in the area of new samplers and algorithms for MCMC that we know how to do more than we know works (personal communication, Professor Vivek Roy of Iowa State University and Professor Galen Jones of the University of Minnesota, on separate occasions). Another aspect of this pragmatic school of thought that should perhaps be kept in mind is that its seductive simplicity can sometimes be too simple. In previous portions of this book (Chapter 13) and *Intermediate Statistical Methods* (Chapter 15) we have drawn a distinction in hierarchical models between consideration of data model parameters as random variables assigned a mixing distribution, and fixed parameters for different situations assigned a common first-stage prior. The current approach to developing Bayesian inference doesn't really recognize any distinction and the prescription is to make inference about posterior distributions that are relevant to the problem. As discussed in the material just referenced, this fails to provide guidance in the determination of what posterior distributions are relevant. And, as discussed in Chapter 13, it may be that the most suitable distribution for the purposes of inference is the posterior predictive distribution of data model parameters, which is not simply a portion of the joint posterior distribution of all model quantities given the observations.

Decision-Theorists

In a bit of a contrast to the pragmatic approach just described, some statisticians would like to see, at least to the degree possible, methodology developed from the basis of decision theory (Berger, 1985; O'Hagan, 1994; Robert, 2007, e.g.,). When a decision-theoretic development of a procedure is available, it does lend optimality to the procedure, at least under the loss function used. But, as we will see in Chapter 18, there can be problems in attempts to develop even well established inferential procedures on the basis of decision theory. And, the impact of the loss function chosen is immense. Aside from obvious default choices such as squared error, absolute error, or $0-1$ loss functions, reaching a consensus on an appropriate loss function can be difficult. There has been much written about utility functions as a potential device to solve this problem but, to our knowledge, most such attempts have simply shifted potential controversy about loss functions to potential controversy about utility functions. Nevertheless, attempts to justify inferential procedures on the basis of decision theory can provide a better understanding of how such a procedure operates, and what factors might influence its performance.

Pure Subjectivist Bayes

This school of thought might be called radical or extreme subjectivism but, to avoid emotionally charged labels we will refer to as a pure subjectivist approach. We have already mentioned this approach in Chapter 15 in a discussion of representation theorems. Under this approach, the gold standard is not to develop methods from a common mathematical basis, but to develop an analysis (data model, prior, posterior) for a problem from one common

belief structure (Bernardo and Smith, 1994, e.g.,). All probabilities, including those of the data model, can be interpreted as epistemic probabilities. If the data model can be dictated by exchangeability combined with a representation theorem, and if the prior is a true prior based only on one's beliefs, then making inferential statements based on the posterior is the unassailable correct procedure. There is no need for model assessment, wondering about prior-data model compatibility, or agreement with other external sources of information. One has arrived at the perfect Bayesian package.

Objective Bayes

What is generally referred to as objective Bayes is characterized by the use of a prior distribution that is in some sense *minimally informative* about the parameter of interest (Berger, 2006). As already illustrated with the normal one sample problem, analyses using the method of inverse probability in the nineteenth and early twentieth centuries made use of constant priors and thus could be called objective Bayes, although (Feinberg, 2006) points out constant priors were not used because of any motivation to be somehow scientifically "objective", and the salient point is that it took many years for adherents of inverse probability to move beyond constant priors as an integral part of the use of inverse probability. A good portion of the motivation for objective Bayes seems to have been as a marketing tool, to try and convince scientists that Bayesian analysis is not inherently subjective. In discussion of Berger's 2006 paper he was roundly criticized for this stance (Lad, 2006; O'Hagan, 2006). Berger (2006) does list a number of other potential benefits to the use of objective Bayes, but these are all relative to the practical application of statistics and do not deal directly with the justification of Bayesian

procedures in terms of either a given mathematical system or a given logical system, which is our primary concern.

An integral part of the objective Bayes approach is to formulate prior distributions based on formal rules that can be applied across a wide range of particular problems and models. The rules devised have all attempted to achieve some particular notion of being non-informative that has been adopted by their adherents. This in a sense automates the application of Bayesian methodology, and can be used to address concerns of those that find selection of prior distributions to be too arbitrary. Kass and Wasserman (1996) trace the idea of selecting a prior on the basis of convention or standard of reference to Jeffreys (1939). Jeffreys is also generally viewed as promoting non-informative prior distributions as a reflection of ignorance. He adhered to the Laplacean principle of insufficient reason and thus uniform priors for finite parameter spaces, constant priors when the parameter space was continuous and either bounded or over the entire line, and invariance for parameter spaces that were continuous and unbounded on the right, the latter resulting in what we now call Jeffreys priors. What sometimes gets lost in all of this is that Jeffreys considered these principles to apply to problems involving an "initial" state of knowledge and about which the researcher has "no opinion" about possible values of the parameter. The implication is that there should be additional investigation at which point the investigator is not completely ignorant. This foreshadows an argument in which the use of non-informative or weakly-informative priors serves as a preliminary or initial analysis with the presumption being that for problems of sufficient importance, additional analyses lie down the road. This is not a position that we believe most subjectivists would find objectionable. O'Hagan indicates that the use of weakly-informative priors is also sometimes used as an approximation to more

carefully developed Bayesian analysis, and also comments that (O'Hagan, 2006, p. 4447) "Often, there are just a few parameters about which prior information is substantial and worth formulating carefully; for the remainder, weakly informative priors suffice."

As part of the 2006 discussion of objective versus subjective Bayes, Goldstein (2006) presented an excerpt from the Internet Encyclopedia of Philosophy concerning objective versus subjective reasoning,

> "Objective judgment or belief" refers to a judgment or belief based on objectively strong supporting evidence, the sort of evidence that would be compelling for any rational being. A subjective judgment would then seem to be a judgment or belief supported by evidence that is compelling for some rational beings (subjects) but not compelling for others.

[Objectivity, D.H. Mulder, The Internet Encyclopedia of Philosophy, http://www.iep.utm.edu/. As mentioned, and as partially reviewed in Chapter 7 of *Intermediate Statistical Methods* there have been a number of proposals for developing objective Bayesian priors, including Jeffreys priors (Jeffreys 1961), reference priors (Bernardo, 1979; Berger and Bernardo, 1992), maximum entropy priors (Jaynes, 2003), matching priors (Datta and Mukerjee, 2004), admissible priors (Berger et al., 2005) and a number of others. None of these prescriptions have led to general acceptance, meaning the evidence for any one of them would not qualify as "objective" according to the previous definition. Thus, objective Bayes is not, in and of itself, objective as it involves a subjective decision about which type of non-informative or weakly-informative prior to use, a point also made by O'Hagan (2006).

Arguments about objectivity aside, it certainly seems to be the case that

a good number of solid statistical analyses have been conducted using the types of non-informative prior distributions mentioned. But producing some type of scientific objectivity does not justify those analyses. As discussed in the preceeding material, the use of weakly informative priors are useful in that they should approximate more exacting analyses constructed through painstaking construction of a prior via expert elicitation or exhaustive research to locate addition information about a problem that might serve as the basis for a prior. It is also possible that the use of several weakly-informative priors might produce a sensitivity analysis with more meaning than simply varying the parameters of a fixed prior.

Subjective Bayes

Quite distinct from what we have termed pure subjectivist Bayes, the approach embodied under this heading involves the incorporation of information from scientific understanding or consensus, previous studies, and other sources to construct informative prior distributions, but there is very little else that unifies what the phrase subjective Bayes refers to. We have mentioned the curiosity that Feinberg (2006) expressed in terms of how long it took the inverse probability argument to move beyond the use of constant prior distributions. He also indicates that early attempts to do so included arguments by Edgeworth and Karl Pearson that prior distributions need to be based on actual prior data. In our minds, that might be a better definition of objective Bayes than what is in current use, but the ship has sailed on that nomenclature. Goldstein (2006) argues extensively in favor of what he calls subjective Bayes, and much of the material presented is targeted at the need for careful statistical analysis of important problems and incor-

porating relevant information from all possible sources. Examples included assessment of computer simulation models and climate change, and we could add to the list any number of problems involving the reliability of physical systems (Hamada, Wilson, and Martz, 2008). Methods for developing priors include the use of similar previous studies, scientific plausibility and, importantly, prior elicitation from experts. Goldstein (2006) makes a strong case that the process of scientific reasoning should impact formulation of a problem, including drawing a distinction between the analysis of a set of data in a vacuum, and analysis of a set of data within the context of a larger scientific analysis of a system or problem. For the later, the process of prior formulation is a difficult task, potentially requiring additional research into related problems or seeking consensus among subject matter experts. For the former, there is more freedom and less burden on prior specification but there is also an accompanying realization that the purpose of such an analysis "is to provide information which will be helpful at some future time for whoever does attempt to address the real questions of interest" (Goldstein, 2006, p.411).

There is one topic included in (Goldstein, 2006) that does strike at the heart of the types of issues we are primarily concerned with in this section, and that involves the question of whether posterior knowledge, even posterior probability, is necessarily in the form of a conditional probability. This issue is a direct consequence of questioning whether updating prior knowledge by conditioning through Bayes theorem is a necessary part of the subjectivist position. The question arises from broadening the view from producing estimates for a given set of data to the analysis of a scientific question. Bayes theorem is the correct way to update beliefs based only on current beliefs and the next set of observed data. But (Goldstein, 2006, p. 414) argues,

> By the time you observe the data, you may have come across further unanticipated but relevant information (or you may not, and this is also relevant information), and you may well have further general insights about the problem, by study of relevant literature, deeper mathematical treatment or careful data analysis. None of this corresponds to Bayesian conditioning.

A version of posterior probability that is not the traditional conditional probability produced by Bayes theorem can be developed as follows (Goldstein, 1997). Consider an event A of interest, and two other events B and B^c. The conditional probabilities produced by Bayes theorem are $Pr(A|B)$ and $Pr(A|B^c)$. Under something called the *temporal sure preference* principle, which basically says if you know you will prefer E to F in the future you should not prefer F to E now, a stochastic relationship between the conditional probabilities and a posterior for A at a future time t can be given as,

$$Pr_t(A) = Pr(A|B)I(B) + Pr(A|B^c)I(B^c) + R, \tag{16.3}$$

where $I(\cdot)$ is the usual indicator function and R is a random quantity such that,

$$E(R) = E(R|B) = E(R|B^c) = 0.$$

Note that the posterior $P_t(A)$ is not a conditional probability. In fact, (16.3) can be viewed as a mean $[Pr(A|B)I(B) + Pr(A|B^c)I(B^c)]$ plus uncertainty (R).

The critical aspect of the above development is not necessarily its direct practical application, but the fact that it is not conditional. And the importance of that is that then probability and, in particular, conditional probability, is no longer a necessary primitive for a Bayesian analysis. An

approach to Bayesian analysis that takes expectation to be the primitive is called *Bayes linear inference*. Suppose that $\boldsymbol{Y}_t$ is a vector of random quantities that will be observed by time t. Then, given temporal sure preference, your current opinion about your posterior expectation for $\boldsymbol{\theta}$ at time $t\text{cov}(R, \boldsymbol{Y})$ must satisfy,

$$\begin{aligned} \boldsymbol{\theta} &= E_t(\boldsymbol{\theta}) + S \\ E_t(\boldsymbol{\theta}) &= E_{\boldsymbol{Y}}(\boldsymbol{\theta}) + R \\ E_{\boldsymbol{Y}}(\boldsymbol{\theta}) &= E(\boldsymbol{\theta}) + \text{cov}(\boldsymbol{Y}, \boldsymbol{\theta})[\text{var}(\boldsymbol{Y})]^{-1}[\boldsymbol{Y} - E(\boldsymbol{Y})], \end{aligned}$$

and where S and R are random quantities such that

$$E(S) = E(R) = \text{cov}(S, \boldsymbol{Y}) = \text{cov}(R, \boldsymbol{Y}) = \text{cov}(S, R) = 0.$$

In the preceeding, $E_{\boldsymbol{Y}}(\boldsymbol{\theta})$ is called the linear Bayes mean for $\boldsymbol{\theta}$. Thus, updating of belief does not result from updating probabilities but from updating expectations, variances, and covariances. This approach may be particularly attractive in problems that do involve the elicitation of expert opinion or summaries of information from previous studies, such as the previously mentioned complex problems of assessing computer simulation models, climate models, and in problems of Bayesian reliability.

Chapter 17

Testing Hypotheses

Few areas of statistical analysis have been as controversial as the testing of hypotheses, beginning at least around 1900 and continuing to today. It is the belief of a number of statisticians (including the authors of these notes) that a portion of the arguments about hypothesis tests are misplaced because the evolution of testing procedures is not well understood. There are also some deeper philosophical issues that are important, but simply understanding the types of problems that are well approached by different versions of hypothesis tests will go a long way toward providing a clearer picture of the topic. This section is an attempt to present the topic of testing hypotheses in a clear and understandable manner, but may take some liberties with technical or absolute philosophical concreteness to do so. Our intent is to provide connections between philosophical foundations and their practical implications, which we believe have been ignored or misrepresented in much of the material a student in statistics is typically exposed to.

17.1 The Same Hypothesis Doesn't Always Have the Same Role

The first step to understanding tests of hypotheses is recognizing that the same mathematical (or statistical) hypothesis can have different uses, depending on the intent and context of the test procedure being conducted. So a hypothesis such as $\mu = \mu_0$, encountered in an introductory course in statistics, might be used for different purposes within the context of different problems.

Example 17.1: A manufacturer of body armor for police uses a specification for the average thickness of Kevlar-type material in its vests that is given by the value μ_0. For each manufactured lot of vests, a random sample is selected and the thickness of the material is carefully measured. Based on values measured in the random sample, a test is conducted for the hypotheses

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

If H_0 is rejected in favor of H_1 in this test, the entire lot of vests is not shipped. If H_0 is accepted, then the lot of vests is shipped for use. The manufacturer wants to reject no more than 10% of lots that are actually ok (that is, for which we should proceed as if H_0 is true and ship) and wants to be able to detect a difference from μ_0 of 2mm (thereby proceeding as if H_1 is true and not shipping) at least 80% of the time when the true μ differs from μ_0 by this much.

The role of the hypotheses H_0 and H_1 in this example are to provide two alternatives, and we will behave as if one of the two is true. If we ship the lot of vests we are behaving as if H_0 is true. If we do not ship the lot

we are behaving as if H_1 is true. Despite this, the two hypotheses do not play entirely symmetric roles in the test procedure, because H_0 is a specific value for μ, while H_1 is not specific. Notice that we may recognize two types of mistakes we could make in deciding how to behave. We could reject a particular lot of vests that should really be shipped, a mistake we want to make no more than 10% of the time, or we could ship a lot of vests that we shouldn't. How often we make this second type of mistake depends on how far away from H_0 the truth is, and we must specify that before we can attempt to even compute the rate for such mistakes, let alone try to control it. That is, we must essentially replace the vague H_1 with another hypothesis that is just as specific as H_0. In this example we have specified that we want to commit this type of mistake no more than 20% of the time if H_0 is violated by 2mm of thickness. At the same time, we want a test procedure such that, if H_0 is violated within the acceptable level of tolerance, then the chances of detecting that departure would be small.

One might wonder, in this example, why the hypotheses are stated in the way they are, rather than $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$. After all, the intention is to avoid shipping dangerous body armor. If the protective material is thicker than needed, then why reject a manufactured lot for that reason? The answer is that manufacturing processes are designed to have a certain tolerance around given specifications. We want to detect departures from that specification given in H_0 regardless of which direction they are in because such a departure indicates a malfunction in the manufacturing process – things are not working as they were designed to work and are in need of modification.

Example 17.2: A veterinary scientist is trying to develop an improved vac-

cine against Porcine Reproductive and Respiratory Syndrome (PRRS) which infects adult female sows but affects newborn pigs. The research has produced a new method for culturing antibodies to combat the PRRS virus, and a set of identically constructed petri dishes is used to provide replicates of this culture technique. Each petri dish is then inoculated with a certain level of PRRS virus and, after a suitable time, the level of remaining viral activity is measured. The culture technique used to produce the vaccine in common use at the current time is believed to result in an average level μ_0 of residual viral activity. Thus, the set of petri dishes is used to test

$$H : \mu \leq \mu_0.$$

If H is rejected, the method of culturing antibodies is retained for further investigation and the research proceeds to the next step, possibly a trial at the organismal level. If H is not rejected, the search for an improved vaccine goes "back to the drawing board". The role of H in this example is to provide a condition, the rejection of which would indicate scientific progress. We really have no interest in behaving as if H is true and so we would never accept H. If we fail to reject H we simply take no further particular action at all. We have failed in our search for an improvement *with the experiment as it was conducted*. Perhaps H is still false, but we have not provided evidence to that effect. We are not concerned with the rate of any type of error, because the experiment will not be conducted over and over with some decision to be reached each time. We have either rejected or failed to reject H this time, and that is all that matters. If we fail to reject H further scientific investigation might result in a seemingly minor modification to the culture technique, or perhaps even a modification to the process of observation, and the new situation could lead to a new test under those circumstances. Alternatively,

the scientist could completely abandon the antibody culture technique under investigation and move on to other ideas. These are concerns of veterinary science, not statistical analysis.

While subtle, the situations of Example 17.1 and Example 17.2 are actually quite distinct in context, which will be explored further in what follows. A key difference that should be clear even at this point is that the test of Example 17.1 is intended to be conducted many times, once for each lot of manufactured vests, and a particular action results from each repetition of the test. In contrast, the test of Example 17.2 will most likely be conducted only one time and the result deemed either a success or a failure.

17.2 Tests of Significance

Tests of significance were developed by R.A. Fisher beginning in the 1920s and perhaps first laid out in general form in (Fisher, 1925). Under this theory of testing, there is a given hypothesis H and our objective is to determine whether a departure from H in the data could be due to random variability or is indicative of a systematic effect. We consider here primarily hypotheses that concern the value of a parameter. To this end, let the hypothesis to be tested be denoted as $H : \theta = \theta_0$. Suppose we have identified a test statistic $T^*(\boldsymbol{Y}, \theta_0)$, the distribution of which does not depend on any parameters and is known if $\theta = \theta_0$ (that is, if H is true). Do not get T^* in this paragraph confused with a random variable that has a $t-$distribution; that may or may not be the case. Here, T^* simply stands for "test statistic". Let the distribution of T^* under H be denoted as $h(t)$. In a test of significance, we want to compare the observed test statistic $t^* = t(\boldsymbol{y}, \theta_0)$ to $h(t)$. If $t^* = t(\boldsymbol{y}, \theta_0)$ is in a region of high or moderate probability according to $h(t)$,

then we have little evidence that H is incorrect and we fail to reject it. If, on the other hand, t^* is in a region of small probability according to $h(t)$, then we do have some evidence that H is incorrect. The degree of evidence *against* H is quantified by the $p - value$.

Definition: The $p-$value for a test of $H : \theta = \theta_0$ based on a test statistic T^* with reference distribution $h(t)$ is,

$$\begin{aligned} p &= Pr\left[\{-|t^*| \geq T^*(\boldsymbol{Y}, \theta_0)\} \cup \{|t^*| \leq T^*(\boldsymbol{Y}, \theta_0)\} \mid \theta_0\right] \\ &= \int_{-\infty}^{-|t^*|} h(u)\, du + \int_{|t^*|}^{\infty} h(u)\, du. \end{aligned}$$

Notice that for classical test statistics used in testing means, the reference distributions are symmetric about 0 so that the $p-$value can be computed as

$$p = 2 \int_{|t^*|}^{\infty} h(u)\, du.$$

The $p-$value is a *tail area*, and serves as an inverse measure of evidence against H. The smaller the value of p in a test, the more "significant" the test is said to be, and hence the name of this procedure as a test of significance. A common interpretation is that a $p-$value represents the probability of obtaining a test statistic as or more extreme than that we actually obtained, if the hypothesis is true. If that probability is small enough then either (i) the hypothesis is true and something extremely unusual has occurred or (ii) the hypothesis is not true. Since we do not generally believe that our sample is highly unusual, we reject the hypothesis. We can never be absolutely certain the hypothesis is not true, but if our evidence against it is strong we reject it as being plausible.

So how small a $p-$value is small enough to reject the hypothesis? Before computers were available, such as in the 1930s, reference distributions needed

to be tabled at certain quantiles. It became popular to use quantiles 0.90, 0.95, 0.975, 0.995, and sometimes 0.999. A convention developed to call tests with $0.01 < p \leq 0.05$ "significant" and tests with $p \leq 0.001$ "highly significant". Some authors expanded this to include $0.05 < p \leq 0.10$ as "marginally significant". There is nothing magical about these types of rating systems, either this one or others. They are really an artifact of the need to produce tables of reference distributions rather than having a computer calculate an exact $p-$value. Although you still see some applied journals that use this type of significance level ranking system, it is now more common to simply report the actual $p-$value. How small is small enough to declare significance still means $p \leq 0.05$ more often than any other value, but it would be hard to argue that an exact value of $p = 0.0501$ is really less evidence against H than $p = 0.0499$.

Example 17.3: Suppose that $Y_1, \ldots, Y_N$ are a random sample from a normal distribution with expected value μ and variance σ^2 and the resulting sample values based on $n = 25$ observations are $\bar{y} = 9.573$ and $s_u^2 = 14.453$. A test of $H : \mu = 8$ results in the exact theory statistic $t = 2.069$ and corresponding $p-$value $p = 0.0495$.

The amount of evidence against the hypothesis increases as the $p-$value gets smaller. Consider testing $H : \mu = \mu_0$ based on either a normal random sample or a general random sample. A T test statistic or a Z test statistic, depending on the situation, results from

$$T = \frac{\bar{Y} - \mu_0}{(S_u^2/n)^{1/2}}$$
$$Z = \frac{\bar{Y} - \mu_0}{(S^2/n)^{1/2}}$$

With other quantities in these statistics held fixed, the $p-$value will decrease

if

1. The distance between $\bar{Y}$ and μ_0 increases. That is, if $|\bar{Y} - \mu_0|$ increases.

2. The variance S_u^2 or S^2 decreases.

3. The sample size n increases.

Example 17.4: Suppose in Example 17.3 the sample size had been $n = 50$ instead of $n = 25$. The test statistic would have been $t = 2.926$ with an associated $p = 0.0052$. Note that in this exact test the sample size n has influenced both the test statistic and the reference distribution, changing the degrees of freedom to $n-1 = 49$ from the value in Example 17.3 of $n-1 = 24$.

Fisher never voiced concerns that a scientist would be unable to determine what an appropriate test statistic is for a given situation. As indicated by Lehmann (1995, p. 1245) Fisher commented that " . . . the same data may contradict the hypothesis in any number of ways", followed by "The notion that different tests of significance are appropriate to test different features of the same null hypothesis presents no difficulty to workers engaged in practical experimentation, but has been the occasion of much theoretical discussion among statisticians." The theory of Neyman and Pearson, in contrast, actually starts with the search for test statistics that possess optimal properties. E. Pearson relates that some of the original motivation for development of the Neyman-Pearson approach to testing was dissatisfaction with justifications Fisher gave for the use of various test statistics in particular problems (Pearson, 1955, p. 204-205). Despite this, however, it is not the form of test statistics that separates the two philosophies, particularly as regards tests of location or mean separation among groups.

17.3 Hypothesis Tests for Acceptance Sampling

The body armor problem of Example 17.1 is an example of a general situation often called *Acceptance Sampling*. The statisticians Jerzy Neymann and Egon Pearson developed a theory of hypothesis testing well suited for this general situation. In this approach to testing hypotheses we formulate both a *null* and an *alternative* hypothesis, denoted H_0 (null) and H_1 (alternative). These hypotheses typically form a logical disjunction, meaning they constitute a *partition* of the parameter space Θ, and they have the forms,

$$H_0: \quad \theta = \theta_0$$
$$H_1: \quad \theta \neq \theta_0$$

with the implication in both cases that $\theta \in \Theta$.

The Neyman-Pearson approach to testing hypotheses is based on the need to make a decision – we will either behave as if H_0 is true or as if H_1 is true. Then there are two mistakes, or errors, that can be made, called Type I and Type II errors:

	Truth	
Decision	H_0	H_1
H_0	No Error	Type II Error
H_1	Type I Error	No Error

It turns out that statistically it is much easier to control the level of Type I error than the level of Type II error, and Neyman called Type I error the "error which is the more important to avoid" (Neyman, 1977, p. 104). Thus, when setting up a test procedure, we generally specify an acceptable level

of Type I error to develop the test. We then (at least should) compute the level of Type II error that will occur for some meaningful departure of θ from the hypothesized $H_0 : \theta = \theta_0$ and, if it is larger than acceptable, adjust the procedure (almost always through sample size). Neyman-Pearson hypothesis tests typically use the same test statistics as Fisher's tests of significance, so the distinction between these procedures in in terms of the appropriate framework for inference, not the mathematics of computing statistics used in the procedures.

Using the same general notation as previously, let $H_0 : \theta = \theta_0$ be the null hypothesis and $H_1 : \theta \neq \theta_0$ be the alternative hypothesis to be tested. Suppose that an appropriate test statistic $T^*(\boldsymbol{Y}, \theta_0)$ with reference distribution $h(t)$ has been selected. To compare an observed value of $t^* = t^*(\boldsymbol{y}, \theta_0)$ with $h(t)$ we use the following procedure:

1. Determine the greatest level of Type I error that is acceptable, and denote it as α.

2. Using α, determine *critical values* for the test statistic $c1$ and c_2 as,

$$\begin{aligned} Pr(T^*(\boldsymbol{Y}, \theta_0) \leq c_1 | \, \theta_0) &= \alpha/2 \\ Pr(T^*(\boldsymbol{Y}, \theta_0 > c_2 | \, \theta_0) &= \alpha/2 \end{aligned}$$

3. If the computed test statistic t^* is less than c_1 or greater than c_2, then we accept the alternative hypothesis H_1. If the computed test statistic is such that $c_1 < t^* < c_2$ then we accept the null hypothesis H_0.

Similarly to tests of significance (because they use the same test statistics and the same reference distributions), in the cases for which θ is a mean and the reference distributions are symmetric about 0, the critical values are also symmetric about 0 and $c_1 = -c_2$.

Note that there is no notion of some test statistics being "more significant" than others here (but see Chapter 17.10.2). Any statistic greater than c_2, for example always leads to the same decision, regardless of whether it is $c_2+0.1$ or c_2+10. In discussion of tests of significance, we contrasted $p-$values of 0.4999 and 0.0501, indicating that they provide essentially the same amount of evidence against a hypothesis H (which here would be the null hypothesis H_0). The same phenomenon does not occur with critical values. If the test statistic is $t^* = c_2 - 0.001$ we accept H_0. If the test statistic is $t^* = c_2 + 0.001$ we accept H_1. This is because in acceptance sampling we are concerned with *error rates* and the way to control the Type I error rate is to make opposite decisions with a dividing line at c_2 (or c_1 and c_2). It doesn't matter how close to that line we may be, we are either on one side or the other.

Example 17.5: Suppose that $Y_1, \ldots, Y_n$ are a random sample from a normal distribution with expected value μ and variance σ^2, and the resulting sample values based on $n = 15$ observations are $\bar{y} = -22.6$ and $s_u^2 = 4.9$. A test of $H_0 : \mu = -23.0$ versus $H_1 : \mu \neq -23.0$ gives a T test statistic of $t^* = -0.6999$. With a specified Type I error rate of $\alpha = 0.05$, critical values are

$$\begin{aligned} c_1 &= T_{0.025;14} = -2.1448 \\ c_2 &= T_{0.975;14} = 2.1448 \end{aligned}$$

Since $-2.1448 < t^* < 2.1448$ we would accept $H_0 : \mu = -23.0$.

Now, to examine the Type II error rate we need to specify how far some possible true value of θ is from the hypothesized value of θ_0. This is because the Type II error rate is

$$\beta(\theta) = Pr(c_1 < T^*(\boldsymbol{Y}, \theta_0) < c_2 \,|\, \theta),$$

where c_1 and c_2 are determined under the assumption that $\theta = \theta_0$.

Notice that the Type II error rate is a function of the true value of θ, which we do not know. What is called the *Power* of the test is the probability of accepting H_1 if the true parameter value is $\theta \neq \theta_0$, Power$(\theta) = 1 - \beta(\theta)$. In selection of a test procedure, it is typical is to pick a range of values of $\{\theta = \theta_0 \pm k\delta : k = 1, \ldots, K\}$ and then compute power for each of those values. For the test procedures we have discussed, power is an increasing function of the difference between the hypothesized value θ_0 and the true but unknown value θ. Thus, if power is acceptable for a distance of 3δ, say, it will also be acceptable for any greater distance $s\delta$ for $s > k$.

Suppose that we are dealing with a one sample normal problem and a hypothesis about the mean, which implies use of the test statistic

$$T(\boldsymbol{Y}, \mu_0) = \frac{\bar{Y} - \mu_0}{(S_u^2/n)^{1/2}}.$$

If $\mu = \mu_0$ then T has a $t-$distribution with $n-1$ degrees of freedom and it is under this assumption that we compute the critical vlaues c_1 and c_2. But if $\mu = \mu_0 + \delta$ for some δ, then it is actually the quantity

$$T(\boldsymbol{Y}, \mu_0 + \delta) = \frac{\bar{Y} - \mu_0 - \delta}{(S_u^2/n)^{1/2}}$$

that has a $t-$distribution with $n-1$ degrees of freedom. Now, the probability of a Type II error is

$$\begin{aligned}
\beta(\mu_0 + \delta) &= Pr\left[c_1 < T(\boldsymbol{Y}, \mu_0) < c_2\right] \\
&= Pr\left[c_1 < \frac{\bar{Y} - \mu_0}{(S_u^2/n)^{1/2}} < c_2\right] \\
&= Pr\left[c_1 - \frac{\delta}{(S_u^2/n)^{1/2}} < -\frac{\bar{Y} - \mu_0 - \delta}{(S_u^2/n)^{1/2}} < c_2 - \frac{\delta}{(S_u^2/n)^{1/2}}\right] \\
&= Pr\left[c_1 - \frac{\delta}{(S_u^2/n)^{1/2}} < T(\boldsymbol{Y}, \mu_0 + \delta) < c_2 - \frac{\delta}{(S_u^2/n)^{1/2}}\right]
\end{aligned}$$

To compute a numerical value for $\beta(\delta)$ requires having values to plug in for

all quantities other than μ_0, which is given in H_0.

Example 17.6: Consider again the situation of Example 17.5. Suppose we want to be able to detect a difference from $\mu = -23.0$ of 0.5 with probability 0.80 or greater. Then we want a Type II error rate of no greater than 0.20 if $\mu = 23 + (-0.05) = -23.5$ or $\mu = -23 + 0.05 = -22.5$. Beginning with the last line of the previous expression and $\delta = 0.05$,

$$\begin{aligned}
\beta(-22.5) &= Pr\left[c_1 - \frac{\delta}{(S_u^2/n)^{1/2}} < T(\boldsymbol{Y}, \mu_0 + \delta) < c_2 - \frac{\delta}{(S_u^2/n)^{1/2}}\right] \\
&= Pr\left(-2.1448 - \frac{0.5}{(4.9/15)^{1/2}} < T(\boldsymbol{Y}, \mu_0 + 0.05) < 2.1448 - \frac{0.5}{(4.9/15)^{1/2}}\right) \\
&= Pr\left[-3.0196 < T(\boldsymbol{Y}, \mu_0 + \delta) < 1.2700\right]
\end{aligned}$$

where $T(\boldsymbol{Y}, \mu_0 + \delta)$ follows a $t-$distribution with $n - 1 = 14$ degrees of freedom. Then

$$\beta(-22.5) = 0.8830.$$

Similarly, if $\delta = -0.5$,

$$\beta(-23.5) = Pr\left[-1.2700 < T(\boldsymbol{Y}, \mu_0 + \delta) < 3.0196\right] = 0.8830.$$

Then the chances of detecting a difference between the hypothesized value of $\mu_0 = -23.0$ and either $\mu = \mu_0 + 0.5 = -22.5$ or $\mu = \mu_0 - 0.05 = -23.5$, is $1 - 0.8830 = 0.1170$, the power of the test. Notice that this is quite low power.

The only factors under our complete control that could increase the power of the test in Example 17.6 are the specified level for Type I error α and the sample size n. Increasing either α or n will increase power, but we assume that α was set to a given value for some reason and we don't want to

change it. A traditional use of power calculations, then, is to determine the sample size needed to achieve an acceptable power for a certain difference from the hypothesized value of the parameter in H_0. In the case of testing for the value of a mean, power calculations are computed for a range of values $\mu + \mu_0 \pm k\delta : k = 1, \ldots, K\}$. A plot of power versus $k\delta$ is then called a *power curve*. A complication in power calculations with exact tests is that n influences both the magnitude of the denominator of the test statistic and the reference distribution (through degrees of freedom). Because of this, in conducting such calculations it is common to simplify the test structure by assuming that σ^2 is known, in which case the test statistic becomes

$$Z = \frac{\bar{Y} - \mu_0}{(\sigma^2/n)^{1/2}},$$

and the reference distribution is then standard normal. This is often used even if the test to be employed in the study proper (i.e., to test H_0) will be based on exact theory and the reference distribution will be a $t-$distribution.

Example 17.7: As a continuation of Examples 17.5 and 17.6, suppose that we want to conduct the test with $\alpha = 0.05$ and would like power of at least 0.75 for $\delta = 0.5$. We will assume that $\sigma^2 = 5.0$. To determine how large n would need to be for this, note first that critical values become $c_1 = -1.96$ and $c_2 = 1.96$ regardless of the value of n. Then, from previous derivations,

$$\begin{aligned} \beta(\mu_0 + \delta) &= Pr\left[-1.96 + \frac{\delta}{(5.0/n)^{1/2}} < \frac{\bar{Y} - \mu_0 + \delta}{(\sigma^2/n)^{1/2}} < 1.96 + \frac{\delta}{(5.0/n)^{1/2}}\right] \\ &= Pr\left[-1.96 + \frac{\delta}{(5.0/n)^{1/2}} < Z < 1.96 + \frac{\delta}{(5.0/n)^{1/2}}\right] \end{aligned}$$

Setting $\beta(\mu_0 + \delta) = 0.75$ and $\delta = 0.5$ implies that we desire

$$\begin{aligned} 0.375 &= Pr\left[-1.96 + \frac{0.5}{(5.0/n)^{1/2}} < Z\right] \\ 0.375 &= Pr\left[Z < 1.96 + \frac{0.5}{(5.0/n)^{1/2}}\right] \end{aligned}$$

The 0.375 quantile of a standard normal distribution is -0.3186, so then either of these gives

$$-0.3186 = 1.96 + \frac{0.5}{(5.0/n)^{1/2}}$$

which implies that

$$n = \frac{5(-1.96 - 0.3186)^2}{0.5} = 51.9$$

so that we would need a sample size of at least $n = 52$ to achieve the power we want to detect a difference between $\mu_0 = -23.0$ and either $\mu = -22.5$ or $\mu = -23.5$.

17.4 Testing Confusion and a Way Past It

17.4.1 Points of Confusion

We have presented Fisher's approach to Tests of Significance as both conceptually and operationally distinct from the Neyman-Pearson approach to Hypothesis Tests for Acceptance Sampling, and we believe this is the best way to keep ideas clear and to be certain we understand what statistical tests are all about. But, as mentioned in the introductory paragraph to this chapter, there is a great deal of controversy and discussion surrounding the topic of statistical tests of hypotheses. And most, if not all, books on introductory statistical theory and methods present a mix of ideas without

any distinction between the approaches, further confusing the topic. Why is this? There are different answers offered by different statisticians, and some of the issues involved can become quite subtle and sophisticated. We will not attempt to cover all aspects of why tests of hypotheses has been perhaps the most abused topic in all of statistics, but will point out a few issues that make it easy to confuse the approaches of Fisher's Tests of Significance and the Neyman-Pearson Theory of Tests for Acceptance Sampling.

1. Duality of $p-$values and Critical Values.

   One thing that has promoted sloppiness in maintaining the distinction between tests of significance and tests for acceptance sampling is that it is easy (and perfectly acceptable) to formulate critical values for acceptance sampling in terms of $p-$values. If we desire a test with critical values such that $Pr(\text{Type I error}) = \alpha$, then we can express those critical values on the scale of the test statistic, as we have done in discussion of the Neyman-Pearson approach, or we can reach the same conclusions by first computing a $p-$value and then using a decision rule stated as,

   $$\begin{aligned} &\text{Accept } H_0 \quad \text{if} \quad p > \alpha \\ &\text{Accept } H_1 \quad \text{if} \quad p \leq \alpha. \end{aligned}$$

   Interpretation of the $p-$value as a *quantitative* measure of evidence against H_0 is irrelevant in acceptance sampling, because a $p-$value of 0.04 and a $p-$value of 0.001 would lead to exactly the same conclusion if $\alpha = 0.05$. There have been arguments about whether the size of a $p-$value is meaningful on a continuous scale, some claiming yes and others no. But almost all of these arguments have taken place under

the assumption there is *one answer* that fits all situations. In contrast, we suggest here that the answer under the approach of tests of significance is yes, the size of the p-value is meaningful, while the answer under the approach of tests for acceptance sampling is no, the size of the p-value is not meaningful aside from whether it is smaller or greater than the specified value of α.

2. One Hypothesis Versus Two Hypotheses.

   We have presented tests of significance in the form developed by Fisher, in which there is only one hypothesis H . To make scientific progress the objective is to reject H. But the versions of H used in most tests are in exactly the same form as the null hypotheses used in tests for acceptance sampling, H_0. In later writings, even Fisher called his hypothesis a null hypothesis. So the hypothesis of Significance Testing is the same as the null hypothesis of Tests for Acceptance Sampling. The claim of Tests of Significance is that there is no alternative hypothesis. But if one rejects $H_0 : \mu = \mu_0$ is not that the same as accepting $H_1 : \mu \neq \mu_0$? It would seem that Tests of Significance have an *implied* alternative hypothesis which, in many and perhaps even most cases, would result in the same H_0 and H_1 used in Tests for Acceptance Sampling. This has been, in fact, a criticism of Fisher's approach, both historically and today. There is a counter-argument we will pursue in Chapter 17.7, but our point at this stage is just that this is an issue that serves to blur the line between the approaches we have presented as distinct procedures. It is important because, if there is no alternative hypothesis H_1, then there can be no Type II error, and both Type

I and Type II errors are central to the behavioral procedure involved in tests for acceptance sampling. It can be argued that the issue of existence of implied alternative hypotheses, and thus two types of errors that can be committed, is something of a red herring. It is not so much the existence or non-existence of these potential errors that is key to the difference in testing approaches, it is whether there exists an *error rate* that is important.

3. The Use of Common Thresholds.

   In tests of significance we may want to judge a $p-$value against some standard thresholds, and 0.10, 0.05 and 0.01 have been the most popular, used by Fisher and everyone since. In tests for acceptance sampling we set an acceptable level of Type I error, and the most popular values have been 0.10, 0.05 and 0.01. This, compounded by the duality between $p-$values and critical values already mentioned, further obscures the distinction between the two approaches.

4. Muddled Phraseology.

   In a Test of Significance we would never reach a conclusion of accepting the hypothesis. If we do not reject the hypothesis we simply *fail to reject* it. This phrase has become a part of the standard presentation of testing hypotheses, regardless of whether the approach being presented is Tests of Significance, Tests for Acceptance Sampling, or some hybrid (usually the latter in most textbooks, as already mentioned). One can find in any number of texts that the decision is between *rejecting* H_0 and *failing to reject* H_0. The implication is that, even under an approach connected with acceptance sampling, it is acceptable to equate rejecting H_0 with *accepting* H_1 but it is not acceptable to equate *failing*

to reject H_0 with *accepting* H_0. This is true despite the fact that one of the developers of Tests for Acceptance Sampling, J. Neyman, wrote repeatedly of "accepting the null hypothesis" and the entire approach is based on making a decision to behave as if either H_0 is true or as if H_1 is true. This is simply another failure to maintain a distinction between approaches that further muddies the water. That is, it is an attempt to apply a notion that is correctly involved with Tests of Significance to Tests for Acceptance Sampling, where it is not a valid notion.

17.4.2 Avoiding Confusion

Despite a lot of overlap in computational process, and some overlap in conceptual basis, it is not all that difficult to keep the approaches of Tests of Significance and Tests for Acceptance Sampling distinct and clear. The way to do this is to recall the very fundamental intentions of the approaches:

- Tests of Significance have the goal of providing a quantification of the evidence against a hypothesis that has resulted from a particular study.
- Tests for Acceptance Sampling have the goal of providing a decision rule for choosing between two hypotheses, in such a way that we can control the rate of Type I errors and achieve acceptable power for meaningful alternatives.

Operationally, we may be able to obtain evidence against a hypothesis from a single test conducted with data from a single study, but we can only control the *rate* of making mistakes if we conduct the same test repeatedly with similar data from repeated studies.

These characteristics of the two approaches to testing hypotheses help us

avoid confusing one for the other. In scientific investigation it is unusual (not unheard of, but unusual) for a study to be repeated in exactly the same way many times. What we desire is a measure of evidence against a hypothesis we would like to reject. In a manufacturing setting, on the other hand, we usually will repeat the same test with similar data many times, where in each case a decision to take some action is needed. These two settings define endpoints of a spectrum, or a gradient, along which most situations that require a test of hypotheses lie. By considering where on that gradient a particular situation might be located we can often make a reasonable decision about which approach to testing hypotheses might be more appropriate.

Example 17.8: A large, highly automated, pasta factory produces ten thousand packages of lasagne noodles each day. These packages are packed into lots of 100 packages each for distribution and each package contains a label that says "16 oz". Before shipment of the day's production, each lot has a sample of 2 packages removed from it which are weighed. From the 200 sampled packages data are obtained to test whether the mean weight is 16 oz or differs from 16 oz by a meaningful amount, which the company has determined is 0.25 oz.

This situation fits the tests for acceptance sampling approach nearly perfectly. We would like to test $H_0 : \mu = 16$ versus $H_1 : \mu \neq 16$ and would like to be able to detect a true difference from 16 of $\delta = 0.25$ oz. The sample size of $n = 200$ is large enough we could justify use of a test based on approximate results and use a test statistic that has a standard normal reference distribution. We could compute power or determine the sample size needed to detect a difference from the nominal value (16) of 0.25 with some high power. Similarly to Example 17.1 on testing thickness of kevlar in body

armor, we use a two-sided alternative because our goal is to determine when the manufacturing process is out of tolerance or control.

Example 17.9: A scientist at a large company believes she has discovered a material that, when used in water filters, will remove cyanobacteria (toxic algae) more effectively than a currently used material. She asks the manufacturing division of the company to construct 20 water filters using each type of material, and plans to filter water purposely contaminated with $10\mu g/L$ of cyanobacteria. The concentration of cyanobacteria in the filtrate is the response of interest.

This situation is certainly more closely related to tests of significance than to tests for acceptance sampling. There is no indication that the study will be repeated at all, let alone a large number of times. The intention is to determine whether use of the new material is worthy of further pursuit, dealing with any number of issues beyond effectiveness such as cost, durability, and so forth. The hypothesis will most likely be of the form $H : \mu_1 = \mu_2$.

Example 17.10: A plant pathologist conducts a study to determine the amount of yield loss that is caused by a particular soybean disease. At each of four research farms, 20 plots are planted with a given variety of soybeans. At a certain time in the growing season, 10 of the plots are inoculated with the disease under study. At the end of the growing season the plots are harvested and the yield (in km/ha) is measured for each plot.

This example again represents a situation more in line with tests of significance than tests for acceptance sampling. While the study involves four research farms and there may be a test of significance conducted for each farm, four tests is not a sufficient number to begin worrying about error rates. In addition, the intention is to estimate yield loss relative to undis-

eased plots. There really is no decision about how to "behave" involved. This is another case for which the hypothesis will most likely be of the form $H : \mu_1 = \mu_2$.

Example 17.11: A technician at a county water treatment facility in the Central Valley of California is responsible for testing *incoming water* for fecal bacteria (coliform and fecal sreptococci) each day. If levels are low enough, then the water may be used for irrigation of vegetable crops without treatment. The county uses a criteria that no more than 4% of samples taken should exceed a level of 2.2 coliforms per 100 ml. The technician selects 20 samples per day.

Although this is not a manufacturing example, it fits the scenario of tests for acceptance sampling quite well. The test is repeated each day, and a decision is made to allow the water to be used for irrigation of vegetables without further treatment, or to prohibit such use. Controlling the rate at which errors of Type I and Type II will be committed is a natural concern in this problem. The hypotheses of interest are most likely $H_0 : \theta \leq 0.04$ versus $H_1 : \theta > 0.04$ where θ is the mean of a set of independent and identically distributed binary random variables (that is, a binomial proportion).

17.5 Likelihood Ratios as Relative Support

Recall the intuition offered for determining a point estimate by maximizing the likelihood function. In the case of discrete random variables, the likelihood function is numerically equal to the (joint) probability of a set of observed data for whatever value of the parameter it is evaluated at (i.e., the argument of the likelihood function). It is intuitive, then, to give as

an estimate that value of the parameter that makes the probability of the observed data as great as possible. In the case of continuous random variables (for which the integral mean value theorem applies to a continuous density), the likelihood function is proportional to the probability the random variables corresponding to each datum are in some small interval, again for the parameter value at which the likelihood is calculated. As long as the proportionality factor remains constant, which it will for a fixed sample size and with all data measured or observed with the same precision, the same intuition applies. The statistician A.W.F. Edwards was perhaps the foremost advocate of considering the likelihood as what he called a measure of "support" of the observed data for various parameter values. It is clear, then, that the maximum likelihood estimate for a given set of data is that value of the parameter that the data support to a greater degree than any other possible value of the parameter.

Now consider any two particular parameter values, say θ_1 and θ_2 for the joint pmf or pdf of a random sample $f(\boldsymbol{y}|\theta)$. The *likelihood ratio*

$$R(\theta_1, \theta_2) = \frac{L(\theta_1|\boldsymbol{y})}{L(\theta_2|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\theta_1)}{f(\boldsymbol{y}|\theta_2)},$$

can then be interpreted as a measure of support in the data for θ_1 *relative* to the support for θ_2. Now, the intuitive argument given for likelihoods as proportional to probabilities of the data makes it easy to be misled into thinking that $R(\theta_1, \theta_2)$ is a ratio of the probability that $\theta = \theta_1$ to the probability that $\theta = \theta_2$. This is to be avoided, as it is not true. The intuition based on probabilities that is attached to likelihoods is that of probabilities of the *data*, for particular values of θ. But $R(\theta_1, \theta_2)$ is a valid quantification of the relative degrees to which the data $\boldsymbol{y}$ are in concert with the two parameter values.

It is typical to work with log likelihoods, rather than likelihoods directly,

and the measure of relative support given above then becomes

$$r(\theta_1, \theta_2) = \log\{R(\theta_1, \theta_2)\} = \log\{f(\boldsymbol{y}|\theta_1)\} - \log\{f(\boldsymbol{y}|\theta_2)\}.$$

We will follow Edwards and call $r(\theta_1, \theta_2)$ the relative support for θ_1 over θ_2.

17.5.1 Tests Based on Support

In this subsection we present a basic procedure by which statistical tests may be conducted using a relative support function in the form of a difference in log likelihood functions. These tests will, in fact, be quite distinct from those based on sampling distributions, always in interpretation, and often in mathematical form. We divide this material into parts based on the degree of specificity of the hypotheses being considered. What are called *simple hypotheses* are hypotheses that give a point value for a parameter, $H : \mu = \mu_0$ for example. In contrast, what are called *composite hypothesis* are hypotheses that specify only that a parameter is in some portion of the parameter space, $H : \theta \in \Theta_0 \subseteq \Theta$.

Simple Null Hypothesis

Consider testing a hypothesis of $H_0 : \theta = \theta_0$ in some problem, which is a simple hypothesis. In Tests of Significance, we eschewed stating an alternative hypothesis, although one viewpoint is that the negation of a hypothesis serves as a de facto alternative if one is to reject the hypothesis. In tests for acceptance sampling we formed an alternative hypothesis as a logical disjunction $H_1 : \theta \neq \theta_0$. For the approach we will call Tests Based on Support, an explicit alternative hypothesis is always required. If both null and alternative hypotheses are simple, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, then the relative support

function $r(\theta_0, \theta_1) = [r(\theta_1, \theta_0)]^{-1}$ can be used directly to reach a conclusion about which value of θ has greater support in the data. We simply choose whichever hypothesis has greater support, with the understanding that this choice is relative to the set of only two possibilities, θ_0 or θ_1. This follows the advice of Jeffreys (1931, p. 396) that "if we must choose between two definitely stated alternatives we should naturally take the one that gives the larger likelihood." Notice, however, that this prescription is based on an entirely different objective, judging support based on the observed data, than is a Neyman-Pearson test where control over error rates is the goal.

If the null hypothesis is simple and the alternative is composite, such as typically occurs in the Neyman-Pearson procedure, a test based on support changes the problem so that the question is whether there exists another simple alternative in H_1 with greater support than H_0, and how much greater that support can possibly be. In the case of a simple null hypotheses about the parameter θ we know the value of θ that the data support more than any other is if we took θ equal to the maximum likelihood estimate $\hat{\theta}_n$. Suppose, for example, that in a given problem our interest is in testing $H_0 : \theta = 5.0$ versus $H_1 : \theta \neq 5$, and in the set of data collected we estimate $\hat{\theta}_n = 8.0$ using maximum likelihood. We might then consider that the value of θ in H_1 that has the greatest support is $\theta = 8.0$ and the maximum increase in support for any θ given by H_1 over $H_0 : \theta = \theta_0$ is $r(8, \theta_0)$. All that is needed, then, is some principled manner to determine how to judge the relative support for θ_0 against the maximum increase possible, which is $r(\hat{\theta}_n, \theta_0)$. Edwards (1972, p. 182) suggests a benchmark of 2 so that any θ_0 with $r(\hat{\theta}_n, \theta_0) \geq 2$ is rejected, while any θ_0 with $r(\hat{\theta}_n, \theta_0) < 2$ is not rejected. Edwards' preference for a value of 2 in judging relative support seems to have been fairly arbitrary, just like the significance level or error rate of 0.05, but he seemed quite pleased that in

a normal one sample problem it matches that other scale reasonably closely. Suppose that a value of $\mu = \mu_0$ differs from $\bar{y}$ by two standard deviations, $\mu_0 = \bar{y} + 2(\sigma^2/n)^{1/2}$. Then the relative support for $\bar{y}$ over μ_0 is $r(\bar{y}, \mu_0) = 2$, and 2 is nearly the critical value of 1.96 for a test of $H_0 : \mu = \mu_0$ using the statistic $(\bar{y} - \mu_0)/(\sigma^2/n)^{1/2}$ and with $\alpha = 0.05$. Edwards, of course, claimed that this "vindicated" a value of 2 in tests based on sampling distributions rather than justifying a value of 2 in a Test Based on Support (Edwards, 1972, p.182)

Example 17.12

Consider a one sample normal problem with known variance, $Y_1, \ldots, Y_n \sim$ iid $\mathrm{N}(\mu, n)$. Suppose our hypotheses are $H_0 : \mu = 5$ and $H_1 : \mu \neq 5$, and that the maximum likelihood estimate of μ based on the observed data is $\bar{y} = 8.0$.

Figure 17.1 shows log densities of $\bar{Y}$ for $\mu = 5$ and $\mu = 8$ in our one sample normal problem with known $\mathrm{var}(\bar{Y}) = \sigma^2/n = 1$. The support of the data, in which we observed $\bar{y} = 8$, for $\mu = 5$ is depicted as the circle on the solid curve and has the value $\log\{f_{\bar{Y}}(8|\mu = 5)\} = -5.4189$. Similarly, the support of the data for $\mu = 8$ is depicted as the triangle on the dashed curve and has the value $\log(f_{\bar{Y}}(8|\mu = 8)\} = -0.9189$. The relative support for $\mu = 8$ over $\mu = 5$ is then,

$$r(8, 5) = (-0.9189) - (-5.4189) = 4.50.$$

If we follow Edwards prescription, we have $r(\mu_0, \hat{\mu}_n) > 2$ and we would reject $H_0 : \mu = 5.0$.

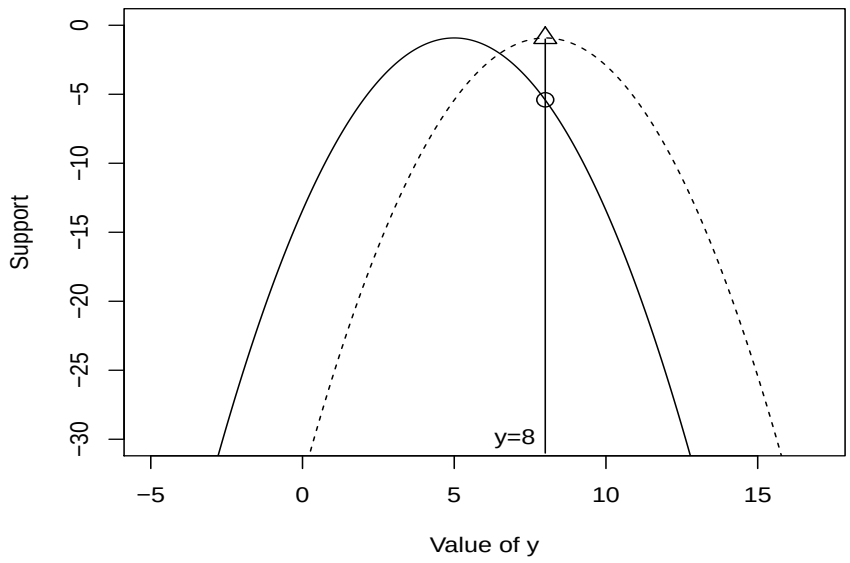

Figure 17.1: Support curves given $\mu = 5$ (solid curve) and $\mu = 8$ (dashed curve) in a one sample normal problem with $\sigma^2/n = 1$. The points depicted give the support for $\mu = 5$ and $\mu = 8$ given by an observed value of $\bar{y} = 8.0$.

Two Composite Hypotheses: A Transition to Model Selection

In the Neyman-Pearson context, simple versus simple hypotheses are rather odd ducks, despite the fact that the classic version of the Neyman-Pearson lemma is formulated for that situation. We say that setting is odd because the underlying concept in a Neyman-Pearson framework is to arrive at a behavioral decision rule; behave as if H_0 is true or behave as if H_1 is true. It does not take much imagination to wonder how we might have gotten to the point where we know there are only two very specific realities or to imagine that there may be other possibilities that are superior to either of

the stated hypotheses. But the conclusion of a Neyman-Pearson test with two simple hypotheses must be either to accept H_0 or accept H_1; there is no other way to control error rates, the foundation of Neyman-Pearson theory. In contrast, testing two simple hypotheses using a support test results in a conclusion to accept (or prefer) H_0 over H_1 or to accept (or prefer) H_1 over H_0; I prefer the word prefer here. It is entirely possible that there may exist another hypothesis, H_2 say, that we would prefer to either H_0 or H_1. A distinction between the procedures then is that for simple versus simple hypotheses, Neyman-Pearson declares the hypotheses to constitute a logical disjunction, while support tests make no such assertion. In tests of a simple null hypothesis versus a disjunctive composite alternative, support tests recast the problem in terms of two simple hypotheses, one of which can only be determined after an analysis of the observed data, making the conclusion a relative comparison of a typical hypothesized value for a parameter with a fitted model that has an estimated parameter. In Example 17.12 the null hypothesis has no estimated parameters, the alternative has one. And the scale for judging relative support changes from one in which the hypotheses are compared head-to-head to one that recognizes the alternative hypothesis now leads to the largest likelihood possible. It is not much of a stretch to extend this situation to having two composite hypotheses which then amounts to a relative comparison of two fitted models, and in which those models do not exhaust the possibilities.

Testing composite versus composite hypotheses within the framework of model comparison does involve some restrictions on the models that can be included. We will assume that one model contains more parameters than the other. The model with a greater number of parameters will be called the *Full* model, and that with a lesser number of parameters will be called the

Reduced model. Full and reduced models are required to be *nested* in that the reduced model can be obtained from the full model by placing restrictions on the full model parameters. Relative to hypotheses, the reduced model corresponds to a null hypothesis while the full model corresponds to the alternative hypothesis.

Example 17.13

Consider a multiple linear regression model with normally distributed error terms, $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \sigma\epsilon_i$, where $\epsilon_i \sim$ iid $\mathrm{N}(0,1)$ for $i = 1, \ldots, n$. We will consider this to be the full model, and might like to test whether the covariate $x_{i,2}$ makes a meaningful contribution to this model. We can then formulate a reduced model through the restriction $\beta_2 = 0$ to arrive at $Y_i = \beta_0 + \beta_1 x_{i,1} + \sigma\epsilon_i$. A statistic representing the support of the full model over the reduced model is then,

$$\begin{aligned} r(F,R) = &-\frac{n}{2}\log(\hat{\sigma}_F^2) + \frac{n}{2}\log(\hat{\sigma}_R^2) \\ &- \frac{1}{2\hat{\sigma}_F^2}\sum_{i=1}^{n}\{(y_i - \hat{\beta}_{0,F} - \hat{\beta}_{1,F}x_{i,1} - \hat{\beta}_{2,F}x_{i,2})^2\} \\ &+ \frac{1}{2\hat{\sigma}_R^2}\sum_{i=1}^{n}\{(y_i - \hat{\beta}_{0,R} - \hat{\beta}_{1,R}x_{i,1})^2\}. \end{aligned}$$

Example 17.14

Two distributions that are often used to model time-to-event or lifetime data are the exponential and gamma distributions. Suppose that $Y_1, \ldots, Y_n$ are are random variables connected with the lifetimes of water pumps when run continuously under controlled conditions. We will assume the pumps are operated in such as way that it would be reasonable to believe that these random variables constitute a random sample from a common lifetime distribution. A model that takes the distribution to be gamma with parameters

$\alpha > 0$ and $\beta > 0$ has probability density function,

$$f_Y(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y);\ y > 0.$$

The joint density is then $f_Y(\boldsymbol{y}|\alpha,\beta) = \prod_{i=1}^n f_Y(y_i|\alpha,\beta)$. We will consider the model with gamma distributions to be the full model. We might like to test whether the simpler exponential distribution (with constant hazard) provides a representation of the data that is not dramatically inferior the gamma model. We obtain an exponential model by placing the restriction $\alpha = 1$ on the full model to obtain, for $\beta > 0$,

$$g_Y(y|\beta) = \beta \exp(-\beta y); \quad y > 0,$$

and the joint density is again $g_y(\boldsymbol{y}|\beta) = \prod_{i=1}^n g_Y(y_i|\beta)$. The relative support function in favor of the full model over the reduced model is,

$$r(F,R) = \log\{f_Y(\boldsymbol{y}|\alpha,\beta)\} - \log\{g_Y(\boldsymbol{y}|\beta)\}.$$

In general, we suppose that a full and a reduced model take distributions of response random variables to have the same form but with parameter spaces such that Θ_R is nested within Θ_F, or Θ_R is a manifold in Θ_F, two ways of saying that we can obtain Θ_R by placing one or more restrictions on the parameters contained in Θ_F. The relative support in the data for the full model over the reduced model is given as,

$$r(F,R) = \max_{\boldsymbol{\theta}\in\Theta_F} \log\{f_Y(\boldsymbol{y}|\boldsymbol{\theta})\} - \max_{\boldsymbol{\theta}\in\Theta_R} \log\{f_Y(\boldsymbol{y}|\boldsymbol{\theta})\}. \tag{17.1}$$

Note that the situation of a simple hypothesis tested against the value of a maximum likelihood estimate such as in Section 17.5.1 is a special case of (17.1) in which the reduced model has parameter given by the hypothesis, and the log likelihood of the full model is maximized over the entire parameter space.

In both Example 17.13 and Example 17.14 the full model contains one more parameter than does the reduced model so $\dim(\Theta_F) - \dim(\Theta_R) = 1$. This makes Williams criterion of $r(F, R) \geq 2$ a reasonable benchmark against which to judge the test statistics. But for many pairs of nested models that we might be interesting in comparing, the full model contains more than one additional parameter than the reduced model. A difficulty that arises is due to the fact that placing restrictions on the values of elements of a parameter vector is guaranteed to decrease the maximized log likelihood value. That is, a reduced model nested within a full model will always have a smaller maximized log likelihood than the full model, meaning that (17.1) will always be positive. This then implies that a reduced model obtained from two restrictions on the parameters of a full model will have a smaller maximized log likelihood than a reduced model obtained from only one such restriction. Thus, the more parameters a full model contains relative to a reduced model, the greater the value of (17.1). This will be true regardless of whether the additional parameters in the full model contribute anything of actual value to how much better the full model represents the data than the reduced model. For this reason, before we conclude that we prefer a full model to a reduced model, we want the full model to have *sufficiently greater* support from the data than the reduced model, where the term sufficiently must be assessed relative to the number of parameters in the full and reduced models. Another way to understand this is to invoke a principle of *simplicity*, sometimes referred to as *Occam's Razor*. The idea is that models with more parameters are more complex than models with fewer parameters. We should prefer a simpler explanation or model to a more complex explanation or model unless the more complex explanation is sufficiently better than a simpler explanation. A quote often attributed to Albert Einstein is, "Everything should

be made as simple as possible, but not simpler." The key to determining whether a full model represents data to a degree sufficiently greater than a reduced model, then, is having a scale against which to judge (17.1), and this must take into account the difference in the number of parameters contained in the full and reduced models. If a benchmark of 2 is a reasonable value for a test of a reduced model corresponding to a simple hypothesis against an full model with unconstrained parameter space, then a reasonable benchmark for the more general setting is to favor the full model over the reduced model if $r(F, R) \geq 2(p_F - p_R)$, where p_F is the number of parameters in the full model and p_R is the number of parameters in the reduced model.

Notice that in our discussion of likelihoods and support functions we have avoided using the random versions of any quantities. All values $\hat{\theta}$ have been maximum likelihood estim**ates**, not maximum likelihood estim**ators**. Support, relative support, and all likelihood or log likelihood functions have been observed versions, written in terms of $\boldsymbol{y}$, not random versions using $\boldsymbol{Y}$. This is because we are not using *sampling distributions* to provide meaning for tests based on support. In fact, as we will see, making use of sampling distributions in the assessment of test statistics renders a procedure one of pre-data precision rather than post-data precision.

An additional point that is relevant here is the relation between Neyman-Pearson tests of hypotheses (and Neyman's confidence intervals) and the use of likelihood ratios as support functions. Recall that the Neyman-Pearson lemma gives a most powerful test of one simple hypothesis versus a second simple hypothesis as a likelihood ratio. But that likelihood ratio is not interpreted in terms of support in the data for the two hypotheses. In fact, Neyman made this clear, saying "I do deal with likelihood function . . . However, I do so not as a matter of principle, but only in cases when the fre-

quency properties of the estimators fit my purposes" (Neyman, 1977, p.100). Thus, Neyman used likelihood ratios as a way to construct procedures to control error rates, not in terms of their interpretation as measures of support for hypotheses.

17.6 Likelihood Ratio Tests

We now turn to the use of the random version of likelihood ratios to provide a reference scale against which to judge the value of a test statistic. In 1938, Sammuel Wilks provided a result that gives the asymptotic distribution of a log likelihood ratio. We state that result here for completeness.

Likelihood Ratio Asymptotic Distribution

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ be a random sample from a distribution with probability density or mass function $f_Y(y|\boldsymbol{\theta})$ for $y \in \Omega$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_S)^T \in \Theta$. Let a full model have parameter $\boldsymbol{\theta}_F \in \Theta_F \subseteq \Theta$ and a reduced model have parameter $\boldsymbol{\theta}_R \in \Theta_R \subset \Theta_F$. The models must be nested in the sense that the reduced model can be obtained from the full model by placing restrictions on one or more elements of $\boldsymbol{\theta}_F$. Let $\hat{\boldsymbol{\theta}}_{n,F}$ and $\hat{\boldsymbol{\theta}}_{n,R}$ be sequences of maximum likelihood estimators of $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_R$, respectively. Then, under the reduced model,

$$T_n(\boldsymbol{Y}) = -2[(\ell_n(\hat{\boldsymbol{\theta}}_{n,R}, \boldsymbol{Y}) - \ell_n(\hat{\boldsymbol{\theta}}_{n,F}, \boldsymbol{Y})] \; right\overset{d}{arrow} \; \chi^2(p-q), \qquad (17.2)$$

where $\chi^2(p-q)$ denotes a Chi-squared distribution with $p-q$ degrees of freedom, $p = \dim(\Theta_F)$ and $q = \dim(\Theta_R)$. Note that for an observed sample $\boldsymbol{y}$, $T_n(\boldsymbol{y}) = 2r(F, R)$ where $r(F, R)$ is given in (17.1). The result (17.2) provides an approximate sampling distribution for a likelihood ratio test statistic.

The χ^2 distribution in (17.2) could then be used to provide critical values in the context of a Neyman-Pearson test of hypotheses or to compute a $p-$value in the context of a Fisherian test of significance. We will argue, however, that likelihood ratio tests should be considered a distinct approach to testing in their own right. If one accepts that a $p-$value in Fisher's tests of significance can be interpreted as a measure of evidence against a hypothesis (more on this later in the chapter) then there should be an analogous interpretation for a likelihood ratio test. Although a $p-$value can certainly be computed on the basis of (17.2), as $Pr[\chi^2(p-q) \geq T_n(\boldsymbol{y})]$, what is the interpretation of that value? The test statistic T_n cannot be computed without knowledge of both the full and reduced models. The Chi-squared limit distribution for T_n is under the reduced model, so the reduced model plays the role of the hypothesis or null hypothesis in the test procedure. But the $p-$value now can be interpreted only as a measure of evidence against the reduced model relative to the full model. If the reduced model is considered a hypothesis, the full model must be considered an alternative hypothesis. But full and reduced models are not necessarily disjunctive in the sense that they do not exhaust the possibilities. This fits neither the framework of Neyman-Pearson nor that of Fisher, even if one insists that rejection of a hypothesis in tests of significance implies the negation of that hypothesis as an alternative. Likelihood ratio tests, then are perhaps reminiscent of both tests of significance and tests for acceptance sampling, but are analogous to neither. They are best considered within the context of model selection.

Example 17.15

To reinforce the notion that likelihood ratio tests are best thought of as a model selection procedure rather than hypothesis tests, consider a situa-

tion in which we have a two sample problem with beta distributions. That is, we have random varibles $\{Y_{i,j};\, i = 1, 2;\, j = 1, \ldots, m_i\}$ such that, for $j = 1, \ldots, m_i$, $Y_{i,j} \sim$ iid Beta(α_i, β_i). We would like to conduct a group comparison by testing a reduced model in which $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ against a full model with four distinct parameters. We can achieve the computations for this test by estimating the parameters of the full model already given and estimating α and β from the reduced model which is $Y_{i,j} \sim$ iid Beta(α, β). Suppose the maximized log likelihood for the full model is $\ell_F = \ell_1(\hat{\alpha}_1, \hat{\beta}_1) + \ell_2(\hat{\alpha}_2, \hat{\beta}_2) = -118.38$ while the maximized log likelihood for the reduced model is $\ell_R = \ell_1(\hat{\alpha}, \hat{\beta}) + \ell_2(\hat{\alpha}, \hat{\beta}) = -125.64$. The likelihood ratio test statistic (17.2) is then $T_n = -2(\ell_R - \ell_F) = 14.52$ which has the associated $p-$value $p = 1 - Pr(\chi^2(2) \leq T_n) = 0.0007$. Because the $p-$value is small, we would prefer the full model over the reduced model. Now consider a third model in which we reparameterize the beta distributions using $\mu = \alpha/(\alpha + \beta)$ and $\phi = 1/(\alpha + \beta + 1)$ and in which we take $Y_{i,j} \sim indepBeta(\mu, \phi_i)$, that is, each group has its own expected value but the beta distributions will differ in shape. Our previous full model still corresponds to having four distinct parameters and, due to invariance of likelihoods, will still have the same maximized log likelihood. Suppose our newly introduced model is now considered the reduced model and analysis results in a maximized log likelihood $\ell_R = 124.11$. A likelihood ratio test of this model against the previous full model then has $T_n = 3.06$ with associated $p = 0.2165$. Our likely conclusion now would be to prefer the new reduced model over the full model.

17.7 Assumption or Hypothesis

An issue that is perhaps not always given its due in discussion of tests concerns which parts of the structure of the test procedure are considered assumptions of the procedure, and which parts are considered parts of the hypothesis or hypotheses under consideration. This issue impacts interpretation of test results, is involved in further examination of the relation between likelihood ratio tests and other procedures, and may even impact our willingness to entertain Fisher's tests of significance as procedures that involve only one hypothesis.

The setup for a Neyman-Pearson test of hypotheses is typically presented as involving a number of assumptions. For example, a test of $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ may assume that we have random variables that are independent and identically distributed with a common normal distribution having expectation μ. A test of $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ may assume that we have two groups of independent and identically distributed random variables with common normal distributions within each group such that the variances of those distributions are equal. The test only has meaning within the context of its assumptions, and if those assumptions are violated to any meaningful extent we declare the test procedure to be invalid and consider that it has provided no useful information about the problem. Indeed, if one is to calculate power one must have an assumed distributional form, independence and identical distribution, with a change allowed only in the value of a parameter. This framework of assumptions that render a procedure valid has led to extensive investigation of the robustness properties of statistical procedures, such as what is known as the *Princeton robustness study* (Andrews, Bickel, Hampel, Huber, Rogers, and Tukey, 1972).

As already noted, Fisher was criticized for not recognizing an alternative hypothesis (Edwards, 1972; Kennedy-Shaffer, 2019), but it is possible that a portion of this criticism is due to applying the test-assumption framework to tests of significance in an uncritical manner. For example, Edwards, immediately after commenting that "Fisher's tests imply alternative hypotheses of particular types" (Edwards, 1972, p. 180) begins an example of a test for a normal observation with the preface that "We assume, initially, that the Normal distribution and its variance are part of the model, not to be questioned on this occasion, and that the hypothesis concerns μ." If one takes this view of Fisher's framework, then the existence of an alternative hypothesis involving μ does seem unavoidable. If one rejects $H : \mu = \mu_0$ then one must certainly be accepting an unstated alternative that $\mu \neq \mu_0$. And the use of tail areas to compute the $p-$value strongly suggests that it is a shift in location that is responsible for a rejection of H. There is some evidence, however, that at least on some occasions Fisher was including what are rightly considered assumptions of the procedure in tests for acceptance sampling, rather as part of the hypothesis being tested in a test of significance (e.g., Inman, 1994, p. 7). Indeed, Fisher (1955, p. 75) stated that "The hypothesis is sometimes called a model, but I should suggest that the word model should only be used for aspects of the hypothesis between which the data cannot discriminate." Here, it seems clear that Fisher would not consider distributional form or perhaps even equality of variances to be part of what he would call the model. Regardless of whether Fisher was or was not including aspects of the situation we sometimes think of as assumptions of the test as part of the hypothesis, consider what happens if we do so. The hypothesis being tested in a one sample normal problem then becomes $H : \boldsymbol{Y}$ is a random sample from a normal distribution with expected value $\mu = \mu_0$

and variance σ^2. The negation of this hypothesis involves many possibilities other than that $\boldsymbol{Y}$ is a random sample from a normal distribution with expected value $\mu \neq \mu_0$. The hypothesis H and its implied alternative $\mu \neq \mu_0$ no longer form a logical disjunction. Attempting a similar expansion of the null hypothesis in a Neyman-Pearson framework vitiates the entire procedure, as how does one behave under a decision that a null hypothesis is not entirely true, but any one of an array of other possibilities might be? If one allows the possibility that parts of what Edwards calls *the model* might be instead parts of the hypothesis under examination, it is still true that not all types of ways that H might be violated necessarily lead to $p-$values of similar magnitudes. One might surmise that this difficulty was at least a part of what prompted Edwards to claim that the major flaw in tests of significance was the failure to identify an explicit alternative (or, really, model). Finally, we also note that on this issue, as any number of others, Fisher sometimes seemed self-contradictory. At one point, Fisher was rather adamant, saying "In fact, errors of the second kind are committed only by those who misunderstand the nature and application of tests of significance" (Fisher, 1933, p. 474). At other points, however, Fisher did refer to his hypothesis as a null hypothesis, and offered a discussion of the use of alternatives Fisher (1935, chapter 10).

For support tests and likelihood ratio tests, we usually take specified distributions and other 'specifications about the setting to be part of model formulation, but rather than model formulation serving as assumptions for tests about particular parameter values, the test choose between alternative model formulations. In example 17.15, the test between a model with two distinct beta distributions with four parameters and a model with one distinct beta distribution with two parameters could be considered a test of,

two hypotheses

H_0 : Reduced Model: $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ and the response variables are random samples from beta distributions,

versus

H_1 : Full Model $\alpha_1 \neq \alpha_2$ and $\beta_1 \neq \beta_2$ and the response variables are random samples from beta distributions.

In contrast to traditional tests for means, the null and alternative hypotheses immediately above do not exhaust the possibilities because random samples and the distributional forms specified are part of the hypotheses rather than considered assumptions under which the test is valid. It could be, for example, that the response variables do not follow gamma distributions, but rather belong to some other family of distributions. In this case it may be that neither the reduced model nor the full model are able to describe the data well, and yet those are the only two possibilities that the test is comparing. Even if this would be the case, however, the likelihood ratio test is not considered invalid but rather is interpreted as an indication of which of the two models considered we should prefer. It may be that neither is much good, but the test procedure does not address that, it only compares the full and reduced models relative to each other. Similarly, if the observations do not correspond to independent random variables, that also does not invalidate the test, which is only comparing two possible models both of which may be flawed in specifying independence. The situation is not quite as wild and unconstrained as this might imply, however. It is presumed that if one has chosen a model that is truly horrible at describing the data, one will encounter difficulties in locating maximum likelihood estimates, except

perhaps for the simplest models. One additional caveat can also be offered. The use of an asymptotic chi-squared reference distribution in a likelihood ratio test may be compromised if the models are poorly chosen because that reference distribution only holds under the specified reduced model. If the true situation deviates from that model not only in the number and values of parameters, but also in other areas of model formulation, the meaning of $p-$values might be hard to discern.

17.8 Bayesian Tests

We covered Bayes Factors in a separate chapter because they can appear in settings that are not necessarily decision theoretic in nature, and the usual approach to developing a formal theory for testing hypotheses in a Bayesian framework is to cast the problem in terms of decision theory. A typical progression (e.g., Berger, 1985; O'Hagan, 1994; Robert, 2007) is to define the testing problem as consisting of a choice between $H_0 : \theta \in \Theta_0 \subset \Theta$ and $H_1 : \theta \in \Theta_1 \subset \Theta$. Let d_0 denote the decision to accept H_0 and d_1 the decision to accept H_1. Start with $0-1$ loss, $L(d_i, \theta) = 0$ if $\theta \in \Theta_i$ and $L(d_i, \theta) = 1$ if $\theta \in \Theta_j$; $j \neq i$. The associated Bayes expected loss is,

$$E_B(d_i, \theta) = E[L(d_i, \theta)] = \int L(d_i, \theta) p(\theta|\boldsymbol{y})\, d\theta = \int_{\Theta_j; j \neq i} p(\theta|\boldsymbol{y})\, d\theta, \quad (17.3)$$

which, for d_0 and d_1 are just the posterior probabilities $Pr(\Theta_1|\boldsymbol{y})$ and $Pr(\Theta_0|\boldsymbol{y})$, respectively. Expected loss is minimized in this problem by choosing H_0 if $Pr(\Theta_1|\boldsymbol{y}) < Pr(\Theta_0|\boldsymbol{y})$ and choosing H_1 if $Pr(\Theta_1|\boldsymbol{y}) > Pr(\Theta_0|\boldsymbol{y})$. Usually, Θ_0 and Θ_1 partition Θ and the decision is sometimes framed as accept H_0 if $Pr(\Theta_0|\boldsymbol{y}) > 0.50$ and reject H_0 if $Pr(\Theta_0|\boldsymbol{y}) < 0.50$.

The next step in the progression is to replace $0-1$ loss with,

$$L(d_i, \theta) = \begin{cases} 0 & \text{if } \theta \in \Theta_i \\ k_i & \text{if } \theta \in \Theta_j;\ j \neq i \end{cases}, \tag{17.4}$$

The Bayes expected loss for decision d_0 is then $k_0 Pr(\Theta_1|\boldsymbol{y})$ and that for d_1 is $k_1 Pr(\Theta_0|\boldsymbol{y})$ and we would accept H_0 if $k_0/k_1 < Pr(\Theta_0|\boldsymbol{y})/Pr(\Theta_1|\boldsymbol{y})$ and accept H_1 if $k_0/k_1 > Pr(\Theta_0|\boldsymbol{y})/Pr(\Theta_1|\boldsymbol{y})$. If Θ_0 and Θ_1 partition Θ then we would equivalently accept H_0 if $Pr(\Theta_0|\boldsymbol{y}) > k_0/(k_0 + k_1)$ and accept H_1 if $Pr(\Theta_1|\boldsymbol{y}) > k_1/(k_0 + k_1)$. If one wishes to draw an analogy with Neyman-Pearson testing, a rejection region for H_0 can be defined as (Berger, 1985, p. 164),

$$C = \left\{ \boldsymbol{y} : Pr(\Theta_1|\boldsymbol{y}) > \frac{k_1}{k_1 + k_0} \right\} \tag{17.5}$$

which can be difficult to determine in practice unless $n = 1$, that is, $\boldsymbol{y} = y_1$. This analogy with Neyman-Pearson tests for acceptance sampling seems to contradict the spirit of Bayesian analysis, as (17.5) constitutes a decision rule that is more in line with a pre-data precision approach than a post-data precision approach, and it is the position of Bayes procedures as the most clearly post-data procedures in statistics that gives Bayesian inference its force. It seems that, in presenting (17.5), Berger was simply trying to point out that if one insists on a Neyman-Pearson-like rule, minimization of Bayes risk provides a rational way of choosing critical values (Berger, 1985, p. 165) and should be preferred to the arbitrary 0.10, 0.05, 0.01 choices so often employed in frequentist testing. This assumes, of course, that choice of k_0 and k_1 somehow avoid arbitrary selection.

The framework can be broadened further by allowing the decision space to contain more than two elements. Consider, for example three possibilities, $\theta \in \Theta_0$, $\theta \in \Theta_1$and $\theta \in \Theta_2$. Let d_i denote the decision to accept H_i and again

use the loss function (17.4), now with $i, j = 0, 1, 2$. Bayes expected loss is still given by (17.3) and we would reach decision d_i (accept H_i) if, for $i = 1, 2, 3$,

$$Pr(\Theta_i|\boldsymbol{y}) \left(\sum_{j \neq i} k_j \right) = \min_h \left\{ Pr(\Theta_h|\boldsymbol{y}) \left(\sum_{j \neq h} k_j \right) \right\}.$$

The procedure with a decision space having three elements allows tests that include a region of indifference or equivalence. Consider a specified value θ^* and a value $\delta > 0$ such that we consider any θ within δ of θ^* to be equivalent to θ^*. Then we might construct a test of $H_0 : \theta < \theta^* - \delta$, $H_1 : \theta > \theta^* + \delta$ and $H_2 \theta^* - \delta \leq \theta \leq \theta^* + \delta$. Situations in which this type of procedure might be appropriate include θ representing the difference in mean improvements in some medical condition resulting from two drugs, or the difference in mean concentrations of some contaminant in sediment between a stretch of river being dredged and the terrestrial receiving zone for the dredge spoil.

Example 17.16

In our discussion of Bayes Factors in Chapter 14, extensive use was made of the example of sex ratio at birth in guanacos. Recall that, based on a claim that mortality in the first year of life was 10% greater for males than females, the sex ratio at birth should be 0.524 in favor of males. We formulated models to test $\theta > 0.50$ versus $\theta < 0.50$, $\theta > 0.524$ versus $\theta < 0.524$, $\theta = 0.50$ versus $\theta = 0.524$, $\theta = 0.50$ versus $\theta \neq 0.50$, and $0.523 < \theta < 0.525$ versus $0.499 < \theta < 0.501$. Some of these Bayes factors were computed using prior odds under the beta$(35, 35)$ prior used in the overall model, and some under a modified or contrived prior odds of 1.0. In each case, we were unable to conclude there was more than only a hint of evidence in favor of the model (hypothesis) corresponding to a higher probability of male than female birth, but neither was there much if any evidence for the competing

model. Here, we can attempt to formalize the problem using a Bayesian test with three hypotheses. If the proportion of male births is 5% greater than that of female births, then the probability of a male birth should be 0.515. If we are willing to consider any increase in male birth probability of less than this amount biologically unmeaningful, then our hypotheses become $H_0 : 0.515 < \theta$, $H_1 : \theta < 0.485$, and $H_2 : 0.485 < \theta < 0.515$. The posterior for this problem is a beta distribution with parameters 184 and 172 (see Chapter 14.1). If we are willing to assign equal weight to different losses we would take $k_0 = k_1 = k_2 = 1$ and the estimated Bayes risks for accepting the hypotheses are $H_0 : 0.472$, $H_1 : 0.885$ and $H_2 : 0.643$ which would lead to the decision d_0 to accept H_0 that the probability of a male birth is greater than 0.515. If we felt that reaching conclusions that the sex ratio at birth differs from 1.0 when those conclusions are incorrect is more seriously wrong than concluding it does not differ in a meaningful way from 1.0 when that is incorrect, we might define the loss function using $k_0 = k_1 = 1.25$ and $k_2 = 1.0$, which would lead to risks of $H_0 : 0.590$, $H_1 : 1.107$, and $H_2 : 0.643$, again leading to a decision to accept H_0. Increasing the loss of incorrect decisions that differ from H_2 to $k_0 = k_1 = 1.50$ and leaving $k_2 = 1.0$ produces estimated risks of $H_0 : 0.707$, $H_1 : 1.328$ and $H_2 : 0.643$ and a decision to accept H_2, the hypothesis of indifference. The use of Bayesian test procedures derived directly from a formal expected loss criterion has greatly clarified what we are able to conclude about our guanaco friends of Torres del Paine National Park in Chili.

An interesting aspect of the three hypothesis procedure just illustrated is that it allows not only a decision rule for fixed choice of k_0, k_1 and k_2, but also an assessment of how much more serious an error in certain directions must be relative to a error in determining indifference before the hypothesis

of indifference is the selected decision. In the guanaco example, an error in either direction from the window of indifference would need to be judged as 1.363 times more serious than an error in accepting indifference before the Bayes expected risks are equal for the three possible decisions. This can assist in reaching a conclusion that can then no longer be written as a fixed rule before the observation of data.

There have been attempts to generalize the decision space for a Bayesian test from a discrete set of possible actions $d = \{0, 1\}$ to the interval $d \in [0, 1]$ in which case the decision is a probability about a given hypothesis $H : \theta \in \Theta_0$. A further generalization is then to expand the hypothesis to be $H : \theta \sim f(\theta);\ \theta \in \Theta$ and the decision to also be a density $d(\theta)$. A class of loss functions appropriate for such settings are called scoring rules $S(x, Q)$ which quantify the difference between a distribution Q and an observed event $X = x$. If X has a true distribution P, then the expected value of $S(X, Q)$ is denoted as $S(P, Q)$ (e.g., Gneiting and Raftery, 2007; Dawid, 2014). A scoring rule is called *proper* if $S(P, P) \geq S(P, Q)$ for all Q and P. In the context of forecasting, Q is the forecast distribution for P. In the context of prior elicitation from experts, Q is the stated distribution by the expert and P is the true distribution believed by the expert. In our context, an expected scoring rule would be used to compare $d(\theta)$ to $f(\theta)$, as $E[S(\theta, d(\theta))] = S(f(\theta), d(\theta))$. Among other uses, scoring rules can be used to motivate Kullback-Liebler divergence as the difference between expected loss for a scoring function $L(d, \theta) = -\log\{d(\theta)\}$ (which is a log likelihood) and the minimum expected score, which is the entropy of the true density $f(\theta)$ (O'Hagan, 1994, Chapter 2.54). Thus, entropy and Kullback-Libler divergence are given meaning within a decision theoretic framework. Robert (2007, Chapter 5.4) uses a related idea of proper loss functions, but with

an objective of developing a context within which to assess posterior probabilities and frequentist $p-$values relative to their justification as inferential procedures within a formal decision theoretic framework. Using quadratic loss, Robert shows that posterior probabilities are optimal Bayes estimators of $I(\theta \in \Theta_0)$, where $I(\cdot)$ is the usual identity function. With an additional restriction to problems involving natural exponential families, he also shows that $p-$values are admissible in one-sided but not two-sided tests.

We close this section by indicating that not all statisticians willing to use a Bayesian approach adhere to the need for procedures to be justified within a decision theoretic framework. Indeed, if I interpret a posterior distribution as a quantification of my beliefs regarding the possible values of a parameter or other quantity, why do probability statements made on the basis of that distribution need justification in terms of a framework more formal than that of probability itself? This type of viewpoint is compatible with one in which the very notion of testing has little use. To make inference about any particular event E, I compute its posterior probability $Pr(E|\boldsymbol{y})$. If I want to call E a hypothesis, that is merely semantics, and relative measures for multiple events such as odds are simply a less informative summary than reporting posterior probabilities for the events; that is $Pr(E_1|\boldsymbol{y})/Pr(E_2|\boldsymbol{y})$ contains less information than $Pr(E_1|\boldsymbol{y})$ and $Pr(E_2|\boldsymbol{y})$. For quantities that are continuous, events consisting of a single point or finite set of discrete points have probability 0, so there is no sense considering events such as $\theta = 0$.

17.9 Goodness of Fit Tests

A statistical model consists of a number of parts and we have sometimes referred to systematic and random model components or large scale and small scale model structures when discussing model construction and model assessment. There are any number of ways in which a model can fail to provide what we judge to be an adequate description of the data, including these various model components in total, or even more specific specifications such as the relation between expected values and variances dictated by an assigned random component. Our intent in this section is not to provide an overview of model assessment. See Chapter 11 for a somewhat general discussion. Rather, we want here to describe one portion of the overall process of model assessment consisting of tests of goodness of fit. We left discussion of these tests until the last section of this chapter because they represent a unique use of statistical hypotheses.

The most famous goodness of fit test (GOF) in statistics is the Chi-squared test developed by Karl Pearson (Pearson, 1916, 1922), and we will use this as vehicle by which to discuss issues that apply more generally to other GOF tests as well. Pearson developed his test to "enable a scientific worker to ascertain whether a curve by which he was graduating observations was a reasonable fit." (cited in Inman, 1994). In our modern versions of goodness of fit tests, we frame the procedure in terms of a posited model forming a hypothesis to be tested, and this does seem to be a fair way to think of Pearson's tests. At the point in time that Pearson, Fisher, and others were arguing about the use of GOF tests, however, there was a closer association drawn between scientific hypotheses, such as the Medelian theory of heredity, and statistical hypotheses about the observed frequencies of

phenotypic traits. Thus, when Pearson asserts he is not involved in assessing the truth or falsity of hypotheses we need to understand he is talking at least as much about scientific theories as statistical hypotheses as we understand them. Still, when interpreted relative to statistical hypotheses, many of his thoughts align well with our modern understanding of what hypotheses represent.

> If we narrow ourselves down to to asking whether a normal curve will reasonably *graduate* the material and find that it does, are we to follow it up by asserting that either the sample or the parent population follows a normal distribution? I should say: Certainly *not*. I have never found a normal curve fit anything if there are enough observations! . . . The fact is that all these descriptions by mathematical curves in no sense represent 'natural laws'. They have nothing to do in this sense with 'hypothesis' or 'reverse hypothesis'. They are merely *graduation curves*, mathematical constructs to describe more or less accurately what we have observed.
>
> (K. Pearson, 1935, cited in Inman, 1994, p. 4)

If we cast a GOF test in terms of our modern use of the word hypothesis, a theoretical frequency distribution (i.e., a graduation curve) becomes the hypothesis to be tested; like Fisher, K. Pearson used only one hypothesis. However, the scientific goal was not to reject the hypothesis, as in Fisher's tests of significance. The goal was to be unable to reject the hypothesis, although it was never really accepted in the behavioral sense of Neyman, either, in part because there was not an alternative such that the hypothesis and its alternative defined a logical disjunction. Rather, a theoretical

distribution that could not be rejected in a GOF test was deemed as an acceptable representation of the distribution of the sample and other similar, but unobserved, samples. This was not to imply that there could be only one acceptable theoretical model, and Pearson used the GOF criterion to assess the relative adequacy of competing theoretical distributions to describe the observed data. In this type of comparison, the larger the $p-$value the better. This attitude seems to have arisen from consideration of multinomial likelihoods for grouped data. In that "model", if the expected cell frequencies equaled the observed frequencies exactly, one would obtain $\chi^2 = 0$ and $p = 1.0$ and this would be the model with maximum support from the data; compare this idea with the development of deviance for exponential dispersion family distributions. Thus, the closer to 1.0 the $p-$value, the better the representation of the data by the theoretical distribution. Fisher countered by pointing out that, if the hypothesized model were true, then a $p-$value of 0.999 was just as unlikely as a $p-$value of 0.001 (Inman, 1994). An applied statistician named Buchanan-Wollaston, who prompted a good deal of the exchange between K. Pearson and Fisher by publishing a brief note in *Nature* (Buchanan-Wollaston, 1935) questioning the usefulness of GOF tests, argued for a criterion based on how close the test statistic is to the mode of its chi-squared reference distribution. Buchanan-Wollaston also questioned whether it was justified to concluded that a model fits the data well just because it cannot be rejected in a GOF test.

How are GOF tests interpreted in today's world? First, the debate among Pearson, Fisher, Buchanan-Wollaston and perhaps others about how to use the $p-$value in interpretation of a GOF test has largely been cast aside (but see Seidenfeld, 1979). Those who are uncritical in their thinking, or perhaps just unaware of the intent of a GOF test, interpret critical values (often in the

form of $p-$values) just as in a Neyman-Pearson test for acceptance sampling. If $p > 0.05$ then the (null) hypothesis is accepted, meaning the posited model is declared to "fit" the data. Those willing to accept $p-$values as measures of evidence against a hypothesis interpret the $p-$value in accordance with that concept, but with considerable unease in terms of standard guidelines for significance; one breathes a sigh of relief if the $p-$value turns out to be somewhere between 0.30 and 0.70 and a difficult assessment can be avoided. This seems to be a bit similar to the idea of Buchanan-Wollaston that acceptable $p-$values be in the central portion of the reference distribution. If the $p-$value turns out to be between 0.05 and maybe 0.15, then a conclusion is reached accompanied by at least some degree of unease.

There are not a wide variety of Bayesian approaches to GOF tests. We know of only two general structures that have been proposed. One is to embed a hypothesized parametric model in a large family of nonparametric alternatives and then compute the Bayes factor in favor of the hypothesized model (Berger and Guglielmi, 2001). The question in this approach is how to specify the class of nonparametric alternatives or, more specifically, the prior for a generic class of nonparametric distributions. Suggestions have included use of a Dirichlet process prior (Florens, Richard, and Rolin, 1996), mixtures of Gaussian priors (Verdinelli and Wasserman, 1998), mixtures of triangular distributions (McVinish, Rousseau, and Mengersen, 2008), and Polya tree processes (Berger and Guglielmi, 2001). The other approach, which seems to have been widely adopted, is the use of posterior predictive distributions to compute tests based on traditional measures of discrepancy such as χ^2 or deviance (Gelman, Meng, and Stern, 1996). This same basic strategy can be used to assess a model for representation of particular characteristics of the data (Gelman et al., 1995). The use of posterior predictive distributions for

model assessment is discussed in *Intermediate Statistical Methods, Chapter 7.7*, and illustrated for a hierarchical model in Chapter 13.3.5 of this book.

17.10 Statistical Inference Revisited

We now review each of the approaches to testing considered in this chapter in terms of how they fit into inferential frameworks consisting of a combination of concepts of probability and concerns of pre-data versus post-data precision. In so doing, we have the intention of identifying the fundamental premises on which they depend, as well as the principle issues that may limit their acceptability as a general solution to the problem of how we best make and defend statistical inferences from testing procedures.

17.10.1 Tests of Significance

Tests of Significance present $p-$values as an evidential measure against a hypothesis (or null hypothesis). The interpretation of this measure may be informed by the use of benchmark values (e.g., 0.10, 0.05, 0.01) but is neither bound nor exhausted by those values. That is, two $p-$values of 0.049 and 0.051 present essentially the same amount of evidence against a hypothesis, even if one is thinking of 0.05 as a benchmark. This is the difference between what we are calling a *benchmark* and the *critical values* associated with tests for acceptance sampling. A critical value is an absolute dividing line for drawing conclusions. A benchmark is, well, a benchmark. Values that are very close to a benchmark but on opposite sides still both have essentially the same meaning as the benchmark.

Of great importance for this approach is the connection between $p-$values

and probabilities. The usual interpretation is that a $p-$value represents the probability that the test statistic would assume a value *as extreme or more extreme* than the actual value resulting from the data, if the hypothesis being tested is true. This is the primary way that $p-$values can be claimed to be a measure of evidence based on relative frequency probability (but see the connection between $p-$values and Bayes factors discussed later as well). There are several issues with this interpretation.

1. The concept of probability under which a $p-$value is to be given meaning is hypothetical limiting relative frequency. So the interpretation of a $p-$value as a probability depends on repeated sampling, at least hypothetical sampling from the theoretical model under which a problem is being considered. A claim that this is fundamentally different from the repeated testing needed under the approach of Tests for Acceptance Sampling is not completely uncontroversial, at least in a historical context. It seems that the advent of modern computational ability has made "hypothetical sampling from a theoretical population" reasonably easy to understand, since we can actually do it with simulation. But in the early-to-mid 1900s, what might be meant by hypothetical repeated sampling was less clear. Hacking (1965, p. 25-26) was particularly harsh,

   > Only excessive metaphor makes outcomes of every chance set-up into samples from an hypothetical population . . . to describe a chance process in terms of samplings from populations, you probably need a hypothetical infinite array of hypothetical infinite populations. Chimaeras are bad enough, but a chimaera riding on the back of a unicorn cannot be

tolerated.

There are still statisticians who would be hesitant to abandon the idea that we sample populations and that theoretical probability distributions arise when those populations grow large. And that is still the explanation offered in any number of elementary statistics texts.

2. A related issue is the reliance of $p-$values on what are called tail probabilities. Fisher promoted the interpretation that a small $p-$value could be taken as meaning that either the hypothesis is true and an exceptionally rare event has occurred, or the hypothesis is not true. And the smaller the $p-$value, the more rare the occurrence (of the test statistic value) under the hypothesis. But the tail area involves events that, while possible, have not happened. Especially if one claims there is really no alternative hypothesis involved, why should the tail area greater than a positive observed statistic (or the tail area less than a negative observed statistic) be the region that represents a rare event? Could not any region of small probability represent a rare event under the hypothesis? Note the similarity of this issue with what prompts the selection of confidence intervals endpoints to be determined by tail probabilities and the definition of highest posterior credible intervals in Bayesian analysis. The use of tail areas also can fuel the claim that tests of significance must involve implied alternative hypotheses, at least in the case of means, since it is under location shifts of the reference distribution that the probability in tail regions would increase. A counter-argument to this point is that distributional shape and higher moments also affect tail behavior in distributions so it is only under an assumption that the true distribution of the test statistic is in the same

family as the reference distribution that this criticism has force. Note that there are also connections here with whether portions of model formulation must serve as assumptions that affect the validity of a test (and allow the computation of power) or whether components of model formulation can be included in hypotheses (see Chapter 17.7).

3. Fisher claimed that, although to give a $p-$value meaning in terms of probability required hypothetical repeated sampling, the $p-$value itself was somehow primitive, in the sense that it really did not need to be based on anything else to have meaning that could be "communicated to and understood by rational minds." This has not met with universal acceptance. It should be noted that the two issues listed above seemed to bother Fisher himself who, at the same time he was arguing for the $p-$value as a primitive measure of evidence, was also developing his theory of *fiducial probability* in an attempt to address these issues. That theory has never been widely adopted, nor Fisher's version of it widely understood, by the statistics community. The sampling from a hypothetical population involves another connection that is not always recognized. Fisher asserted that the repeated sampling and testing required to define error rates in the Neyman-Pearson theory must produce samples "in all relevant aspects like the one observed" [*need citation and page number*] and this is usually quite difficult with real, physically existing populations. A hypothetical population on the other hand can achieve such samples, albeit they are only hypothetical in nature. Again, in our modern version of this, think about simulation from theoretical probability distributions on a computer.

Interestingly, there has been some thought that $p-$values might also have

interpretation under epistemic probability as a weight of evidence against a hypothesis. Good (1992), after stating that he did not think epistemic probabilities have sharp values (i.e., can be applied to points) indicated that for point hypotheses one "might have to fall back either on P values, with some modification or surprise indexes" (Good, 1992, p. 599). Good offers an example of a binomial (θ, n) data model with a uniform $(0, 1)$ prior, and the null hypothesis $H_0 : \theta = 1/2$. For a range of values of binomial sample size n, observed values y, resultant $p-$values, and numerical values of the Bayes factor against H_0, the quantity $BF\, p\, n^{1/2}$ is nearly constant, meaning the $p-$values were (nearly) proportional to BF. He concludes that "Thus we have empirical evidence that sensible P values are related to weights of evidence . . ." Good suggested what he called *standardized* $p-$values, $p_{stan} = \min(1/2, p\sqrt{n}/10)$ to take into account the sample size involved in their calculation.

If one is willing to admit Fisher's argument that $p-$values depend on repeated sampling from a hypothetical but not real population, or that a $p-$value has some primitive force as a measure of evidence against a hypothesis, or that a $p-$value has meaning as a weight of evidence under epistemic probability, then tests of significance can be claimed to involve post-data precision. To conclude our discussion of tests of significance, we believe that it would not be unreasonable to take the position that, despite some clear deficiencies as a general approach for producing statistical inference, such procedures do have a certain amount of inherent appeal. The idea that a $p-$value represents a continuous measure of evidence offered by a given set of data against a hypothesis seems quite natural and intuitive, even if giving it meaning in terms of an absolute scale cannot be done in a convincing manner without invoking hypothetical repeated sampling and tail areas. As

a testing procedure, it is supported by a clear and valid logical argument from deductive logic, leading to statements of statistical inference (which, as statements, are inductive) through the introduction of probability to replace implication at a crucial step.

17.10.2 Tests for Acceptance Sampling

Tests for acceptance sampling are based on a foundation about which there is little controversy. They are designed, under certain assumptions, to first and foremost control the rate of Type I errors in repeated testing situations, and that they can do so is not a source of controversy. Given such control, they are designed to maximize power, which is not possible in all cases without some restrictions on the set of procedures with which they are allowed to compete. In situations for which control of error rates in making binary decisions is an appropriate concern, it would seem difficult to claim that they are not a solution with a great deal to recommend them. The difficulty with tests for acceptance sampling is that the situations in which control of error rates is appropriate are limited, and do not contain what are probably the majority of scientific investigations. Fisher (1955, p. 69-70) said,

> Now, acceptance procedures are of great importance in the modern world . . . I am casting no contempt on acceptance procedures, and I am thankful, whenever I travel by air, that the high level of precision and reliability required can really be achieved by such means. But the logical difference between such an operation and the work of scientific discovery by physical or biological experimentation seem to me so wide that the analogy between them is not helpful, and the identification of the two sorts of operations

> is decidedly misleading.

Egon Pearson denied that the theory developed with Neyman had acceptance sampling as its original motivation, stating that "it was not until after the main lines of this theory had taken shape . . . that the fact that there was a remarkable paralellism of ideas in the field of acceptance sampling became apparent" (Pearson, 1955, p. 204). He did, however, also acknowledge the need for what Fisher had termed a population of samples like that observed in all relevant aspects, stating that (Pearson, 1955, p. 205),

> If probability is to be justly applied to the analysis of data, it follows that a random process must have been introduced or been naturally present at some stage in the collection of these data.

As previously noted, neither Fisher nor E. Pearson's father, Karl Pearson, necessarily adhered to this principle. In some ways, then, Fisher and the elder Pearson seem to have presaged what we now call *emulation*.

For his part, Neyman seems to have eventually drifted away from both the need for repeated sampling from an existent population similar to that observed, and the concept that critical values are unconcerned with "levels of significance", although not until about 15 years after Fisher's death. The second of these was approached in a rather trivial manner in (Neyman, 1977, p. 107-108) where significance level was equated with the size of a test; $\alpha = 0.05$ corresponded to significant and $\alpha = 0.01$ to highly significant. But there was no relaxation that, within a particular test, $p = \alpha + 0.001$ leads to a diametrically opposed conclusion than $p = \alpha - 0.001$. Perhaps more meaningfully, in this same paper Neyman denied that repeated testing of the same or closely related hypotheses was necessary to provide meaning to the concept of Type I and Type II error rates. His framework in 1977

was to consider a long sequence of "situations" $\{S_i : i = 1, \ldots\}$ in which one tests hypotheses $H_{0,i}$ against $H_{1,i}$, where these hypotheses form logical disjunctions for each situation. The situations themselves could vary widely, referring to "problems in of astronomy, others to highway traffic, still others to radiation biology, some to problems of big cities and slums or to weather modification, etc." (Neyman, 1977, p. 108). Neyman then argues that if a hypothesis test for situation S_i has been conducted with an acceptable level of Type I error α_i, then the central limit theorem implies that the frequency of Type I errors in the sequence of tests will be close to the mean of the set of α_i values chosen, and leads to a similar approximation for some type of (not well defined) Type II errors. How this renders error rates relative for meaningful inference in a given situation is not explained. Again, Neyman focuses only on the mathematical properties of statistical procedures and not the logical basis for reasoning that they might provide.

We have already noted, in the section Assumptions or Hypotheses, that in order to control error rates and maximize power, Neyman-Pearson procedures must take all aspects of model formulation other than the null hypothesis and make them assumptions required for the procedure to be valid. In addition, the physical repeated testing situations must all adhere to these assumptions to a reasonable degree. This, we believe, is what Fisher meant when he claimed the repeated sampling required by tests for acceptance sampling must be from populations that produce samples that are the same in all relevant aspects, as mentioned in the subsection on Tests of Significance just ended. This is difficult to achieve but is probably most easily justified in the context of manufacturing settings, which is where the term acceptance sampling originated. In this type of situation, constructed sampling units may well form groups that can be considered as populations that are the

same in all relevant aspects, at least for finite time frames. These are the situations that we believe lend themselves most readily to analysis using tests for acceptance sampling, as demonstrated by the examples of Chapter 17.4.2.

Control of error rates is concerned with pre-data precision, and Tests for Acceptance Sampling lack a mechanism for providing a data-based measure of evidence about the hypotheses under consideration. There have been arguments presented by adherents of this approach to the effect that concern with error rates can somehow lead to a type of inductive *behavior* if not actual inductive *reasoning*, and that this serves the purposes of scientific investigation. Whether there is more to these arguments than mere semantics remains an open question. What we can say for certain is that the approach of tests for acceptance sampling is really well suited for problems that involve acceptance sampling. Perhaps that should be sufficient.

17.10.3 Fisher, Neyman-Pearson and Deductive Syllogisms

The influential philosopher of science, Karl Popper, proposed a theory of how science makes progress that was based on two foundational principles. The first was the assertion that all logic is deductive; there is really no inductive logic. The second principle was that theories can be disproven, but never proven. The theories Popper dealt with were primarily deterministic scientific theories, not those of statistical tests, but he did mention statistical hypotheses once or twice, and there certainly are analogies. Most authors who have drawn connections between the ideas of Popper and the use of tests in statis-

tics have aligned Popper with Neyman because both staunchly insisted that all valid arguments were deductive in nature, while Fisher espoused tests of significance as a vehicle for inductive reasoning (e.g., Lehmann, 1995). And this certainly is a huge difference between the philosophies of Fisher and Popper. Yet, the basic concept of Fisher's tests of significance is remarkably similar to Popper's theory of science. That is, in a test of significance, there is a single hypothesis about a parameter that the scientist is generally hoping to reject, such as the situation of Example 17.2. A statistical test produces a result that provides evidence *against* the hypothesis. If that evidence is strong enough, we *reject* the hypothesis. Otherwise, we *fail to reject* the hypothesis, but we never accept the hypothesis. In this regard, Popper is more closely aligned with the approach of Fisher than with that of Neyman-Pearson in which we accept either the null or alternative hypotheses.

We find it interesting that, although Neyman claimed that all reasoning is deductive, there does not appear to be a formal argument in deductive logic that supports the full procedure involved in acceptance sampling. And, although Fisher claimed that Tests of Significance involved inductive reasoning, there is a formal deductive argument that seems to fit his procedure. Consider propositions A, B, and C. An argument called *modus tollens* in logic is,

- Posit A.
- If A, then C.
- Not C.
- Therefore not A.

These steps match the procedure involved in tests of significance after changing *Not C* to Fisher's famous *either something extremely rare has occurred or Not C*. Consider a normal one sample problem and the hypothesis that $\mu = \mu_0$. Then we have

- $H : \mu = \mu_0$.
- If H, then T has a $t-$distribution with $n-1$ degrees of freedom.
- Given a small $p-$value, conclude that T does not have a $t-$distribution with $n-1$ degrees of freedom.
- Therefore, reject H.

This deductive argument does not match the Neyman-Pearson procedure used in tests for acceptance sampling because there is no path for acceptance of H (or A in the general syllogism). The conclusion 'Therefore not A' is a deductive statement in the general syllogism that follows absolutely from 'Not C', while the conclusion 'Therefore, reject H' is an inductive statement in the version appropriate for a test of significance, because in the statistical version we run into Fisher's dichotomy that either a rare event has occurred, or T does not follow the reference distribution derived under H. Our conclusion can therefore not be made with certainty, although the argument for reaching it is clearly deductive in form.

If one insists that Fisher's tests of significance must involve an implied alternative hypothesis that is the negation of what now becomes a null hypothesis, the logical syllogism is expanded as,

- Either A or B (either H_0 or H_1)
- If A then C (if H_0 then $T \sim t(n-1)$)

- Not C (small $p-$value)
- Therefore not A (reject H_0)
- Therefore B (accept H_1)

This version of the syllogism leads to a conclusion of accepting H_1 (B), but still does not provide a means for accepting H_0 (A). Thus, it does not apply to tests for acceptance sampling any more than the previous unexpanded version.

17.10.4 Tests Based on Support

Tests Based on Support are clearly post-data procedures, constructing a measure of evidence literally by definition. Again, note that this is a consequence of interpretation, not the algebraic form of the test statistics used. Given that likelihoods and log likelihoods quantify the degree to which a given set of data support specified parameter values, a ratio of likelihoods or a difference of log likelihoods constitutes a clear measure of relative support for opposing hypotheses or models used to construct the measure.

The principle issue that arises in the use of support tests is also quite clear, being how to determine an appropriate scale by which to judge numerical values of relative support. While the notion of likelihoods as measures of support is based in probability, likelihoods are not themselves probabilities that apply to possible values of parameters. Edwards, the foremost proponent of the use of support, claimed that statisticians and scientists would become accustomed to judging differences in support against an arbitrary scale, in the same way that we understand differences in temperatures relative to the arbitrary scales of degrees Centigrade or degrees Fahrenheit. It

does not appear that statisticians and scientists have done so. While arbitrary benchmarks such as 2 may make it easy to dismiss tests based on support as ungrounded, doing so would be a mistake. First, one can argue that scales against which to judge relative support are no less arbitrary than judging $p-$values against values of 0.10 or 0.05, they are just more honestly arbitrary. In addition, the relative portion of relative support invites us to consider competing hypotheses more as problems of model selection than problems of making statements about values of unknown parameters. This is particularly attractive in problems for which distributional parameters are not direct translations of meaningful attributes in the real world.

If one accepts the concept of tests as a vehicle for model comparison, then discarding the desire to formulate hypotheses that define a logical disjunction also becomes quite palatable. There really is no reason that there should be only two models possible for any given situation. The notion that tests of significance might allow a hypothesis that includes elements of problem formulation beyond the value of a parameter is also in concert with this viewpoint, although perhaps in a bit different way that does not involve an explicit alternative, either hypothesis or model.

Another major departure from other approaches to testing hypotheses, including that of likelihood ratio tests to follow, is the manner in which support tests depend on probability. This is actually the source for interpretation of likelihoods and log likelihoods as measures of the support contained in a set of data for different models. Probability enters relative support by virtue of the fact that likelihoods, while functions of parameters, are numerically proportional to the probability of a given set of data under the model with those parameter values. This is fundamentally different than the way in which the other approaches rely on probability, which is through sampling distributions

of estimators and test statistics constructed from those estimators.

In conclusion, support tests do not offer any more of a generally acceptable procedure for making statements of inference than either tests for acceptance sampling or tests of significance. They do, however, give a completely data-centered and evidential interpretation to test statistics that are often algebraically equivalent to those used in the other approaches and that, at least under tests for acceptance sampling, draw meaning only from repetition of the procedure. As such, they provide a valuable piece for the overall puzzle presented by the problem of making statistical inference.

17.10.5 Likelihood Ratio Tests

We provided a transition from tests based on support to likelihood ratio tests by first presenting the results of Wilks on asymptotic distributions of statistics that are equal to twice relative support (see Chapter 17.6), as a means to choose benchmark values for support tests. We then took the step of abandoning the support interpretation of these test statistics for one that relies instead on sampling distributions. The resulting procedures of likelihood ratio tests thus become procedures that rest on the same probability basis as tests for acceptance sampling and tests of significance.

In some ways, then, Likelihood Ratio Tests become something of a double agent. The test statistics used are proportional (by a factor of 2) to measures of relative support, and could be interpreted as such. At the same time, the approximate sampling distributions allow the computation of $p-$values that can be interpreted in much the same way as $p-$values in tests of significance, with one extremely important distinction. A $p-$value computed for a likelihood ratio test may be interpreted as evidence against the reduced model

only relative to the full model. We usually reverse this to be evidence in favor of the full model over the reduced model. It might be possible to use $p-$values in likelihood ratio tests as a scale for expressing critical values as in tests for acceptance sampling but, except in cases where the reduced and full models constitute a partition of an overall parameter space, it would be difficult to define Type I and Type II errors, and reach conclusions in the same manner as in tests for acceptance sampling.

17.10.6 Bayesian Tests

Bayesian treatments of hypothesis testing are clearly the most evidence-based and post-data precision procedures of any of the approaches we have considered. As shown previously, if we take a decision-theoretic approach under $0-1$ loss, then the Bayesian procedure to test one inferential statement S_1 against another S_2 is a comparison of the posterior probabilities $Pr(S_1|\boldsymbol{y})$ and $Pr(S_2|\boldsymbol{y})$ and we prefer whichever has the higher probability. This basic procedure avoids seemingly all of the pitfalls of the other testing prescriptions we have covered. It does not require disjunctive hypotheses (models), although one can make up silly hypotheses that would be meaningless to test. It does not require repeated sampling, either real or hypothetical, for interpretation. And it and provides an absolute scale for judgment, that being probability.

Although Bayesian testing procedures involve post-data precision by nature, they are turned into quasi-pre-data precision procedures when scales provided for interpretation are used as absolute dividing lines rather than helpful guideposts. Thus, if one declares that a Bayes factor of 2.9 provides no evidence in favor of one model over another, but a Bayes factor of 3.1 pro-

vides some evidence, we have vitiated the post-data characteristics of Bayes factors. This is actually also true if we assert that we will accept whichever of two hypotheses has higher posterior probability, as in the previous paragraph. The result is that one hypothesis with posterior probability 0.5001 is accepted while the other with posterior probability 0.4999 is not. This is exactly the same phenomenon that occurs with $p-$values. If one is willing to allow an evidential interpretation to $p-$values, claiming to conduct a Fisherian test of significance but then using $p-$values with an absolute significant/not significant criterion of 0.05 is simply a Neyman-Pearson hypothesis test with critical points expressed in terms of $p-$values rather than values of test statistics. So concern with post-data precision is not really an inherent property of Bayesian testing, it is something that is allowed by Bayesian test procedures. In this sense, then, the distinction between a Bayesian test and a Fisherian test is only whether one is willing to accept the inferential summary as a measure of the evidence that a given set of data provide about a hypothesis. For the $p-$value this is neither crystal clear nor universally accepted. For posterior probabilities it is undeniable unless one is unwilling to accept an epistemic concept of probability.

Given the simplicity and appeal of comparing posterior probabilities for alternative inferential statements, one wonders why Bayes factors have garnered so much attention. Under the definition of Bayes factors as relating prior odds of two models to posterior odds, if $Pr(M_1)/Pr(M_2) = 1$ then Bayes factors are equal to the posterior odds $Pr(M_1|\boldsymbol{y})/Pr(M_2|\boldsymbol{y})$, and they are then the same thing as comparing posterior probabilities. They can sometimes be computed in situations for which prior and/or posterior model probabilities are difficult to arrive at or don't exist (this later not without some controversy). They are flexible enough to be applicable in a wide vari-

ety of specific problems. And yet, in moving away from preferring whichever inferential statement (or model or hypothesis) has greater posterior probability they then require a new scale for interpretation, and this reintroduces an arbitrary element to the procedure. Such scales could better be replaced with windows of indifference left on the numerical $0-1$ scale of probability, but this would be functional only in situations for which prior odds are unity and discarding Bayes factors for anything beyond cases in which they are equal to posterior odds. As we have seen in Chapter 14, Bayes factors are also subject to heavy influence by prior specifications that can sometimes cause them to lead to inferential conclusions that are not in agreement with inferential statements made directly from the posterior. For this last reason, in particular, we view Bayes factors with a bit of skepticism.

Overall, conducting tests on the basis of Bayesian procedures has a great deal to recommend it. There is no real controversy that they are based on measures of evidence provided by a given observed data set. We have expanded the setting of testing hypotheses about data model parameters to a broader context of comparing potential inferential statements. This is facilitated by our ability to simulate from posterior and posterior predictive distributions so that we may, by means of Monte Carlo, provide probabilities for a wide range of events such as $\theta \in (\theta - \delta_1, \theta + \delta_2)$, $(\theta_1 \leq \theta_2)$, $\min\{|\alpha_j| \leq \delta;\, j = 1, \ldots, K\}$, and a host of others involving one group, two groups, multiple groups, quantiles and characteristics of data. Thus, our preference is to rely on posterior probabilities for potential inferential statements for which they can be computed directly, such as hypotheses concerning parameter values. For model selection, we rely primarily on posterior predictive probabilities that reflect how well a fitted model represents meaningful structure in the observed data.

17.10.7 Goodness of Fit Tests

Goodness of fit tests defy attempts to put their inferential basis into a neat organization based on concepts of probability and pre-data versus post-data precision. While it is quite natural for us to think of interpretation of $p-$values in GOF tests according to either the traditions of Fisherian or Neyman-Pearson testing, Karl Pearson certainly did not. One has to remember that he pre-dates the development of either of those schools of thought, writing about the role of probability in scientific reasoning in the 1890s (Pearson, 1892) and at that time inverse probability was the standard for reasoning. Pearson seems to have accepted inverse probability, although he seldom used it in applications and insisted that prior probability distributions have a frequency interpretation based on previous experience (Inman, 1994). Thus, Pearson's view of probability included epistemic content but such that "A measure of my belief in the occurrence of some event in the future is thus based upon my statistical experience of its occurrence or failure in the past." and "The means by which statistics are taken in practical life are human and they become subjective and individual in the process of taking and applying them." (Pearson, 1941, p. 93-94). Note that these quotes, while taken from a 1941 article in Biometrika, are from a set of lectures on the theory of probability delivered by Karl Pearson in 1892. The 1941 paper is a printed version of the first of these lectures, given on November 1 1892 at Gresham College. So how did Pearson interpret a $p-$value associated with a GOF test? The key is a randomization argument, but one that is quite distinct from an assumption that the data were generated by some random mechanism. In essence, Pearson seemed to overlay a hypothetical random generating mechanism on the results of a GOF test, regardless of whether

the data were actually obtained from such a mechanism or not. This then allowed a frequentist interpretation of a $p-$value such as 0.79 that you have a "reasonably good graduation because 79 percent of random samples would, were the "hypothesis" true, give a worse result than the observations do." (Pearson, 1935, p. 550). The application Pearson was referring to were observations of orbital motions of stars, which were clearly not obtained from a random sampling plan, nor could the problem be subject to repeated sampling. As Inman (1994, p.9) observes, "In terms of a test of significance, this distinction appears to make the assumption of random sampling part of the hypothesis under test instead of a necessary condition for the statistical test." And the random sampling under consideration is not necessarily actual random sampling, but part of a hypothetical structure under which to interpret results.

Where does this leave us in terms of understanding the logical content of a GOF test? Unfortunately, the answer seems to be 'left wandering in the wilderness'. The strategy of reaching a decision between two hypotheses that form a logical partition is clearly not applicable. The strategy of interpreting a $p-$value as a measure of evidence against a hypothesis has some force, but our usual guideposts for declaring $p-$values as meaningful under that strategy are poorly suited for use with GOF tests. We offer additional thoughts relative to the use of GOF tests in Chapter 19.

Chapter 18

Interval Estimation

We turn now to the topic of constructing interval estimates of parameters. On the frequentist side of the coin are the classic confidence intervals of Jerzy Neyman, other intervals constructed from sampling distributions (especially asymptotic), in certain specific problems Fisher's fiducial intervals, and intervals based on the use of likelihoods as support functions. On the Bayesian side are credible intervals, either central or highest posterior density.

18.1 A Historical Note

There appears to have been a good deal of confusion about the interpretation of interval estimates shortly after R.A. Fisher promoted his fiducial argument beginning in 1930 and Jerzy Neyman introduced his construction of confidence intervals in 1933. At the time, intervals were the purview of methods of inverse probability (the term Bayesian had not yet been coined). According to Bartlett (1984) both Fisher and Neyman were working to show that intervals could be constructed without recourse to formal use of a prior

distribution. Neyman (1977, p. 128) recounts his early attempts as starting with a prior distribution and then deriving an interval which would be independent of the prior. In the mid 1930s, it was known that for certain problems, the resulting interval is identical under all three approaches, what we now call Neyman's classical intervals, what we now call Bayeisan credible intervals, and Fisher's fiducial approach. This contributed to misconceptions that Neyman's confidence intervals were essentially re-formulations or modifications of Fisher's fiducial intervals, or that Fisher's intervals were the same as intervals constructed from inverse probability with certain priors. Part of the difficulty also arose from rather incomplete exposition of the philosophical differences between the approaches of Fisher and Neyman, which were still very much in the early stages of development. Over the subsequent decade, distinctions became clearer. It can also be argued that the dispute between Fisher and Neyman regarding the foundational basis for their intervals led to an explosion in personal animosity that is well recognized in the history of statistics.

To illustrate a simple problem in which the approaches based on inverse probability, Fisher's fiducial argument, and Neyman's confidence intervals all produce the same result, consider a random sample $\{y_1, \ldots, y_n\}$ from a normal distribution with expected value μ and known variance σ^2. With a uniform (improper) prior on μ, the posterior distribution of μ is normal with mean $\bar{y}$ and variance σ^2/n. An $\alpha(100)\%$ credible interval for μ is then,

$$Pr(|\mu - \bar{y}|) = 1.96\sqrt{\sigma^2/n}. \tag{18.1}$$

Following the fiducial argument outlined by Fisher (1930), **?** shows that in this problem the fiducial distribution for μ is the same as the posterior just given, namely, $N(\bar{y}, \sigma^2/n)$ and the 95% fiducial interval for μ is again (18.1).

Jeffreys (1940) comments that "In fact, the fiducial argument, when completed, and the inverse probability argument, are simply different ways to saying the same thing; the hypotheses are identical and so are the results." We briefly treat the fiducial argument which, at best, lacks in general applicability, in the sequel. The importance of it for our current topic is that it represents the first attempt to present interval estimation from a frequentist viewpoint although, as just indicated, the distinction between fiducial and posterior probability was not always recognized. Neyman (1934) first introduced his theory of confidence intervals as a secondary topic in a much larger paper contrasting random sampling with purposeful sampling. But the solution for a 95% confidence interval for μ in a normal one sample problem was again (18.1). As the discussions connected with that paper show, neither Neyman nor Fisher were fully aware of the wide gulf between their approaches and there was an indication that Neyman considered his development to have generalized the argument of fiducial probability, extending the limited situations under which Fisher applied his approach. However, the first hints of how radically different the two approaches are does appear in Neyman's reply to discussion pieces in that 1934 paper.

18.2 Neyman's Confidence Intervals

Neyman's motivation in the development of confidence intervals was the same as in the development of most powerful tests, namely to develop the theory of statistics as arising from a solid mathematical basis. As Bartlett (1984, p. 170) points out, Neyman rejected an epistemic concept of probability such as that promoted by Jeffreys (1931), wanting a pure mathematical basis for statistics to develop as a discipline. He also rejected likelihood as a

measure of support or evidence and felt the fiducial argument of Fisher "is a conglomeration of mutually inconsistent assertions, not a mathematical theory" (Neyman, 1977, p. 100).

Suppose $Y_1, \ldots, Y_n$ denotes a random sample from some distribution $f(y|\boldsymbol{\theta})$ depending on a parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ and our interest is in producing an interval estimate of θ_1. The "problem of the confidence interval" was presented by Neyman (1977) as determination of functions $L(\boldsymbol{Y})$ and $U(\boldsymbol{Y})$ such that, if θ_1^0 denotes the true value of θ_1, then,

$$Pr[L(\boldsymbol{Y}) \leq \theta_1^0 \leq U(\boldsymbol{Y})] = \alpha, \tag{18.2}$$

for some specified $0 < \alpha < 1$ and regardless of what the values of $\theta_2, \ldots, \theta_p$ are. For a fixed parameter θ_1, the functions $L(\boldsymbol{Y})$ and $U(\boldsymbol{Y})$ in (18.2) are random variables, and when they are replaced with observed values from a sample, we obtain an observed confidence interval and assert that, with confidence α,

$$L(\boldsymbol{y}) \leq \theta_1^0 \leq U(\boldsymbol{y}), \tag{18.3}$$

a statement that no longer contains any random quantities and for which the probability is either 0 or 1. This then, is the origin of the use of the word confidence rather than probability to label the interval. It was recognized that for a scalar parameter $\boldsymbol{\theta} = \theta_1$, there exists an infinite number of intervals satisfying (18.2). If the dimension of $\boldsymbol{\theta} \geq 2$, then the solution requires the existence of a set of sufficient statistics for $\theta_2, \ldots, \theta_p$ and then there are again an infinity of solutions.

Neyman expended a great deal of effort in developing some notion of optimality for intervals that satisfy (18.2), an obvious choice being length of the interval. This was presented by Neyman as two separate one sided intervals $L(\boldsymbol{y}) \leq \theta_1^0$ and $U(\boldsymbol{y}) \geq \theta_1^0$ having the properties that, for any

$\theta_1' < \theta_1^0$ and $\theta_1'' > \theta_1^0$,

$$Pr[L(\boldsymbol{y}) \leq \theta_1'] = \min_{\tilde{L}(\boldsymbol{y})} Pr[\tilde{L}(\boldsymbol{y}) \leq \theta_1' | \theta_1 = \theta_1^0]$$
$$Pr[L(\boldsymbol{y}) \geq \theta_1''] = \min_{\tilde{L}(\boldsymbol{y})} Pr[\tilde{L}(\boldsymbol{y}) \geq \theta_1'' | \theta_1 = \theta_1^0].$$

It turns out that these properties can be produced in some problems in the same manner as determination of uniformly most powerful one sided tests. Also similar to the development of Neyman-Pearson hypothesis tests, for a simultaneous two sided interval the shortest interval does not exist unless the class of intervals considered is restricted to be unbiased, which for intervals means that for any $\theta_1' \neq \theta_1''$,

$$Pr[L(\boldsymbol{Y}) \leq \theta_1' \leq U(\boldsymbol{Y}) | \theta_1 = \theta_1'] = \alpha \geq Pr[L(\boldsymbol{Y}) \leq \theta_1' \leq U(\boldsymbol{Y}) | \theta_1 = \theta_1''],$$

which says that the probability the interval covers a value θ_1' is the greatest (and equal to α) when θ_1' is the true value of θ_1. Shortest unbiased intervals can be determined for problems in which the dimension of $\boldsymbol{\theta}$ is small (1 or 2), again relying on results from the development of tests under the approach of Neyman and Pearson.

Neyman repeatedly made it clear that the use of confidence intervals rested on a decision to assert that for an observed interval $[L(\boldsymbol{y}), U(\boldsymbol{y})]$ the parameter of interest, θ_1 in this section, lies within the interval. The statement should have nothing to do with *plausible* values of the parameter or degrees of belief about where it might lie, the correct conclusion is that the parameter is, in fact, contained in the interval. That this conclusion might be in error makes it an inductive statement, but not the result of any type of inductive reasoning or assessment of evidence. Like tests, an interval is a mechanism for making *decisions*.

> . . . we may decide to behave as if we actually knew that the true value [is in the interval]. This is done as a result of our decision and has nothing to do with 'reasoning' or 'conclusion'. The reasoning ended when the functions [used to give endpoints of the interval] were calculated. Neyman (1941, p. 134)

The final sentence of this quote holds the key concept, namely that the force of confidence intervals is based on *the procedure*, not any particular interval that we may compute. What the procedure does is allow us to decide to behave as if the true value is contained in that interval. Decide absolutely, not with some degree of "belief" or "plausibility". And, if this is our rule for determing behavior, the procedure indicates that we will be correct $(1-\alpha)100\%$ of the time in the long run. This is a statement of precision that is made "pre-data". That is, it does not depend on having observed anything in particular, nor does it change from having observed something in particular.

There are many alternative interpretations of confidence intervals that have been attempted, almost all of which are not correct. In particular, if one given interval is shorter than another given interval that does not imply that the first interval has been estimated with greater precision than the second. Consider a situation in which we obtain a sample of size n from a normal distribution with mean μ and variance σ^2. We divide our sample into two equally sized sets of size $n/2$ and compute a standard 95% confidence interval using the sample means and variances of the two groups and the 0.975 quantile of a $t-$distribution with $n/2-1$ degrees of freedom. One of our two intervals will be shorter than the other. But clearly, that interval has no greater precision than the wider interval – they both came

from the same sample. The mistake is thinking that precision is a property of an estimate (given interval) rather than the estimator (random version of interval endpoints) that produced it. Note that it is true that a procedure that produces intervals that are shorter on average is more precise than another procedure that produces intervals that are wider. But this correct statement is very different than saying an interval (3.2, 4.1) gives a more precise estimate of the parameter than does an interval (2.5, 4.6), which is not a correct conclusion. Similarly, it is sometimes asserted that values farther from the center of a equal-tailed confidence interval are less likely or less plausible as values for the unknown parameter than are values closer to the center. This fallacy is easily dispelled with a simple Monte Carlo exercise that is rather fun to conduct as a class exercise. Perhaps the most easily discredited misnomer about the interpretation of confidence intervals is that a parameter lies inside a given interval with a specified probability. This is taking an observed confidence interval (18.3) and ascribing to it the properties of (18.2). A related misconception, on the other hand, seems to be among the most difficult for many people to grasp is incorrect, and that is the notion that values inside an interval are more *plausible* as values of the true parameter than are values outside the interval. The difficulty here is that the word plausible invokes an interpretation of belief, both figuratively and literally. As already discussed, there is no notion of belief, likelihood (in a general sense), or evidence in the construction of confidence intervals. In summary, incorrect notions in the interpretation of confidence intervals involve (1) statements about *an interval*, not the procedure used to obtain the interval, and/or (2) claims about the unknown and unobservable true parameter, not the data.

Interestingly, although it does not appear to have been the genesis for

Neyman's development of confidence intervals, perhaps the most straightforward interpretation of confidence intervals rests on the relation between such intervals and Neyman-Pearson hypothesis tests of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. In particular, we can see that any hypothesized value θ_0 that lies in the interior of a $(1 - \alpha)100\%$ confidence interval would result in acceptance of H_0 with a specified test size of α, while any hypothesized value θ_0 outside the interval would result in acceptance of H_1. In this way, the classical confidence intervals of Neyman can be thought of a nothing more than a family of hypothesis tests.

18.3 Likelihood Intervals

As was also the case for testing, likelihoods can be used in several ways to construct intervals. We cover these in turn.

18.3.1 Likelihood Support Intervals

One way to obtain intervals from likelihood functions is to consider interpretation of likelihoods or log likelihoods as support functions. Consider random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ having distribution f with a scalar parameter θ. In Chapter 17.5 we defined the relative support in a set of observed data $\boldsymbol{y}$ in favor of θ_1 over θ_2 as,

$$r(\theta_1, \theta_2) = \log\{f(\boldsymbol{y}|\theta_1)\} - \log\{f(\boldsymbol{y}|\theta_2)\}. \tag{18.4}$$

As for tests, the most highly supported value of $\theta \in \Theta$ is the maximum likelihood estimate $\hat{\theta}_n$. We can form a Δ_S interval of support for θ as,

$$\{\theta_0 : r(\hat{\theta}_n, \theta_0) \leq \Delta_S\}. \tag{18.5}$$

Here, we interpret the interval as containing all values of the parameter for which the support differs by no more than Δ_S of the value most highly supported by the data, namely the mle $\hat{\theta}$. This interpretation is clearly a post-data statement of precision, drawing its meaning directly from the fact that the likelihood is proportional to the probability of the data under whatever parameter value at which it is evaluated. And here we can go further than what is available for many classical (i.e., Neyman) confidence intervals and state that values in the interval further away from the maximum likelihood estimate (which may or may not be the center) have less support from the data than values nearer the maximum likelihood estimate. Note here that this is a statement about the *data* not the true parameter value.

If the distribution of response random variables depends on a vector-valued parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, we can make use of normed profile likelihoods as discussed in the book *Intermediate Statistical Methods*. Suppose we desire a Δ_S interval of support for an element of $\boldsymbol{\theta}$, θ_j, say. To construct such an interval, let $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^T$ and replace $r(\hat{\theta}_n, \theta_0)$ in (18.5) with the log normed profile likelihood,

$$\ell_p(\theta_{j,0}) = \sup_{\boldsymbol{\theta}_{-j}} \log\{f(\boldsymbol{y}|\theta_{j,0}, \boldsymbol{\theta}_{-j})\} - \log\{f(\boldsymbol{y}|\hat{\theta}_n)\}.$$

18.3.2 Intervals from Inversion of LRT

We are familiar with constructing intervals for scalar parameters from inversion of likelihood ratio tests,

$$\{\theta_0 : -2[\ell(\theta_0|\boldsymbol{y}) - \ell(\hat{\theta}|\boldsymbol{y})] \leq \chi^2_{1-\alpha}(1)\}, \tag{18.6}$$

where $\ell(\theta|\boldsymbol{y}) = \log\{f(\boldsymbol{y}|\theta)\}$. While superficially similar to an interval of support, as it would appear all we have done is multiple by -2, there is

a potentially dramatic difference in interpretation between (18.5) and (18.6) caused by the selection of the benchmark value as an arbitrary Δ_S on the one hand or a quantile of the approximate (asymptotic) sampling distribution of the log likelihood ratio as $\chi^2_{1-\alpha}(1)$ on the other. As soon as we invoke the sampling distribution of a statistic to provide interpretation, we place ourselves in the position of relying on at least hypothetical repeated sampling to provide meaning and the focus changes from *these data* to the behavior of the procedure under repeated sampling. Intervals constructed through the inversion of likelihood ratio tests as in (18.6) are properly interpreted in the same manner as the classical confidence intervals of Neyman. The additional interpretation that values farther from the maximum likelihood estimate have less support in the data than values closer to that estimate does continue to hold for this type of interval estimate, and one could certainly claim that is a post-data characteristic. So similar to likelihood ratio tests, intervals constructed from inversion of likelihood ratio tests do not necessarily fit neatly into one of the pre-data post-data categories.

It is not unreasonable to wonder, if Δ_S in (18.5) is arbitrary, why we could not arbitrarily choose it to be $\Delta_{LR} = \chi^2_{1-\alpha}(1)$ and retain the interpretation of an interval of support. The potential pitfall in this prescription is to believe both that Δ_{LR} is arbitrary and that it is motivated by the sampling distribution of the likelihood ratio test statistic. If it is not desired to rely on the sampling distribution for interpretation, then why use the factor -2 in the definition of (18.6)? How does one then motivate arbitrary choices of $\Delta_{LR} = 3.8414$ (the 0.95 quantile of a $\chi^2(1)$ distribution or $\Delta_{LR} = 2.7055$ (the 0.90 quantile of a $\chi^2(1)$ distribution? Why not choose $\Delta_S = 4$ or $\Delta_S = 3$, which seem like much more reasonable choices unless one is concerned that the sampling distribution needs to be adhered to? So trying to use the

sampling distribution of the likelihood ratio to select a benchmark value and at the same time claim that repeated sampling and behavior of the procedure are not now central to interpretation is the classic scenario of wanting to have your cake and eat it too.

As for intervals of support, when the parameter is vector-valued, intervals constructed from inversion of likelihood ratio tests for individual elements of the parameter can be obtained through the use of normed profile likelihoods, with a benchmark value selected based on the approximate sampling distribution of that profile likelihood. See *Intermediate Statistical Methods* for details.

18.4 Fisher's Fiducial Approach

We include a brief overview of the fiducial approach to interval estimation for completeness and because it played a rather important role in the historical development of intervals. We do not claim to delve into all of the intricacies involved, or to provide a totally clear explanation of the motivation or justification that Fisher believed he had discovered. We are not certain anyone has arrived at such an explanation, including Fisher himself. We also note that Fisher's own formulation and defense of fiducial analysis evolved considerably over time, both in terms of what was required and in terms of justification (see Zabell, 1992).

Fisher's goal in the development of the fiducial approach was to arrive at a purely post-data inferential procedure in which one can make inference about a parameter θ in the light of only the data at hand, but to do so without requiring a prior distribution. Fisher (1922) professed the belief that probabilities corresponded to relative frequencies in hypothetically infinite

populations. And yet, he also asserted that the fiducial argument allowed probability statements to be made about the values of fixed but unknown parameters (e.g., Fisher, 1930, p. 533). As Zabell (1992, p. 374) postulates, Fisher seemed to view probability as a concept defined in terms of objective relative frequencies but one that also has connotation in terms of an epistemic measure of rational belief.

The original formulation of the fiducial argument was put forth by Fisher (1930) and had the following form. Suppose that T is a continuous statistic whose distribution depends on only a single parameter θ. Let $F(t, \theta) = Pr(T \leq t|\theta)$. Suppose that $F(t, \theta) = p$ implicitly defines functions $t_p(\theta)$ and $\theta_p(t)$ such that $F(t_p(\theta), \theta) - p = 0$ and $F(t, \theta_p(t)) - p = 0$. Then, as long as $\theta_p(t) \leq \theta$ if and only if $t \leq t_p(\theta)$,

$$Pr[\theta_p(T) \leq \theta] = p$$

is a statement of fiducial probability about θ. Note that the condition on the relation between $\theta_p(t)$ and $t_p(\theta)$ simply says that they increase or decrease together. The fiducial $100(1-p)$ percent value corresponding to t was defined by Fisher to be $\theta_p(t)$.

By 1935, Fisher's formulation of the fiducial argument had evolved to include the need for T to be a pivotal and constructed from a sufficient statistic. For a simple problem, Fisher's modified fiducial argument proceeds as follows (cf. Fisher, 935a; Edwards, 1976; Sidenfeld, 1992). We will take D to represent our background statistical knowledge, such as that $X \sim N(\mu, 1)$. We know, for this model that X is sufficient for μ, and $Z = (X - \mu) \sim N(0, 1)$ is a pivotal quantity. These two conditions were required by Fisher, sufficiency to produce uniqueness, and a pivot containing the sufficient statistic to be used in a modified version of the original 1930 formulation. Given Z and its

distribution we have the direct probability statement $Pr(0 < Z|D) = 0.50$, for example. Now, if we observe $X = x$, we still have the same probability statement about Z, namely $Pr(0 < Z|D, X = x) = Pr(0 < Z|D) = 0.50$. So observation of $X = x$ has provided no information about Z. But we also know that $Z = (X - \mu)$ so that the event $[0 < Z|D, X = x]$ may be written as,

$$[0 < Z|D, X = x] = [0 < (X - mu)|D, X = x] = [0 < (x - \mu)|D]. \quad (18.7)$$

Thus, if $Pr(0 < Z|D, X = x) = 0.50$ then we must also have $Pr(x - \mu|D) = 0.50$. This is a probability statement containing no random variable and is then a statement of fiducial probability about μ, in an epistemic context. We are, of course, interested in more than singular probability statements about μ, and so are interested in the fiducial probability distribution for μ given X which is fairly immediately,

$$\begin{aligned} Z|D, X = x &\sim \text{ N}(0, 1) \\ \Rightarrow (x - \mu|D, X = x) &\sim \text{ N}(0, 1) \\ \Rightarrow (\mu|D, X = x) &\sim \text{ N}(x, 1). \end{aligned}$$

So the fiducial distribution of μ given $X = x$, $p(\mu|x)$, is normal with expected value x and variance 1. As such, a 95% fiducial interval for μ is

$$(x - 1.96,\, x + 1.96),$$

which agrees exactly with a classical Neyman confidence interval and a 95% credible interval resulting from an improper uniform prior on μ. The difference with the classical interval is that we may properly assert that $(x - 1.96 < \mu < x + 1.96)$ as a statement of fiducial probability about μ. The

difference with the credible interval is that no prior based solely on belief was used to construct the fiducial interval.

What may be called the *direct* fiducial argument then has the following basic form. Let $\boldsymbol{X}$ denote a collection of independent and identically distributed random variables with distribution depending on a parameter θ and let $T = T(\boldsymbol{X})$ denote a sufficient statistic for θ. Let $Z(T, \theta)$ be a pivotal quantity based on T and let E be some event specified in terms of $Z(T, \theta)$, such as $Z(T, \theta) \leq c$. Let D be, as before, the set of statistical information available including, but not necessarily limited to the distribution of $\boldsymbol{X}$. Then, under the principle of irrelevance,

$$Pr[Z(T,\theta) \in E|D] = Pr[Z(T,\theta) \in E|D, \boldsymbol{X} = \boldsymbol{x}] = Pr[Z(t,\theta) \in E|D]. \tag{18.8}$$

Solving $Z(t, \theta)$ for θ then makes (18.8) into a statement of fiducial probability about θ.

Example 18.1

Suppose that random variables $Y_1, \ldots, Y_n$ are independent and identically distributed with a common exponential distribution having parameter $\beta > 0$,

$$f(y|\beta) = \beta \exp(-\beta y); \;\; y > 0.$$

The distribution of $T = \bar{Y} = (1/n) \sum_{i=1}^{n} Y_i$ is a gamma distribution with parameters n and β, and the distribution of $Z = \beta\bar{Y}$ is gamma with parameters n and n,

$$g(t) = \frac{n^n}{\Gamma(n)} t^{n-1} \exp(-nt); \;\; t > 0.$$

A simulated sample of size $n = 25$ from an exponential distribution with parameter $\beta = 2$ resulted in $\bar{y} = 0.7518$. To construct a 95% fiducial interval

we would determine the 0.025 and 0.975 quantiles of the distribution of T, which are $q_0.025 = 0.6471$ and $q_{0.975} = 1.4284$. Since $t = \beta\bar{y}$, the fiducial interval for β is

$$(q_{0.025}/\bar{y},\ q_{0.975}/\bar{y}) = (1.6325, 3.6032).$$

For comparison, an approximate 95% confidence interval based on asymptotic normality of the maximum likelihood estimate $\hat{\beta} = 1/\bar{Y}$ is (1.5337, 3.5114), which compares favorably. The distinction is in terms of what we can say relative to interpretation. To produce the entire fiducial distribution of β for this example, we first compute a sequence of quantiles $\boldsymbol{q}_T = \{q : q_0.01, q_0.02, \ldots, q_0.99\}$ for the distribution of T, which is gamma$(25, 25)$. These are then transformed into quantiles of the fiducial distribution of β as $\boldsymbol{q}_\beta = \boldsymbol{q}_t/\bar{y}$. A plot of the fiducial distribution function, namely the probabilities $(0.01, 0.02, \ldots, 0.99)$ against the quantiles $\boldsymbol{q}_\beta$ is shown in Figure (**??**), on which the asymptotic normal distribution associated with maximum likelihood is also shown. It can be seen from this figure that the fiducial and the approximate normal-based sampling distribution are quite similar.

To generalize the basic idea of using a pivotal quantity to formulate fiducial distributions, Fisher attempted to use the probability integral transform as a generic pivotal quantity for continuous univariate distributions. It turns out, however, that more is needed for a fiducial distribution to be available, even in principle. The key step in 18.8 is $Pr(Z(T, \theta)|D) = Pr(Z(T, \theta)|D, \boldsymbol{X} = \boldsymbol{x})$, and this is where problems might arise. Sidenfeld (1992) gives an example of a random varible X with a uniform distribution on $[0, \theta]$ for some $\theta > 0$. Suppose that D consists of this model and some additional information that $\theta \leq \theta^*$ for some $\theta^* < \infty$. The distribution function of X is $F(x|\theta) = x/\theta$. Let $V = F(X, \theta)$ be our chosen pivot so that

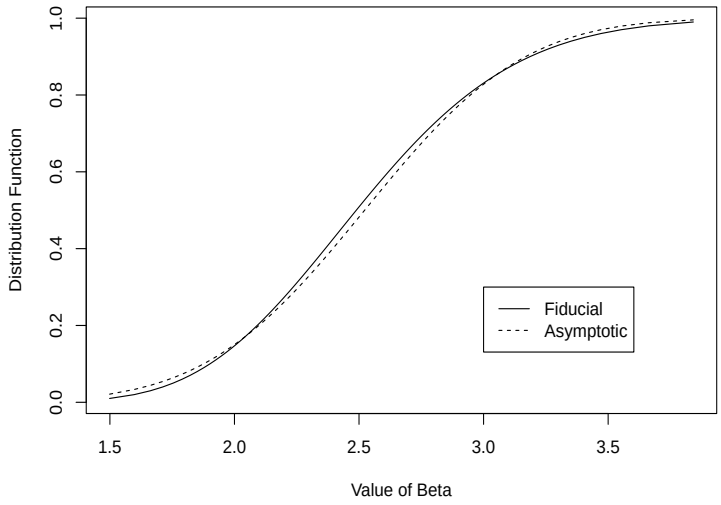

Figure 18.1: Fiducial distribution function for parameter of an exponential distribution based on a sample of size $n = 25$.

$p(v|D) = 1$; $0 < v < 1$ and then $Pr(V < 1/2|D) = 0.50$. The fiducial argument fails in this case because $p(v|D) \neq p(v|D, X = x)$. To see this, note that for a given $0 < x < \theta$, $x/\theta^* < v = x/\theta < 1$ because $\theta < \theta^*$. Thus, observation of $X = x$ is not irrelevant to the distribution of the pivotal quantity V. When the observation is irrelevant to the distribution of the pivotal quantity it is said that the *Principle of Irrelevance* holds. According to Edwards (1976) it was Ian Hacking who gave the Principle of Irrelevance its name with respect to the fiducial argument. And, it appears that the principle of irrelevance holds only for location-scale distributions (Edwards, 1976, p. 24). In particular, it does not hold for the primary example used to introduce the fiducial approach in Fisher (1930) which was the correlation coefficient in a bivariate normal model. And, as illustrated by the previous uniform example, it may

not hold if the background information available involves more than only the model for response random variables. The principle of irrelevance was later referred to by Fisher as the problem of recognizable subsets.

It was in extending the direct fiducial argument (18.8) to the multiple parameter case that Fisher encountered difficulties that seem to have ultimately proven insurmountable. Fisher set out to solve the Behrens-Fisher problem using fiducial probability, which is determination of an interval for $\mu_1 - \mu_2$ in a normal two sample problem with unknown variances σ_1^2 and σ_2^2. As recounted by Zabell (1992) Fisher's presentation and justification of the fiducial approach for this problem shifted in response to a number of criticisms by Bartlett (1936, 1937, 1939). During this time frame, Fisher offered a number of sometimes conflicting rationales for the fiducial argument, and an evolving set of conditions needed for it to be valid (Zabell, 1992). The confluence of the back-and-forth between Bartlett and Fisher seems to have been Fisher's arrival at the concept of recognizable subsets. For Fisher, probabilities required specification of a reference set such that the event of interest was an element of that reference set. These were not required to be physically existing populations, but could be the sets leading to the *hypothetical repeated sampling* used to give meaning to $p-$values in tests of significance. What we have called here the principle of irrelevance is equivalent to the assertion that there are no recognizable subsets of the reference set such that the probability of an event is different in different subsets.

It may have been the identification of relevant subsets in several problems that led to the evolution of Fisher's thinking about the fiducial argument. As stated, Fisher focused on the Behrens-Fisher problem, but we can illustrate the situation with the related but simpler problem of computing an interval

for μ in a one sample normal model when the variance σ^2 is unknown. There were two avenues for the application of the fiducial argument in this problem, one that can be called the *step-by-step* development and the other the *direct* development, both of which Fisher (9173) presented.

Fisher asserted that the direct development could be used to obtain a fiducial interval for μ by taking, in 18.8, $T = (\bar{Y}, S^2)$ and then

$$Z(T, \mu) = \frac{\bar{Y} - \mu}{(S^2/n)^{1/2}}.$$

The event E is $t_{\alpha/2;n-1} < Z(T, \mu) < t_{1-\alpha/2;n-1}$ and then application of the argument leads to the fiducial probability statement,

$$Pr(\bar{y} - t_{1-\alpha/2;n-1}\sqrt{s^2/n} < \mu < \bar{y} - t_{\alpha/2;n-1}\sqrt{s^2/n}) = 1 - \alpha.$$

What Fisher saw as necessary for this development to hold is sufficiency of T, the pivot $Z(T, \mu)$ and that "knowledge of μ and σ *a priori* is absent." Sufficiency and absence of prior knowledge were believed (at that time) to ensure that there existed no recognizable subsets (i.e., the principle of irrelevance was met). But Buehler and Feddersen (1963) showed that recognizable subsets did exist for this problem (Sidenfeld, 1992; Zabell, 1992).

The step-by-step development in this problem is to recognize that the elements of the the sufficient statistic $T = (\bar{Y}, S^2)$ are conditionally independent and that,

$$p(\bar{y}, s^2|\mu, \sigma^2) = p(\bar{y}|\mu, \sigma^2)\, p(s^2|\sigma^2).$$

Consider first, $p(s^2|\sigma^2)$. $T_2 = (n-1)S^2$ is sufficient for σ^2 and $Z_2(T_2, \sigma^2) = T_2/\sigma^2$ is pivotal having a Chi-squared distribution with $n-1$ degrees of freedom,

$$p(z_2) = \frac{(1/2)^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}\right)} z_2^{(n-1)/2-1} \exp\left(-\frac{1}{2}z_2\right).$$

Applying the fiducial argument to this piece results in the fiducial distribution for σ^2,

$$p(\sigma^2|s^2) = \frac{[(n-1)s^2/2]^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{(n-1)/2+1} \exp\left(-\frac{(n-1)s^2}{2}/\sigma^2\right), \tag{18.9}$$

which is an inverse gamma distribution with parameters $(n-1)/2$ and $[(n-1)s^2]/2$. As for $p(\bar{y}|\mu, \sigma^2)$, consider everything conditional on σ^2. Given σ^2, $T_1 = \bar{Y}$ is sufficient, $Z_1(T_1, \mu) = (T_1 - \mu)/\sqrt{\sigma^2/n}$ is pivotal, and

$$p(z_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_1^2\right).$$

Applying the fiducial argument to this piece results in the conditional fiducial distribution for μ given σ^2,

$$p(\mu|\bar{y}, \sigma^2) = \frac{1}{(2\pi\sigma^2/n)^{1/2}} \exp\left[-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right]. \tag{18.10}$$

With (18.9) and (18.10), we may obtain the marginal fiducial distribution of μ as,

$$p(\mu|\bar{y}) = \int p(\mu|\bar{y}, \sigma^2)\, p(\sigma^2|s^2)\, d\sigma^2. \tag{18.11}$$

Following the derivation used in Gelman, Carlin, Stern, and Rubin (1995, p. 68-69) to obtain the marginal posterior of μ in a Bayesian analysis with an improper prior (more on this in the sequel), the integral in (18.11) turns out to be a non-central $t-$distribution with $n-1$ degrees of freedom,

$$p(\mu|\bar{y}) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)(2\pi\sigma^2)^{1/2}} \left[1 + \frac{1}{n-1}\left(\frac{(\mu - \bar{y})}{\sqrt{s^2/n}}\right)^2\right]^{-(n/2)}. \tag{18.12}$$

The expected value of this fiducial distribution for μ is $E(\mu) = \bar{y}$ and the variance is $\text{var}(\mu) = s^2/n$.

Note that the fiducial distribution of $(\bar{y} - \mu)/\sqrt{\sigma^2/n}$ is again a central $t-$distribution with $n-1$ degrees of freedom, and and interval for μ reduces to,

$$Pr(\bar{y} - t_{1-\alpha/2;n-1}\sqrt{s^2/n} < \mu < \bar{y} - t_{\alpha/2;n-1}\sqrt{s^2/n}) = 1 - \alpha,$$

which is the same as found using Fisher's *direct* development. But, as already described, the direct development is flawed in that there do exist recognizable subsets or, equivalently, the principle of irrelevance is violated. One might take this to mean that the step-by-step development is then equally flawed, but Sidenfeld (1992) argues that is not necessarily the case, using a measure-theoretic argument that is more involved than we wish to get into here. The crux of the matter, however, rests on the fact that there is an indisputably valid analysis based on a Bayesian formulation of the problem that is equivalent, not just in end product, but in virtually its entire development, to the step-by-step fiducial procedure, but not the direct procedure. And this Bayesian formulation of the problem can be arranged to put probability 0 on the types of recognizable subsets that Buehler and Feddersen (1963) identified. Specifically, if one considers the model $Y_1, \ldots, Y_n \sim$ iid $\mathrm{N}(\mu, \sigma^2)$ and assigns the joint prior $\pi(\mu, \sigma^2) = 1/\sigma^2$, the marginal posterior of σ^2 is the same as (18.9) and the conditional posterior of μ given σ^2 is the same as (18.10). The marginal posterior of μ is then the same as (18.12), the fiducial distribution for μ given $\bar{y}$. Thus, the step-by-step procedure is validated, while the direct procedure advocated by Fisher and repeated by others, including Edwards is not.

In closing, although the fiducial argument is not taken seriously as a method of practical importance and has been called "Fisher's one great failure" (Zabell, 1992) it is instructive to consider what Fisher was trying to

achieve and the impact arguments about the fiducial approach had on the development of interval estimation as an topic in statistical inference.

18.5 Bayesian Intervals

As with the topic of Bayesian inference in general, there are two broad approaches to how credible intervals can be given meaning. One, which we call the pragmatic approach, is based on intervals as summaries of posterior distributions in the same way that means, variances, and quantiles are summaries of posterior distributions. The other attempts to embed the development of credible intervals in a formal decision theoretic framework, hoping thereby to provide greater mathematical justification for the development of intervals.

18.5.1 The Pragmatic Approach

This approach is based on the simple assertion that posterior distributions encapsulate everything we can know or believe about the possible values of a parameter. In terms of inferential content, we then need nothing more than the recognition that an interval containing a given proportion of the posterior distribution represents a set to which we believe the parameter belongs with the given proportion of our entire belief set. Rules for the construction of posterior intervals are then just conventions designed to improve communication of the summary represented by an interval. In particular, we generally agree that,

1. Posterior intervals should be contiguous rather than consisting of disjoint pieces.

2. Posterior intervals should be relatable to other summaries of posterior distributions such as quantiles.

3. There are no *optimality* criteria that can be used to distinguish different possible intervals.

There are, of course, potential scenarios that might cause difficulties with these principles, such as bimodal posteriors, but many of these also reflect a deficiency of the model to represent the data generating mechanism.

The basic principles enumerated above provide guidelines, not mathematical criteria, for the construction of posterior intervals. The concept of highest posterior density intervals, as discussed in Intermediate Statistical Methods, is perfectly acceptable under this approach, in particular for posteriors that are moderately skewed. At the same time, some practitioners (e.g. Gelman et al., 1995) prefer central intervals consisting of the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution. One reason for this preference is that quantiles are invariant to parameter transformations, while the endpoints of highest posterior intervals are not.

We have mentioned, in the historical development of confidence and fiducial intervals, that prior to the 1930s, intervals were largely the domain of inverse probability (what we now call Bayesian analysis). It appears as if those early intervals were interpreted along the lines of this approach to Bayesian interval development. Jaynes (1976) reports that one of the (if not the) earliest use of interval estimation was conducted by Laplace in an analysis of the mass of Saturn at the end of the 18^{th} century. He apparently applied Bayes theorem with a uniform prior density on mass (as calculated from observations of mutual perturbations of Jupiter and Saturn and the motion of their moons) to arrive at a posterior distribution for the mass of

Saturn, M. Laplace then reported that "it is a bet of $11,000$ against 1 that the error of this result is not $1/100$ of its value" Jaynes (as quoted in 1976, p. 180) which we take to mean that, if $\tilde{M}$ was the posterior mean, median, or mode, then $Pr(\tilde{M} - \tilde{M}/100 < M < \tilde{M} + \tilde{M}/100) = 0.9999$.

18.5.2 The Decision Theoretic Approach

It is not clear to us when the problem of interval estimation from a Bayesian viewpoint was first cast in the context of decision theory, but it does not seem to be universally accepted that doing so is particularly valuable. Berger (1985, p. 145), commenting on the lack of invariance of highest posterior density intervals asserts

> . . . we do not view credible sets as having a clear decision-theoretic role and are therefore leery of "optimality" approaches to selection of a credible set. We mainly view credible sets as an easily reportable crude summary of the posterior distribution, and, for this purpose, are not unduly troubled by, say, nonuniqueness of HPD credible sets.

Similarly, Robert (2007, p. 266) comments that

> We do not pursue any further the decision-theoretic study of Bayesian confidence regions. Indeed, an important aspect usually overlooked in the derivation of confidence regions deals with how they will be used, although this very use is essential in the construction of the loss function.

Just as Neyman wrestled with the two fundamental properties of confidence intervals coverage, and width, so should reasonable loss functions

consider these two aspects in determining rule for choosing a set C for which the inference is to be $\boldsymbol{\theta} \in C \subseteq \Theta$. One form of such a loss function is

$$L(C, \theta) = a \text{ vol}(C) - I(\boldsymbol{\theta} \in C), \quad a > 0, \tag{18.13}$$

where I is the usual indicator function $I(A)$ that assumes a value of 1 if A is true and a value of 0 otherwise. The factor a is intended to allow differential weighting of loss due to volume and loss due to coverage. The loss function (18.13) is given by Casella, Hwang, and Robert (1993). An equivalent form is given by Robert (2007) as,

$$L(C, \theta) = \text{ vol}(C) + cI(\boldsymbol{\theta} \notin C). \tag{18.14}$$

Aside from the slight difference in where the weight factors a and c appear, these two loss functions differ only by an additive constant of 1. Using the form (18.13), the expected loss becomes,

$$R(C, \theta) = E[\text{vol}(C)] - Pr(\boldsymbol{\theta} \in C). \tag{18.15}$$

Casella et al. (1993) demonstrated a most unpleasant characteristic of (18.13) that they attribute originally to J.O. Berger. We have already described that the classical Neyman confidence interval for a normal mean and the Bayesian highest posterior density interval for the same with Jeffreys prior (and the fiducial interval to boot) are all the same, namely,

$$C_t = \left\{ \bar{y} - t\sqrt{s^2/n} \leq \mu \leq \bar{y} + t\sqrt{s^2/n} \right\}, \tag{18.16}$$

where t is some quantile of the $t-$distribution with $n-1$ degrees of freedom. Using (18.13) the expected loss for this set is,

$$R(C_t, \theta) = a\left(2t\sqrt{\frac{s^2}{n}}\right) - Pr(\mu \in C_t). \tag{18.17}$$

The difficulty is that there is an alternative set estimator with posterior expected loss less than or equal to that of C_t for any data set,

$$C_s = \begin{cases} C_t & \text{if } s \leq \frac{\sqrt{n}}{2ta} \\ Q & \text{if } s > \frac{\sqrt{n}}{2ta}. \end{cases} \tag{18.18}$$

In (18.18), Q can be any real number such as 21, $\bar{y}$ or even $\emptyset$. Then for small s^2, the expected loss $R(C_s, \theta) = R(C_t, \theta)$, but for larger s^2, $R(C_s, \theta) = 0 < R(C_t, \theta)$. As Casella et al. (1993) point out, C_s is a nonsensical estimator that indicates no uncertainty in the value of μ as the variance in the data grows larger. The problem lies in the linear form of the loss function (18.13) which gives too much weight to the volume of C and insufficient weight to coverage. The weighting factor a is largely irrelevant in remedying this deficiency because the loss due to coverage must lie in the interval $(0, 1)$ while that for volume is unbounded.

Note that the interval (18.16) is not the Bayes rule that corresponds to minimization of (18.17). It is, however, a highest posterior density interval. The point of the demonstration just concluded is that under the linear loss function (18.13) the solution that is accepted by frequentists as optimal in the sense of Neyman, which is also a highest posterior density credible interval in a Bayesian formulation of the same problem, that solution can be dominated (in terms of expected loss) by a nonsensical alternative. The problem lies not with the interval, but with the loss function.

Casella et al. (1993) suggested that the problem could be overcome by using loss functions

$$L_H(C, \theta) = H[\text{vol } (C)] - I(\boldsymbol{\theta} \in C), \tag{18.19}$$

where H is a continuous monotone increasing function. Loss functions of the form (18.20) include the problematic linear loss (18.13), but conditions

placed on H can eliminate the type of problematic behavior seen with linear loss. Robert (2007, p. 266) indicates that one of the simplest of this new class of loss functions is, for some $k > 0$,

$$L_R(C, \boldsymbol{\theta}) = \frac{\text{vol}(C)}{\text{vol}(C) + k} + I(\boldsymbol{\theta} \notin C). \tag{18.20}$$

Both terms in 18.20 are bounded by 1. Robert indicates that the Bayes estimators associated with such loss functions are still highest posterior density intervals, but now do not exhibit the behavior of collapsing to a point as the data variance increases. There does remain the need to choose a suitable function H, which might depend on the problem at hand, and there appears to be little in the way of generally applicable guidance for how to construct an appropriate loss of the form (18.19).

There can be some additional difficulties with developing set estimators using decision theory, and some of these are discussed at some length in Joshi (1969) and Casella, Hwant, and Robert (1990). These difficulties are typically quite technical and involve several alternative versions of admissibility. We do not believe a thorough discussion of the issues involved is necessary to grasp the fundamental nature of attempts to develop Bayesian intervals using a decision-theoretic basis. The material presented in this chapter should indicate that such development has not been a straightforward exercise in applying the fundamental principles of decision theory to the derivation of rules for constructing credible intervals and sets. Thus, it is easy to see how the attitudes expressed in the quotes from Berger and Robert at the beginning of this section were formed.

Chapter 19

Epilogue

> Philosophy is consequential because it guides action.
>
> Kass (2006, p. 437)

In the fourth part of this book we have presented material connected with actions that are encompassed by statistical analysis such as testing, interval estimation, and identifying appropriate interpretations of results. We have done so from the viewpoint of applied statisticians, not philosophers, and we are certain true philosophers will find much to criticize, perhaps even ridicule, in our presentation. Our defense is that we are concerned, not with tying together a coherent philosophy of statistical practice, but in trying to make sense of what is a jumble of disparate methodological approaches and their underlying philosophies. We find ourselves willing to make use of different approaches with different problems, rejecting the notion that there is one correct philosophical viewpoint for all of statistics. What we believe is very important, however, is to be clear about what the underlying basis of an analysis or method is, and how that basis translates into the

interpretation of the results of applying the method. It is our opinion, for example, that those who compute credible intervals through the application of MCMC methods but then give those intervals the same interpretation as classical frequentist confidence intervals do damage to the theories and applications of both approaches to the construction of statistical intervals. In this spirit, we give here some concluding remarks about what we take away from our brief journey through some of the foundational issues in statistics. We do so in reverse order of how the topics were presented in the previous chapters.

19.1 Construction of Intervals

There are more approaches to the construction of statistical intervals than one might initially think. Classical Neyman-style confidence intervals and their asymptotic extensions, and Bayesian credible intervals are the major players in the field, but the use of intervals from inversion of likelihood ratio tests and normed profile likeihoods are also common. Support intervals are not common, but perhaps should see greater use, particularly given that 90% and 95% confidence intervals are just a hangover from what we were taught about tests.

Of the choices, intervals based on support and Bayesian credible intervals are the only two that can lay claim to post-data precision or a pure evidential interpretation. We approach the justification of Bayesian intervals from the viewpoint of pragmatic Bayes and are willing to live with either highest posterior density or central intervals as convenient conventions without formal justification through decision theory or other attempts to define optimal properties regarding length or volume.

Intervals based on likelihood ratios and their sampling distributions are standard fare in non-Bayesian analyses. The availability of normed profile likelihoods broadens the applicability of this class of intervals far beyond problems with only scalar parameters. Intervals based on likelihood sampling distributions do have the nice interpretation that values farther away from the maximum likelihood estimate receive less support from the given set of observed data than do values nearer the maximum likelihood estimate. Aside from this property, however, likelihood intervals based on sampling distributions are subject to the same litany of misinterpretations as classical confidence intervals.

Confidence intervals constructed from the theory of Neyman are perhaps the most widely used inferential procedure in statistics, at least if one includes in this category what we have sometimes called Wald theory approximate intervals, those based on asymptotic normality of maximum likelihood estimates in regular problems. Confidence intervals of this type are subject to a host of misinterpretations including that the parameter lies inside a given interval with the specified probability, that values inside the interval are more *plausible* as values of the true parameter than those outside the interval, and that values nearer the center of an interval are closer to the true parameter value than those farther away from the center. A subtle misinterpretation that even seasoned statisticians can be drawn into, is the belief that a shorter interval reflects a more precise estimate of the parameter than does a wider interval. In short, conclusory statements made about the possible values of the parameter or about individual intervals (after they are computed) are incorrect interpretations. Only interpretations relative to performance of the procedure are valid. That is, classical confidence intervals are concerned only with pre-data precision. The most consistent way to use confidence intervals

is per the advice of Neyman that "we may decide to behave as if we actually knew that the true value [is in the interval]" (Neyman, 1941, p. 134). This simply treats a confidence interval as a family of hypothesis tests in which every value inside the interval would be accepted as a null hypothesis at the given level of confidence (or Type I error rate).

19.2 Use of Tests

Tests may appear in many different aspects of a statistical analysis. If a scientific hypothesis is directly embodied as a value of some parameter, then a test of the hypothesis that the parameter is equal to that value constitutes a test of the scientific hypothesis. If we wish to determine whether the distributional location of two or more groups differ we may have the classic hypothesis of equal means. If we wish to draw a conclusion that the distributions of a set of response variables is influenced by a covariate we may test the hypothesis that a slope parameter is equal to zero or that an expectation function in total is constant across covariate values. If we wish to determine whether the distributions of response variables in two groups differ we may test a model in which all of the parameters of a common form of distribution are equal between the groups against a model in which at least some of the parameters differ. If we wish to determine whether a posited link function in a generalized linear model appears reasonable we may test the hypothesis that the parameter in a family of link functions has a specified value that reduces it to the desired link function. If we wish to determine whether there is structure in a set of data that can be represented by temporal or spatial dependence we may test that an autoregressive coefficient or a spatial dependence parameter in the model is equal to zero. If we wish to determine

whether some model appears to be a reasonable representation of a distribution appropriate to describe the available data, we may test the hypothesis that the empirical distribution of a set of generalized residuals is consistent with a random sample from the uniform distribution on the unit interval.

To deal with this array of specific settings in which tests may be useful, we break possibilities into situations that involve testing that parameters have specific values, testing alternative models, and tests associated with model assessment. There may be considerable overlap among these situations, particularly the first two, but the division will be useful to help organize thinking. Throughout we will deal with a vector-valued parameter $\boldsymbol{\theta} \in \Theta$, although a hypothesis may involve only a portion of such a parameter.

19.2.1 Testing Parameter Values

In this category we put hypotheses of the form $H : \boldsymbol{\theta} \in \Theta_0 \subset \Theta$,or $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \notin \Theta_0$. Typically, Θ_0 will result in either a simple hypothesis, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ or, in the case of a scalar parameter a one sided hypothesis $H : \theta \leq \theta_0$. If the hypotheses are formulated as a pair of null (H_0) and alternative (H_1) hypotheses, then this pair is taken to form a logical disjunction so that $H_0 \cup H_1 = \Theta$.

Non-Bayesian tests associated with hypotheses in this category may very well take the form of Fisherian significance tests or Neyman-Pearson tests of hypotheses. We consider these to be distinct procedures and reject attempts to meld them into a "hybrid" approach, as we believe doing so damages not only the clarity with which we draw conclusions from tests, but also our ability to communicate how they should be interpreted. In general, we adhere to the advice of Chapter 17.4.2 in deciding which of these approaches to adopt.

For tests of a scientific nature that are undertaken in a non-Bayesian framework, we find tests of significance a useful tool. For tests leading directly to making a decision between two alternatives and that will be repeated in a similar manner many times, we find the demonstrably optimal properties of Neyman-Pearson hypothesis tests in controlling error rates compelling. That is, if a problem calls for a procedure designed to have pre-data precision, that is exactly what tests for acceptance sampling were designed to achieve. There is no reason one of these approaches must prove superior for all testing problems, since not all testing problems involve the same contextual situation.

In terms of the use of tests of significance in scientific contexts, we find sufficient reason to consider only a single hypothesis (see Chapter 17.7) and are willing to allow that a $p-$value conveys some measure of evidence against such a hypothesis. A $p-$value then entails a certain amount of post-data precision as long as guidelines for significance are not taken as sharp divisions. Even so, we do not find $p-$values to be a completely satisfying solution to the search for a quantification of epistemic content in a test procedure. The probability interpretation of $p-$values as the probability under the hypothesis of obtaining a result as or more extreme than that obtained under hypothetical repeated sampling does not provide a direct avenue for status of the $p-$value as a measure of evidence connected to a particular data set. We do not, however, find the reliance of $p-$values on tail areas particularly troubling, in much the same way that we are not overly bothered by the use of central credible intervals in a Bayesian analysis. Overall, for tests of a scientific nature that are undertaken in a non-Bayesian framework, we find tests of significance a useful, if less than perfect, tool.

A question that arises if one adopts the Neyman-Pearson approach of acceptance sampling is how similar repeated applications of the testing sit-

uation need to be before one prefers this approach to an alternative. Fisher argued that the answer was repeated sampling from populations that are the same in all relevant aspects (Chapter 17.9.2), although exactly what this means was not made concrete. Neyman argued at one point that the situations did not need to be similar at all, only the applicability of hypothesis tests. Thus, if one is faced with a large number of decision problems, control over the Type I error rate is desirable even if those decision problems are a set of unrelated decisions. There is some force to this argument, the best example we can think of being the process by which the Food and Drug Administration (FDA) decides whether to grant licenses for the production and sale of new drugs. In this context, there are repeated problems that involve a dichotomous decision and there may be some desire to control error rates in the decisions as a set, even though the situations comprised by the set may be quite different. The difficulty with this argument relative to scientific research is that scientists are not in the business of making decisions about different research questions. Rather than dichotomous decisions, it is evidence that is central to determining the next course of action, which may involve quite different avenues of investigation depending on how much evidence can be produced about a given hypothesis.

Procedures derived from using likelihoods, either as support functions or relative to the sampling distributions of likelihood ratios, seem more well suited to problems framed as model comparison rather than hypotheses about specific parameter values, and we prefer to consider them in that context. Although there are certainly numerous hypotheses that can be represented as either statements about particular parameter values or as the comparison of models, the former would seem to have applicability primarily in situations that involve variation-independent parameters of low dimension. We note

that the use of likelihoods also serves to give a general solution to the problem of what are sometimes unfortunately called "nuisance parameters" in testing situations, without worrying about the identification of pivotal quantities.

In terms of Bayesian tests our procedure of choice is to report the posterior probability $Pr(H|\boldsymbol{y})$ or, if there are two hypotheses, $Pr(H_0|\boldsymbol{y})$ and $Pr(H_1|\boldsymbol{y})$. As stated in a previous chapter, we are a bit cautious about using Bayes factors unless they correspond to the posterior odds of two models and, even then, we believe there is more information to be gained by reporting the two probabilities individually rather than as only a ratio. With this approach we rarely see the need to even frame the problem as a test, and again tend to simply report relevant probabilities. For example, if θ_1 and θ_2 represent the mean decreases in blood pressure under two treatments, rather than frame a comparison as $H_0 : \theta_1 \leq \theta_2$ versus $H_1 : \theta_1 > \theta_2$, it seems more meaningful to report $Pr(\theta_1 \leq \theta_2)$. Similar scenarios apply to at least the vast majority of inferences one might cast in terms of hypotheses to be tested.

19.2.2 Testing Alternative Models

In testing alternative models in a non-Bayesian framework, the typical test statistics used in group comparison are often not available to us. In some cases they are, such as a test that the slope parameter equals zero in a simple linear regression with normal errors. One could equally well consider this to be a test of particular parameter values or a test of alternative models. Similarly, what is often called a *general linear test* is formulated in terms of Reduced and Full models and, again, exact theory results are available for identification of the sampling distribution of the test statistic; this test statistic, incidentally, has the form of a slightly modified likelihood ratio

test for independent normal response variables with linear systematic model component and constant variance. But often, when we consider a problem to be one of testing alternative models, we are talking about models that do not fall neatly into the category of linear normal models with equal variances, or the few other cases of exponential family models for which exact tests can be developed from the Neyman-Pearson lemma. By far and away the most common type of (non-Bayesian) test conducted to compare two models is the likelihood ratio test. Although the test statistics may differ from those often used in testing hypotheses about particular parameter values, the underlying approaches of tests of significance and tests for acceptance sampling still apply. If the reference distribution for a test is derived from the sampling distribution of the test statistic under a reduced model, then we can interpret a $p-$value as a primitive measure of evidence against that model, or we can adhere to sharp boundaries for reaching a decision to accept one model or the other. As already reiterated numerous times in the material of this book, either one or the other of these approaches can be preferred, depending on the context of the testing problem. That said, it may be a bit less common that repeated comparison of non-normal or non-linear models or models with non-constant variance arise in practice than in simpler hypotheses involving a single parameter. As with intervals, we find the idea of testing (comparing) models based on the use of likelihood as support to be attractive, and believe that the use of this approach to a greater degree than it is currently seen might simplify and improve the communication of statistical inference.

In terms of Bayesian model comparison, there is really little to add to what we have already put forth. Bayes factors may well have some truly legitimate role to play in this regard, but the potential pitfalls involved argue for only very careful application. It is perhaps only when computing

the posterior probabilities of models $Pr(M_1|\boldsymbol{y})$ and $Pr(M_2|\boldsymbol{y})$ proves computationally prohibitive that reliance on a Bayes factor is clearly justified. We also find that the careful assessment of models, through the use of posterior predictive checks on model behavior can be a more informative way to compare models than just a one number representation of posterior odds ratio (Bayes factor) or even the actual posterior odds. This is because the intelligent selection of test quantities in a posterior predictive assessment can provide more detailed information about the way in which a model might fail to provide a good representation of structure in the data.

19.2.3 Testing Goodness of Fit

In Chapter 17.10.7 we went into some detail regarding the position of Karl Pearson on interpretation of $p-$vlaues associated with his Chi-squared GOF test and, by extension, other GOF tests based on measures of discrepancy between what is expected under a hypothesized model and what is observed in the data. Why are the opinions of K. Pearson regarding the interpretation of his GOF test relevant for us? First, because it is clear that he never intended the GOF test to be assessed in terms of $p-$values following either of the routes offered by Fisher's tests of significance or Neyman-Pearson tests for acceptance sampling. Perhaps, then, neither should we. Secondly, K. Pearson's interpretation of the results of a GOF test ask us to reconsider our own views about how such tests fit into the overall framework of testing statistical hypotheses. Are either of the Fisherian or Neyman-Pearson approaches to interpretation of $p-$values truly appropriate for GOF tests? If not, then how should the results of GOF tests be interpreted?

We see two avenues for the interpretation of $p-$values calculated from

GOF tests. One is the Fisherian interpretation of a $p-$value as a measure of evidence against the hypothesis, which here is a theoretical probability distribution. Under this interpretation we have the usual conceptualization of looking at the distribution of a test statistic under the assumption that the hypothesized distribution is correct, and determining whether the actual test statistic has a value that would be unusual under that distribution, or not. Unusual generally means the use of tail areas which, as we have discussed previously, is not without some controversy. The controversy over tail areas aside, this conceptualization basically considers the observed data to be a realization from the posited model, and hence Fisher's explanation that a $p-$value represents the probability of obtaining a result as extreme or more extreme than that actually observed if the hypothesis is true. If we choose this route for interpretation of GOF tests we are left with the conundrum of whether a $p-$value of 0.07, say, offers sufficiently little evidence against the hypothesized distribution that we fail to reject it, or whether it offers sufficiently strong evidence against the hypothesized distribution to declare it untenable as a description of the observed data. Perhaps an even more serious disconnect between GOF tests and the Fisherian significance testing framework is that in the Fisherian framework, a hypothesis that we fail to reject is simply ignored whereas, in a GOF test, we wish to declare positive support for a hypothesized model. The typical approach among applied statisticians is to simply ignore our conundrum, since we have no good solution to it, and hope for an unambiguous $p-$value that is either small (e.g., less than 0.05) or decidedly not small (e.g., $p > 0.15$).

The other possible framework under which to interpret $p-$values in GOF tests is essentially that of Pearson. Consider, for simplicity, a one sample problem in which our hypothesized model is that $Y_1, \ldots, Y_n$ are indepen-

dent and identically distributed random variables with distribution function $G(y|\boldsymbol{\theta})$. We have data, $y_1, \ldots, y_n$, which may or may not correspond to the hypothesized model, including the stipulations of independence and identical distribution as well as the distributional form. Now, if the observed data are a random sample from $G(y|\theta)$, then the empirical distribution $F_n(x)$ converges to $G(x|\theta)$. But even if this is not true, and the data are not a random sample from $G(y|\theta)$, it is possible that $G(x|\theta)$ provides a good description of the frequency distribution $F_n(x)$. Our objective, then, is to determine whether the hypothesized model $G(y|\theta)$ provides a good representation of the data, regardless of where the data actually came from. The important point is that we are concerned with testing the degree to which a theoretical model describes the frequency distribution of our data, not whether the data actually arose from the theoretical model. Our test statistic is in the form of a discrepancy measure, such as the χ^2 statistic, deviance, or a Kolmogorov statistic. These quantities all have the propertie s of being non-negative and monotone increasing as the degree of discrepancy gets larger. The $p-$value for a perfect fit (model frequencies equal to data frequencies) is $p = 1.0$. There is no concern for tail areas as representing regions of low probability, only the fact that as the test statistic increases from zero, the $p-$value decreases from 1.0. The value of p is that it functions on the same scale for all suitable problems, while the absolute values of test statistics can be quite different among problems. There is no definitive threshold, "The value of P at which we consider goodness or badness of graduation starts cannot be fixed without regard to the special problem under consideration" (Pearson, 1935). But given a monotone relation between "goodness of graduation" and $p-$values, we can declare relative differences among competing models. This is, although a bit subtle, actually quite a departure from the usual interpre-

tation of $p-$values in tests of significance and not even in the same universe with the interpretation in acceptance sampling in which different $p-$values carry no meaning except as to which side of the critical value they are on.

Simulation-based GOF procedures, such as those discussed in Chapter 11 or those based on posterior predictive $p-$values in a Bayesian analysis can be interpreted by considering $p-$values in much the same manner as for tests of significance. Although this does not solve the problem of whether or not to declare an adequate fit based on a $p-$value of small but not extremely small magnitude (e.g., $0.05 < p < 0.15$), simulation based methods lend themselves to multiple assessments using different characteristics of data structure. For example, in a regression problem with multiple observations at each covariate value, we might construct one test statistic as the average difference between the maximum data value and the third quartile at each level of the covariate to reflect extreme tail behavior. We might then construct another test statistic as the slope in a Box-Cox plot to reflect mean-variance relation, and yet another test statistic as the value of a Kolmogorov statistic to compare generalized residuals to a uniform distribution to reflect distributional form. This would then result in three simulation-based $p-$values of the type discussed in Chapter 11. Far from being the hindrance we are taught about multiple testing (from the viewpoint of Neyman-Pearson theory) having three $p-$values allows us to draw more detailed conclusions about which aspects of data behavior are well represented by the model and about which aspects of data behavior might be not so well represented by the model. It is normal, not unusual, that for any reasonably complex situation, models will describe different parts of data behavior with differential quality. Thus, we may have several $p-$values that vary in magnitude because they are measuring different characteristics of the model. The same is true of tests developed from

posterior predictive distributions in a Bayeisn analysis. In fact, we find the force of this concept so compelling that we are ready to recommend that statisticians abandon the search for one omnibus GOF procedure that will declare a model *adequate* or *inadequate* as a whole, except in the simplest of problems.

19.3 Overall Inferential Frameworks

As we have indicated several times, we do not believe there exists a single approach to statistical analysis that should be followed in every problem. This does not mean, however, that we find any approach equally appropriate regardless of the problem. And, there are methodological areas we have not touched upon, such as functional data analysis, random forests, and nonparametric smoothers, just to name a few. Whether there are definitive philosophical underpinnings for these less traditional methods that can be compared and contrasted with those we have presented here can be considered an open question. It has been our position throughout the book that if the only objective of an analysis is pure data description or prediction, then one is free of concerns connected with the representation of scientific mechanisms by stochastic models. Thus, while not dismissing procedures that focus solely on data description or prediction as lacking value, we do find that their objectives lie somewhat outside of the sphere of statistical inference as it relates to understanding and explanation. In this section, then, we briefly discuss the types of considerations that are relevant in determining the inferential framework one will apply to a given problem.

19.3.1 Working with a Team

It should go without saying that one cannot force a particular philosophical approach onto a problem without agreement from other individuals involved. One can make a case for whatever approach one finds the most applicable, but the viewpoints of others must be accommodated to one degree or another so that a consensus position is possible. For example, in a consulting situation with an applied scientist, if that scientist is, for whatever reason, vehemently anti-Bayesian, it is unlikely that one will be able to convince them that a Bayesian analysis would be a good approach. One must then either excuse oneself from the project or change course and make use of an alternative analysis that, while perhaps not your first choice, can provide ways to achieve the objectives of analysis. If part of a multi-disciplinary research team, such as might be formed in response to a funding announcement or as part of a collaborative effort between government, industry and academia, it is generally helpful to push for the identification of what the objectives of a data analysis are going to be. This then leads naturally to the use of various potential methods and their associated philosophical bases.

It is true that the philosophy of science a group wishes to guide the planning, execution, and analysis of a project is rarely, if ever, a topic of discussion in research group meetings. But if one is consciously aware of the impact that philosophical considerations have on how scientists conduct studies, one will see the importance of those considerations in the suggestions made regarding, for example, study design. Similarly, one does not need to launch into a discussion of the finer points of concepts of probability and procedures designed to achieve pre-data and post-data precision to provide information on the types of conclusions that an analysis will support.

19.3.2 Don't Forget Randomization Procedures

In Chapter 16.2 we outlined direct connections between a number of concepts of probability and statistical methods associated with them. We have said little else about Laplacian probability and randomization/permutation tests, or (finite) relative frequency and survey sampling methods. But when the associated concept of probability is applicable to a problem these methods form very philosophically sound approaches to analysis. And they are not simply approximations to other methods, they are distinct. To illustrate this, consider a problem consisting of two groups of experimental units (e.g., pigs, lab rats) to which two treatments are to be given and a well-defined response variable measured or observed. Suppose there are 10 experimental units total, selected to be physically independent and as homogeneous as possible in terms of characteristics that might be related to the responses of interest. Let $U_1, \ldots, U_{10}$ denote the ten experimental units, and suppose that the units are randomly assigned to the treatment groups $T_1 : U_1, U_5, U_7, U_8, U_{10}$ and $T_2 : U_2, U_3, U_4, U_6, U_9$. Suppose further that the observed responses were $U_1 : 6.22$, $U_2 : 6.08$, $U_3 : 5.27$, $U_4 : 5.13$, $U_5 : 0.24$, $U_6 : 5.03$, $U_7 : 2.93$, $U_8 : 2.36$, $U_9 : 5.31$, and $U_{10} : 2.66$. If we want a non-Bayesian test for a location shift between the groups T_1 and T_2, we could conduct a randomization test or perhaps a $t-$test. For the randomization test, there are 252 possible assignments of experimental units to treatment groups and the randomization $p-$value is $p = 12/252 = 0.0476$. If we assume normal distributions and equal variances for the two groups, a $t-$test results in $p = 0.0599$. If the problem is such that we need to decide whether to behave as if the group means are equal or not we might interpret our $p-$values following Neyman-Pearson theory with $\alpha = 0.05$, in which case the randomization test would accept

the hypothesis of equal means while the $t-$test would accept the hypothesis of unequal means. If the problem is such that we wish to assess evidence provided by the data for a hypothesis of equal means we would interpret our $p-$values following the tradition of Fisher and most likely conclude that the two procedures provide essentially the same amount of evidence against the hypothesis. Given the small size of the study, we would likely conclude that there is sufficient evidence to reject the hypothesis of equal means. A graph of the two reference distributions is shown in Figure 19.1. In this example,

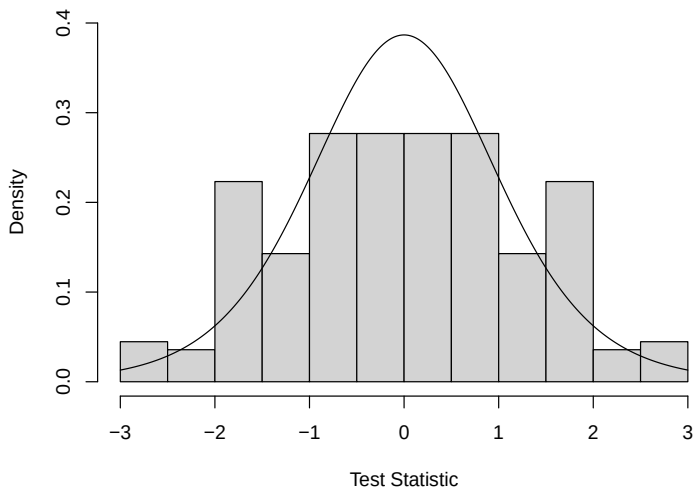

Figure 19.1: Reference distributions for testing hypotheses of equal mean among two groups. Histogram is for randomization test and curve is for traditional $t-$test.

the data used were simulated from two normal distributions with means 3.0 and 4.9 and equal variances of 4.0. Thus, as long as we assume in a physical version of the problem that experimental units were randomly assigned

to treatment groups, we have two procedures that are both completely justified by the circumstances. If one were truly in an acceptance sampling situation, the difference between randomization and parametric $t-$test procedures would be a bit disconcerting since these are not two in a set of similar problems, but rather two wholly justified procedures applied to the same set of data.

19.3.3 Prior Information Should be Used

If, in a given problem, we are privileged to have access to pertinent information about the problem arising from sources other than the current data themselves we should certainly incorporate that information into our analysis. Few statisticians would argue that an analysis should make use of less information than contained in a complete data set, and we see no reason this concept should not apply to information from other sources as well. The question is how relevant auxiliary information should be incorporated into an analysis. The obvious answer, if approaching a problem from a Bayesian viewpoint, is by allowing that auxiliary information to influence the selection of prior distributions. While examination of current data to help choose a prior for its analysis will lead to underestimation of uncertainty in posterior distributions, examination of data from previous or related studies does not suffer the same problem. Data from previous or related studies are, in fact, something of a gold standard in the determination of what one believes prior to seeing data from the current study. As we have commented previously, this situation, which was demanded by Karl Pearson for the use of inverse probability, is what should really be called objective Bayes, although that is not true in our current statistical parlance.

While Bayesian analyses have been criticized, and sometimes rightly so, for rather fanciful construction of prior distributions, frequentist analyses have somehow largely escaped criticism for failing to incorporate prior information in a meaningful way. One presumption seems to be that relevant prior information has been incorporated into study design, measurement technology, and perhaps formulation of response distributions. But all of these modeling concerns can be true for Bayesian analyses as well. If there is available information about the values of parameters or relations among parameters in a model, that information should logically influence the development of likelihoods. But this can be a difficult mathematical exercise if one is unwilling to make use of a prior distribution, and many frequentist analyses will not attempt the incorporation of such information.

It is not the norm for a problem to have true prior information in the form of data from previous or related studies. It might be only slightly more common for scientific opinion to offer a consensus relative to the components of a statistical model. But when there does exist true auxiliary information relevant to a given problem, it seems incumbent upon a statistician to make use of that information, just as it is expected that he or she will make use of all of the values in an observed data set. We have found that actively searching for such auxiliary information can yield positive results. An example is the information found on the websites of natural resource agencies regarding the relation of weight to length in Walleye in Chapter 9.10 of *Intermediate Statistical Methods*. That information was useful in motivating the systematic component of the regression model as well as choosing priors for the data model parameters.

19.3.4 Ignorance is Not Prior Information

In a Bayesian analysis with an unbounded parameter space, an improper prior distribution can be thought of as reflecting a complete lack of knowledge about what values that parameter might assume. If we have a one parameter model, or if an improper prior is placed on a joint parameter space, assuming that the posterior is proper, one might claim to have conducted an analysis from the viewpoint of objective Bayes. We would claim one has not conducted a Bayesian analysis at all. The fundamental conceptualization of a Bayesian analysis is that one takes current belief, expressed in terms of epistemic probability through a prior, combines that belief with information contained in a set of observed data, and produces a quantification of updated belief in the form of a posterior. Typically, the prior and posterior are expressed as distributions and information from the data as a likelihood, but this same progression can be formulated in terms of moments rather than distributions, as discussed under the heading of *Subjective Bayes* in Chapter 16.5.2. The progression cannot, however, be formulated if one of the pieces of prior information, information from observations, and resultant posterior information is missing. Asserting that one is totally ignorant about data model parameters is to acknowledge that one of the fundamental pieces, prior information, is missing. As a result, the mathematical exercise of multiplying a joint distribution by a (non-existent) prior and re-normalizing in terms of the parameter does not constitute a Bayesian analysis. Consider the classical case of a one sample normal problem with known variance, $Y_i \sim$ iid $\mathrm{N}(\mu, 1)$, for which an improper prior, $\pi(\mu) = 1$, results in a posterior $\mu|\boldsymbol{Y} \sim \mathrm{N}(\bar{y}, 1/n)$. This example is often touted as showing objectivity, as inference from the posterior is the same as inference from the sampling distribution of $\bar{Y}$. Our

stance is that it simply shows one is basing inference on likelihood theory, not Bayes theory, which means that interpretation of "posterior" quantities in terms of epistemic probability is really just a fantasy. If one interprets a data model using anything other than epistemic probability, then a likelihood, being a function of the parameter for given values of the random variables, is not a statement of probability about the parameter. Taking a likelihood, multiplying it by 1, and re-naming it a posterior does not somehow magically change this fact and suddenly allow interpretation in terms of epistemic probability. The slight-of-hand involved is reminiscent of Fisher's fiducial probability and the famous quote of (Savage, 1961, p. 578) that the fiducial argument was "a bold attempt to make the Bayesian omelet without breaking the Bayesian eggs ". If one is in a situation for which there seems to be no actual prior information relevant to the data model parameters, one should simply admit that a Bayesian analysis is not possible, and turn to other inferential frameworks such as that associated with maximum likelihood. Doing so neither denies the benefits of Bayesian inference nor misses an opportunity to make use of that framework, it is simply admitting that one cannot construct a brick house from wood, nor a treehouse from brick. If one has a situation perfectly designed for a randomization test for group differences, but the determination of which experimental units received which treatment was not made using a random rule, then one must conclude that a randomization test is not appropriate. If one desires estimation of the mean attribute for a finite population of people but the sample was not chosen according to a probabilistic design, then one must conclude that an analysis based on the methods of survey sampling are not appropriate. Similarly, if one is faced with a situation for which there exists no prior information at all, then one must conclude that a Bayesian analysis is not appropriate.

Note that all of these, randomization procedures, survey sampling methods, likelihood analysis, and Bayesian analysis all involve different but legitimate inferential frameworks. In any particular problem there is no guarantee that all, or even any, of them will be fully justified. A caveat for the message of this paragraph is that it may be useful from a research perspective to examine models with all improper priors to determine properties of the model and what conditions, if any, will yield proper posterior distributions.

Despite the rather damning opinion relative to objective Bayes expressed in the previous paragraph, we do not believe that improper priors are without potential uses. The statistical literature has long referred to the presence of "nuisance" parameters in various problems. While in general we are not fond of designating portions of a parameter vector to be a nuisance rather than part of the overall probabilistic structure assigned to a problem, it can be the case that some parameters are more central to the objectives of an analysis than are others. And, in some sense, making marginal inferences about individual elements of a parameter vector is treating other elements of the vector as nuisance parameters so that one element of the parameter vector may be the focal point for one inferential statement (e.g., interval) and part of the collection of nuisance parameters for statements about other elements of the parameter vector. There is perhaps an analogy (but not an equivalence) here with elements of an overall parameter we might have legitimate prior information about and other elements of the parameter for which we have no such information. This is not uncommon, for example, in complex hierarchical models. In such situations we might formulate a joint prior in product form, with individual components reflecting actual prior information as proper marginal distributions or ignorance as improper priors. The question that immediately arises is, if we think it is acceptable

to use improper priors for some of the elements in a parameter vector as long as others have proper priors, but are quite strongly opposed to using improper priors for all elements of the parameter, then for what proportion of the parameter vector are we willing to entertain improper priors? We have no completely satisfying answer to this question, but will again appeal to the literature on dealing with nuisance parameters in which we want the number of parameters falling into the nuisance category to be smaller than the number not so designated and, ideally, considerably smaller.

19.3.5 A Philosophical Parting Shot

We end this chapter, and the book as a whole, with a plea for students and practitioners of statistical science to not ignore the philosophical connections (and perhaps disconnections) between what we do in the analysis of data, and what it means for problems in the real world. As indicted in the quote of Robert Kass at the head of this chapter, our philosophical outlook guides our actions. We might add that our actions are, in return, only meaningful within the context of their philosophical underpinnings. If one does not know what those underpinnings are, or chooses to ignore them because of the belief they are only important in the abstract, then one does not know what one is doing. That the philosophy of statistical inference is messy, incomplete, and poorly agreed on is no excuse. There has been an explosion of data-related procedures that have been developed in the past several decades. Big Data, Machine Learning, Data Analytics, Deep Learning, Natural Language Processing, and perhaps others have burst onto the scene. There are even individuals who identify themselves as "Data Storytellers" on the internet. Depending on which web site one wishes to believe there are 4 Cs of Big

Data, 5 Vs of Big Data (or 5 Ps), and 4 Ps of Data Analytics. Most recently, Data Science has ruled the day, and many departments of statistics have incorporated data science into their names and curriculum so as not to loose out on potential revenue tied to students attracted to that name. While not denying that there are some seemingly impressive results (mostly predictive) that have been produced by some of these techniques, it is unclear what their foundational bases are. My credit union has alerted me to potential fraud concerning my debit card on a number of occasions, denying payment, several of which have been real and for which I am therefore grateful. My credit union has alerted me to potential fraud concerning my debit card on a number of occasions, denying payment, about 90% of which have not been real and have resulted in me needing to re-initiate a payment process, sometimes with additional expense. I'm certain the payment denials issued by my credit union are driven by some type of data analytic algorithm. So how should I judge their performance? Should I believe that their decisions are based on a solid basis or reason, or just correlations computed from massive amounts of data? I do not want fraudulent charges on my debit card, but I am not happy that more often than not a denial of payment was in error and I am required to fix the situation. And I, along with thousands of other patrons of that credit union, have no idea what the logical basis for their decisions consists of. So, should the field of statistics try to claim these various data-related activities? Are they statistics? Is statistics somehow useful in either improving them or assessing them? We would claim that the answer to these questions, at least from an academic standpoint, if not an administrative one, rests on the relation between foundational (yes, philosophical) issues of statistics and whatever data newcomer is of concern.

Finally, I will indicate that I have earned a degree that carries with it

the title of Doctor. But I am not a Doctor of Statistics. I am a Doctor of Philosophy, in Statistics, and the same is true of the hundreds of thousands of others who have completed a PhD in statistics. And the same will be true for current students in doctoral programs in statistics around the world. This title should mean something. In particular, it should imply that an individual so endowed will spend at least some small portion of their career contemplating the foundational aspects of our discipline. In doing so I have learned a great deal that has impacted both my practice and research. I hope and believe it will do the same for others.

Bibliography

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press.

Bartlett, M. (1936). The information available in small samples. *Proceedings of the Cambridge Philosophical Society* **32,** 560–566.

Bartlett, M. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A* **160,** 268–282.

Bartlett, M. (1939). Complete simultaneous fiducial distributions. *Annals of Mathematical Statistics* **10,** 129–138.

Bartlett, M. (1984). Jerzy neyman obituary ii: Neyman and the theory of statistical inference. *Bulletin of the London Mathematical Society* **16,** 169–176.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 2 edition.

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis* **1,** 385–402.

Berger, J. and Bernardo, J. (1992). On the development of reference priors (with discussion). In *Bayesian Statistics 4*, pages 35–60. Oxford University Press, London.

Berger, J. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model against nonparametric alternatives. *Journal of the American Statistical Association* **96,** 174–184.

Berger, J., Strawderman, W., and Tang, D. (2005). Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Annals of Statistics* **33,** 606–646.

Bernardo, J. (1979). Reference posterior distributions for bayesian inference (with discussion). *Journal of the Royal Statistical Society, Series B* **41,** 113–147.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory.* John Wiley and Sons, Chishester.

Buchanan-Wollaston, H. (1935). Statistical tests. *Nature* **136,** 182–183.

Buehler, R. and Feddersen, A. (1963). Note on a conditional probability property of student's t. *Annals of Mathematical Statistics* **34,** 1098–1100.

Carnap, R. (1950). *Logical Foundations of Probability.* The University of Chicago Press, Chicago.

Casella, G., Hwang, J., and Robert, C. (1993). A paradox in decision-theoretic interval estimation. *Statistica Sinica* **3,** 141–155.

Casella, G., Hwant, J., and Robert, C. (1990). Loss functions for set estimation.

Datta, G. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Springer, New York.

Dawid, A.P. Musio, M. (2014). Theory and applications of proper scoring rules. *Metron* **72,** 169–183.

Edwards, A. (1972). *Likelihood*. Cambridge University Press.

Edwards, A. (1976). Fiducial probability. *Journal of the Royal Statistical Society, Series D* **25,** 15–35.

Ellis, R. (1844). On the foundations of the theory of probabilities. *Transactions of the Cambridge Philosophical Society* **8,** 1–6.

Feinberg, S. (2006). Does it make sense to be an "objective bayesian"? commment on articles by berger and goldstein. *Bayesian Analysis* **1,** 429–432.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society or London Series A* **222,** 309–368.

Fisher, R. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fisher, R. (1930). Inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society* **26,** 528–535.

Fisher, R. (1933). The concepts of inverese and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society of London Series A* **139,** 343–348.

Fisher, R. (1935). *Design of Experiments*. Oliver and Boyd, London.

Fisher, R. (1935a). The fiducial argument in statistical inference. *Annals of Eugenics* **6,** 391–398.

Fisher, R. (1935b). The logic of inductive inference. *Journal of the Royal Statistical Society* **98,** 39–54.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B* **17,** 69–78.

Fisher, R. (1973). *Statistical Methods and Scientific Inference, 3rd ed.* Hafner, New York.

Florens, J., Richard, J., and Rolin, J. (1996). Bayesian encompassing specification tests of a parametric model against a nonparametric alternative. Technical Report 96.08, Universite Catholique de Louvain, Institut de Statistique.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6,** 733–807.

Gneiting, T. and Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102,** 359–378.

Goldstein, M. (1997). Prior inferences for posterior judgements. In Chiara, M., Doets, K., Mundici, D., and van Benthem, J., editors, *Structures and*

Norms in Science: Volume Two of the Tenth Internationa Congress of Logic, Methodology and Philosophy of Science, pages 55–71. Kluwer, Dordrecht.

Goldstein, M. (2006). Subjectivity and objectivity in bayesian statistics: rejoinder to the discussion. *Bayesian Analysis* **1,** 465–472.

Good, I. (1992). The bayes/non-bayes compromise: a brief review. *Journal of the American Statistical Association* **87,** 597–606.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press, Cambridge.

Hamada, M., Wilson, A.G abd Reese, C., and Martz, H. (2008). *Bayesian Reliability*. Springer, New York.

Inman, H. (1994). Karl pearson and r.a. fisher on statistical tests: A 1935 exchange from nature. *The American Statistician* **48,** 2–11.

Jaynes, E. (1976). Confidence intervals vs bayesian intervals. In Harper, W. and Hooker, C., editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. Dordrecht D. Reidel.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

Jeffreys, H. (1931). *Scientific Inference*. Cambridge Univeristy Press.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press, Oxford, England.

Jeffreys, H. (1940). Note on the behrens-fisher formula. *Annals of Eugenics* **10,** 48–51.

Joshi, V. (1969). In-admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *Annals of Mathematical Statistics* **40,** 1042–1067.

Kass, R. (2006). Kinds of bayesians (comment on articles by berger and goldstein). *Bayesian Analysis* **1,** 437–440.

Kass, R. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91,** 1343–1370.

Kennedy-Shaffer, L. (2019). Before p¡0.05 to beyond p¡0.05: Using history to contextulize p-values and significance testing. *American Statistician* **Suppl 1,** 82–90.

Keynes, J. (1921). *A Treatise on Probability*. McMillan and Co., London.

Kolmogorov, A. (1950 [1933]). *Foundations of the Theory of Probability*. Chelsea Publishing Company, New York.

Koopman, B. (1940). The axioms of algebra of intutitive probability. *Annals of Mathematics* **41,** 269–292.

Kyburg, H. (1971). Epistemological probability. *Synthese* **23,** 309–326.

Kyburg, H. (1974). *The Logical Foundations of Statistical Inference*. Reidel, Dordrecht.

Lad, F. (2006). Objective bayeisan statistics...do you buy it? should we sell it? comment on articles by berger and goldstein. *Bayesian Analysis* **1,** 441–444.

Laplace, P. (1814). *A Philosophical Essay on Probabilities, 6th ed.* Dover Publications, New York.

Lehmann, E. (1995). Neyman's statistical philosophy. *Probability and Mathematical Statistics* **15,** 29–36.

McVinish, R., Rousseau, J., and Mengersen, K. (2008). *Scandanavian Journal of Statistics* **36,** 337–354.

Neyman, J. (1934). On the two different aspects of the representative method. *Journal of the Royal Statistical Society, Ser. A* **97,** 558–625.

Neyman, J. (1935). On the problem of confidence intervals. *Annals of Mathematical Statistics* **6,** 111–116.

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika* **32,** 128–150.

Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *Revue d'Institute International de Statistique* **25,** 7–22.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36,** 97–131.

O'Hagan, A. (1994). *The Advanced Theory of Statistics, Vol. 2B: Bayesian Inference.* Edward Arnold, London.

O'Hagan, A. (2006). Science, subjectivity and software. comment on articles by berger and goldstein. *Bayesian Analysis* **1,** 445–450.

Pearson, E. (1955). Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society: Series B* **17,** 204–207.

Pearson, K. (1892). *The Grammer of Science*. Scott, London.

Pearson, K. (1916). On a brief proof of the fundamental formula for testing the goodness of fit of frequency distributions and on the probable error of p. *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science* **31,** 369–378.

Pearson, K. (1922). On the χ^2 test of goodness of fit. *Biometrika* **14,** 186–191.

Pearson, K. (1935). Statistical tests. *Nature* **136,** 550.

Pearson, K. (1941). The laws of chance, in relation to thought and conduct. *Biometrika* **32,** 89–100.

Pollock, J. (1990). *Nomic Probability and the Foundations of Induction*. Oxford University Press, New York.

Ramsey, F. (1931). *The Foundations of Mathematics*. Routledge and Kegan Paul, London.

Reichenbach, H. (1949). *Theory of Probability*. University of California Press, Berkley and Los Angeles.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd ed.* Springer, New York.

Savage, L. (1961). Foundations of statistics revisited. In Neyman, J., editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 575–586. University of California Press, Berkeley and Los Angeles.

Seidenfeld, T. (1978). Direct inference and inverse inference. *The Journal of Philosophy* **75,** 709–730.

Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*. D. Reidel, Dordrecht, The Netherlands.

Sidenfeld, T. (1992). R. a. fisher's fiducial argument and bayes theorem. *Statistical Science* **7,** 358–368.

Venn, J. (1866). *The Logic of Chance. An Essay on the Foundations and Province of the Theory of Probability, with Especial Reference to its Applicaitons to Moral and Social Science*. MacMillan and Co., London and Cambridge.

Venn, J. (1889). *The Principles of Empirical or Inductive Logic*. MacMillan and Co., London.

Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness of fit testing using infinite dimensional exponential families. *The Annalys of Statisitcs* **20,** 1203–1221.

von Mises, R. (1957). *Probability, Statistics and Truth*. Allen and Unwin, London.

Zabell, S. (1992). R. a. fisher and the fiducial argument. *Statistical Science* **7,** 369–387.