# BAYESIAN ANALYSIS OF ORDERED CATEGORICAL DATA FROM INDUSTRIAL EXPERIMENTS

H. Chipman and M. Hamada

IIQP Research Report
RR-93-06

August 1993

# BAYESIAN ANALYSIS OF ORDERED CATEGORICAL DATA FROM INDUSTRIAL EXPERIMENTS

*Hugh Chipman and Michael Hamada*
Department of Statistics and Actuarial Science
and
The Institute for Improvement in Quality and Productivity
University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

August 1993

*ABSTRACT*

Data from industrial experiments often involve an ordered categorical response, such as a qualitative rating. ANOVA based analyses are often inappropriate for such data, suggesting the use of Generalized Linear Models (GLMs). When the data are observed from a fractionated experiment, likelihood-based GLM estimates quite often do not exist, especially when there is an opportunity for making substantial improvement. These difficulties are overcome with a Bayesian GLM, which is implemented via the Gibbs sampling algorithm. Techniques for modelling data and for subsequently using the identified model to optimize the process are outlined. An important advantage in the optimization stage is that uncertainty in the parameter estimates is accounted for in the model. These techniques are used to analyze several data sets. In robust design experiments, the variability in the parameters is easily incorporated with the response modelling approach (Welch, Yu, Kang and Sacks (1989), Shoemaker, Tsui and Wu (1991)).

*Key words: Gibbs Sampler, Generalized Linear Model, Robust Design, Binary Data*

# 1 Introduction

Data sets with an ordered categorical response arise naturally in a number of industrial situations. Such a response can be characterized as a random variable that takes on a number of discrete outcomes or categories, which have an implicit ordering, such as "poor", "average", and "good". While there may be any number of categories, a small number (two to six) is common. Also, a binary response is an important special case.

In industrial situations, such data may be quick surrogates for continuous measurements. This occurs when the main priority is to run the experiment quickly, and if necessary consider a more complete measurement later. For example, we might be interested in the force required to close a door. While devices exist to make such measurements, they may be expensive or time-consuming, so an expert might be used to make categorical judgments, such as "difficult", "acceptable", and "easy". In other situations, such as in assessing the physical appearance of steel, the appearance can be graded but there is no obvious underlying continuous response. Moreover, it might require much ingenuity and cost to develop a device which could measure the appearance appropriately on some continuous scale.

The data considered in this paper have several features specific to industrial experiments. For example, the predictor variables (or factors) are typically set at a relatively small number of levels, and are arranged in an orthogonal array. Also, the number of main effects (and interactions) is typically close to the number of observations made. This means that the models considered will be relatively simple, and identification of important effects and interactions will be a priority.

Interest focuses on modeling the way in which predictor variables influence the distribution of the ordered categorical response. Several such models are currently in use and are outlined in Section 2. All of these approaches have difficulties, namely inappropriate or inaccurate inference, as well as non-existent estimates of factor effects. A Bayesian model is proposed to circumvent these difficulties. Previously, Hamada and Wu (1992) showed the benefits of using a Bayesian approach for analyzing censored data because it avoided the problem of non-existent estimates from standard methods.

The Bayesian model for ordered categorical data is described in Section 3 along with the Gibbs sampler, a computational technique used to fit the model. In Section 4, an analysis strategy for industrial data using the Bayesian model is outlined. The strategy includes a model selection stage and an optimization stage in which factor levels are chosen. The Bayesian nature of the model makes it easy to model uncertainty in the parameters and allows for easy integration of noise factors in the case of a Taguchi-style robust design experiment. Data from industrial experiments are used to illustrate the strategy in Section 5. In Section 6, a summary and conclusions are given.

# 2 Existing Analysis Techniques

Several techniques are available for the analysis of ordered categorical data. Among those used for industrial situations are scored ANOVA, accumulation analysis, and generalized linear models. In this section, we consider these techniques and some of the problems they have.

The most straightforward method of analysis assigns arbitrary (strictly increasing) scores to the ordered response categories, and performs an ANOVA on these scores. While this approach is

familiar, there are some complications. Any conclusions made will depend on the scores used as discussed in Hamada and Wu (1990). Furthermore, the scored data are not continuous and the normal distribution may be an unreasonable distributional approximation.

Another technique, Accumulation Analysis (AA), is an adapted ANOVA which uses the cumulative frequencies of the ordered categories as a response. It was introduced by Taguchi (1974) as an improvement over Pearson's chi-squared test, which does not use the order of the categories. There are a number of serious criticisms raised against AA by Nair (1986), Hamada and Wu (1990) and discussants therein, which include detection of spurious effects, testing for a combination of location and dispersion effects, and reversing the order of factor importance. The method's shortcomings make further discussion inappropriate, and the reader is referred to their papers for further details.

Many of the problems with ANOVA based techniques relate to the validity of the inference they make. A search for statistically valid methods leads naturally to Generalized Linear Models (GLMs) for ordinal data. These models are attractive since software is readily available for fitting them.

We can summarize the method as follows: instead of creating a pseudo measurement to be analyzed, the probabilities of the $J$ ordered categories are modeled as a function of the predictors. McCullagh (1980) suggests a family of models of the form:

$$\text{link}(\Pr(Y \leq j)) = \gamma_j - \mathbf{x}'\beta \qquad \text{for} \qquad j = 1, ..., J-1$$

where "link" is a (known) monotone increasing function mapping the interval (0,1) onto the real line $(-\infty, \infty)$, $\gamma_j$ is a "cutpoint", $\mathbf{x}$ is the vector of main effects and interactions, and $\beta$ is a vector of effects. The $\gamma_j$s may be thought of as intercepts, and consequently there is no intercept term in the vector $\mathbf{x}$. A natural ordering of cells is obtained by modeling cumulative cell probabilities rather than individual probabilities. Although each response category has a corresponding cutpoint, the regression coefficients $\beta$ are constant across outcomes.

Once a link function is specified, estimation of the parameters $(\gamma_1, \gamma_2, ..., \gamma_{J-1}, \beta)$ is carried out via maximum likelihood (ML). One problem with this approach is that estimates for certain coefficients or cutpoints may not exist. For example, suppose we have a binary response and a factor $A$ with two levels. If all of the observations at the low level of $A$ are in the first category, and all the observations at the high level of $A$ are in the second category, then the estimated cell probabilities are zero and one. This corresponds to an estimate of $+\infty$ for the coefficient of $A$. In multifactor experiments, this problem is quite likely to occur, especially when the number of effects in the model is near the number of runs. These problems will manifest themselves as estimates which fail to converge. Tse (1986) gives conditions under which MLEs are infinite.

Despite the non-convergence of estimates, the likelihood does converge, and may be used as a criterion for assessing the relative importance of the factors. Such an approach is discussed in Lawless and Singhal (1978). However, in situations where the model is to be used for prediction, or comparisons between effects are necessary, non-existent estimates are a problem.

One solution would be to assume some sort of prior knowledge of the coefficients, and use Bayesian techniques to fit a model to the data. Even when knowledge about the coefficients is minimal or nonexistent, this approach has justification; we suspect that the parameters are only large, and that there is not sufficient data to distinguish between large and infinite values.

Thus, the Bayesian approach retains the advantages of GLMs, which do not require a choice of scores, and provides a more accurate description of the response distribution. At the same time,

the issue of infinite estimates is resolved; moreover asymptotic approximations are not necessary for inference. The details of this method are provided in the next section, including the use of Gibbs sampling techniques for implementing the analysis.

# 3 A Bayesian Approach

In this section, the theory and tools for fitting a Bayesian GLM are outlined. The main focus is the Gibbs sampler, a computational technique used in the calculation of marginal posteriors of the parameters of interest. A brief summary of the Gibbs algorithm is given before specific results are derived.

## 3.1 The Gibbs Sampler

The Gibbs sampler, which was introduced by Geman and Geman (1984), is a technique for the calculation of marginal distributions of random variables, given a set of conditional distributions. In the Bayesian context, it is used to obtain marginal posteriors of the parameters after conditioning on the observed data $Y$. This technique is seen as a simpler alternative to numerical integration techniques, which may require specialized knowledge and attention to the details of a specific problem. The applications of this technique to statistical problems have been more recently outlined in a number of papers, such as Gelfand and Smith (1990). We note that for the problem considered, a related technique data augmentation (Tanner and Wong (1987)) cannot be used since the response category division as defined by the underlying continuous scale cutpoints are not known.

To illustrate the algorithm, consider three random variables $U_1, U_2$, and $U_3$. The notation $f(\cdot)$ will denote the density of the argument. Thus, $f(x)$ is the density of $X$, $f(y)$ is the density of $Y$, but $f(x) \neq f(y)$.

We wish to determine the marginal distributions $f(u_1)$, $f(u_2)$, and $f(u_3)$. It is assumed that the distributions $f(u_i|u_j, j \neq i)$ are available (i.e., they may be sampled from). We start with an arbitrary set of values $(U_1^{(0)}, U_2^{(0)}, U_3^{(0)})$. We sample $U_1^{(1)} \sim f(u_1|U_2^{(0)}, U_3^{(0)})$, $U_2^{(1)} \sim f(u_2|U_1^{(1)}, U_3^{(0)})$, and $U_3^{(1)} \sim f(u_3|U_1^{(1)}, U_2^{(1)})$. The algorithm then cycles through the conditional distributions repeatedly in the same fashion, always conditioning on the most recent values of $(U_1, U_2, U_3)$.

More generally, for $k$ random variables, we cycle through each of the $k$ distributions of one variable conditioned on the other $k-1$ variables. Geman and Geman (1984) show that these random variables converge in distribution to a sample from the joint distribution. Consequently, any subset of the $k$ variables can be viewed as a sample from the appropriate marginal distribution. The general application of the sampling algorithm described above is to iterate the sampling cycle $m$ times, and treat the final observation $(U_1^{(m)}, ..., U_k^{(m)})$ as an approximate sample from the distribution $(U_1, ..., U_k)$. An arbitrary number of such samples may be generated by repeatedly restarting the algorithm.

## 3.2 The Gibbs Sampler for Ordinal Data

The model to be fit is a Bayesian version of the GLM described in the previous section, with a probit link, a normal prior for the coefficients $\beta_i$, and an ordered normal prior for the cutpoints $\gamma_j$. The derivation of conditional distributions for the Gibbs sampler is similar to Albert and Chib

(1993), but with informative priors. The ability to vary the shape of the prior will allow us to assess the robustness of the conclusions from the analysis.

The form of the fully conditional distributions is simplified by the assumption of an underlying continuous variable associated with the observed categorical response. In industrial applications there is quite often such a variable, and an ordered categorical version of it is observed due to cost or time constraints.

We assume that for each response $Y$, there is an unobserved variable $Z$ on the continuous scale. The correspondence between this variable and the $J$ categories of the response is via "cutpoints" $\gamma_0, \gamma_1, ..., \gamma_{J-1}, \gamma_J$, where $-\infty = \gamma_0 < \gamma_1 < ... < \gamma_{J-1} < \gamma_J = \infty$. If $\gamma_i \leq Z < \gamma_{i+1}$, then $Y = i + 1$ is observed. The distribution of the variable $Z$ is determined by the form of the link function, since we require that $\text{link}(\Pr(Y \leq j)) = \gamma_j - \mathbf{x}'\beta$, or equivalently that $\text{link}(\Pr(Z < \gamma_j)) = \gamma_j - \mathbf{x}'\beta$. The link function $F^{-1}$ maps $(0,1)$ onto $(-\infty, \infty)$ , so if $Z + \mathbf{x}'\beta$ has distribution function $F$, then we obtain $\Pr(Y \leq j) = \Pr(Z \leq \gamma_j) = \Pr(Z + \mathbf{x}'\beta \leq \gamma + \mathbf{x}'\beta) = F^{-1}(\gamma + \mathbf{x}'\beta)$.

Fully conditional distributions are now derived for arbitrary independent priors on $\beta$ and $\gamma$, as well as an arbitrary link function. The fully conditional distributions are required only up to a proportionality constant, since they are used only for random variable generation. Consequently, we begin with the joint distribution of $Y, \gamma, \beta, Z$. This distribution is degenerate, since knowledge of $(Z, \gamma)$ determines $Y$ exactly. Thus, we may write:

$$f(\beta, \gamma, z, y) = f(\beta, \gamma, z) I(y; z, \gamma),$$

where the indicator function $I$ is 1 when the vector $y$ agrees with $(\gamma, z)$, and 0, otherwise. Using conditioning, and assuming independence of priors for $\gamma$ and $\beta$ yields:

$$\begin{aligned} f(\beta, \gamma, z, y) &= f(z|\beta, \gamma) f(\beta, \gamma) I(y; z, \gamma) \\ &= f(z|\beta, \gamma) f(\beta) f(\gamma) I(y; z, \gamma). \end{aligned}$$

Note that the distribution of $Z$ given $(\beta, \gamma)$ does not depend on $\gamma$. This follows, since the only dependence of $Z$ on $\gamma$ is through $Y$, which is not being conditioned upon in the first term of the expression above. Thus, we obtain

$$f(\beta, \gamma, z, y) = f(z|\beta) f(\beta) f(\gamma) I(y; z, \gamma).$$

This yields the fully conditional distributions:

$$\begin{aligned} f(\beta|\gamma, z, y) &\propto f(z|\beta) f(\beta) \\ f(\gamma|\beta, z, y) &\propto f(\gamma) I(y; z, \gamma) \\ f(z|\beta, \gamma, y) &\propto f(z|\beta) I(y; z, \gamma). \end{aligned}$$

The probit link $F(x) = \Phi(x)$ is used in this implementation. This means that $Z$ will be normally distributed, and a multivariate normal prior on $\beta$ will be convenient. The choice of the prior for $\gamma$ is flexible since $f(\gamma)$ only appears in one of the three conditional distributions, and there it is simply truncated. For ease of interpretation, we choose a prior for $\gamma$ in which the variates are normal except for the ordering; that is, one of the form $\phi(\gamma_1, ..., \gamma_{J-1}) \times I(\gamma_1 < \gamma_2 < ... < \gamma_{J-1})$, where $\phi$ is the multivariate normal density. As an alternative, a diffuse prior for $\gamma$ is also considered.

With these choices, we obtain the following fully conditional distributions: $\beta$ is multivariate normal, $\gamma$ is truncated normal (truncated by the $Z$'s), and $Z$ is truncated normal (truncated by the $\gamma$'s).

# 4 Analysis Techniques

This section summarizes techniques for the analysis of industrial data with the Bayesian method. Several of these methods and concepts were explored in Hamada and Wu (1992). For example, their approach to variable selection was to label as significant those factors whose marginal posteriors were far from zero. They also emphasized the importance of robustness of conclusions to the shape of the priors. In this section these techniques are refined, and combined with graphical displays and a new method of optimization using posterior distributions.

The first group of techniques is related to model identification. Assuming that the experiment has already been run, the first task is to identify the factors with the largest effects. A scored ANOVA could be used to determine an approximate ranking of factors. For a more accurate ranking, we examine the posteriors generated by a Bayesian analysis. Although marginal posteriors of all effects should be examined, numerical summaries of the posteriors are useful when the number of main effects and interactions is large.

One convenient measure of factor importance is the proportion of a marginal posterior on one side of zero. The "zero p-value" is defined in terms of this quantity, in a fashion similar to two sided p-values used in standard hypothesis testing:

$$p_k = 2\min(\hat{F}_k(0), 1 - \hat{F}_k(0)) ,$$

where $\hat{F}_k$ is the estimated distribution function of the parameter of interest. Small values of $p_k$ indicate high probability that the corresponding effect is significant. This statistic will be invariant to rescaling of the factors, eliminating the need to choose a scaling metric.

If all factors can be expressed on a comparable scale, other summaries of the posterior may be examined, such as the (marginal) posterior mean or median. These quantities represent the magnitude of the effect, which is also an important consideration in model identification. Quantities such as the posterior standard deviation may provide information about uncertainty in the location measures.

Whatever summary of marginal posteriors is chosen, it will depend on the priors. This dependence may be studied via a new graphical display, in which summaries of all marginal posteriors are plotted as a function of prior variance. For example, if we have prior variance $\sigma^2\Sigma$, the posterior median effects could be plotted for $\sigma \in (0.10, 1.0)$. An example of this is given in Figure 1. While the magnitude of some summaries may change over prior variances, the rankings of factor effects quite often do not. This robustness to prior information, which was first recognized by Hamada and Wu (1992), increases our certainty in the factors and interactions in the model. This type of plot offers many possibilities, since we can study the effect of changing prior variance for one set of parameters with the prior variance held constant for the remaining parameters.

The rationale for the upper limit of $\sigma$ is related to maximum effect a factor can have on a response probability. Consider a normal prior on the $\beta$s, and a factor $x$ with levels $\pm 1$. If $\beta_x$ has a N(0,$\sigma$) prior, then a priori it will fall in the range $(-2\sigma, 2\sigma)$ approximately 95% of the time. Consequently, the effect (on the transformed scale) of changing $x$ from -1 to 1 will be no more than $4\sigma$, 95% of the time. Now it is the effect on the cumulative probabilities that is of interest. Since the link function is nonlinear, the effect of the factor depends on the value of the linear predictor. For links based on symmetric unimodal distributions, such as the logistic or probit, the maximum change on the probability scale will occur when the rest of the linear predictor sums to zero. For

the probit link, the largest magnitude of the effect will be a change from $\Phi(-2\sigma)$ to $\Phi(2\sigma)$. For $\sigma = 1$, we see that this implies a change of no more than 0.025 to 0.975, 95% of the time.

The scaling of the factors clearly influences the choice of this range. Thus it is important that the factors all be scaled prior to modeling. In all the examples presented in this paper, the design matrix $X$ with run size $n$ has been scaled so that $X^TX/n$ is equal to an identity matrix.

Once a model has been identified, it is necessary to draw inferences about the best factor level settings. A general method for this is to consider all combinations of levels for factors identified as important. For each setting, we have a posterior distribution on the response categories. This is an improvement over the more standard procedure of using point estimates, which ignores model uncertainty represented by the posteriors.

When there are many settings, we may first wish to consider category probabilities which are averaged over the posteriors. These probabilities can be plotted in the case of a response with two or three categories. For three categories we may plot Pr(category 1) against Pr(category 2). Depending on our goal, different regions of this plot will be best. For example, if categories (1,2,3) are (good, ok, poor), then we would want high values of Pr(good) with most of the remaining probability taken up by Pr(ok). In Figure 6 we see a plot of the average posterior probability of a good part against the average posterior probability of an "ok" part. Each number represents a specific setting of factor levels, so we may use this plot to identify the factor settings that produce the best (average) posterior probabilities for the response. Once several promising settings are identified, the full posteriors for the response proportions may be examined.

Other performance measures may be used in addition to probabilities of the outcome categories. We may assign a score to each category, and then consider the distribution of these scores due to effect uncertainty. For example, the scores (1,2,3) could be used for a ordered response in three categories. We would examine 1xPr(outcome 1) + 2xPr(outcome 2) + 3xPr(outcome 3) over the posterior of the effects. These performance measures may be easier to interpret than outcome probabilities since they are univariate. For example, rankings of factor levels by the proportion of "ok" or "good" parts is easier if a univariate criterion is used instead of directly examining the probabilities of three outcome groups "poor", "ok", and "good". As with the response probabilities, we may examine means first and distributions afterwards.

All of the techniques mentioned above are also applicable to Robust Design experiments with minor modifications. In this situation, factors are divided into two groups: control and noise factors. The goal of the analysis is to find control factor settings that result in favorable outcomes, regardless of the levels of the uncontrollable noise factors. The model selection stage is unchanged, with control factors, noise factors, and interactions between them all be assessed for their importance. Modeling the response directly as we have done was proposed by Welch, Yu, Kang and Sacks (1989) and named the response model approach by Shoemaker, Tsui and Wu (1991). Once a model has been selected, we assume that the noise factors follow some distribution. Using the model, we seek control factor settings that produce the best results subject to both the model uncertainty and the extra variation arising from the uncontrolled noise factors.

The additional uncertainty due to the noise factors may be dealt with in a straightforward and convenient way. This variation is treated in the same fashion as effect uncertainty, since both are represented by distributions. This flexibility to accommodate both model uncertainty and external variation is another advantage the Bayesian approach offers in analyzing such data.

| | Control Factors | | Noise Factors |
|---|---|---|---|
| A | Drilling (Sequential / Concentric) | H | Brake type (Master / Slave / Control Valve) |
| B | Clamping (top & bottom / side) | I | Position (Left / Right) |
| C | Reamer Set (a / b) | | |
| D | Housing Heat (yes/no) | | |
| E | Cutting Fluid (Synthetic / Mineral) | | |
| F | Feed Speed (min / max) | | |
| G | Piston Material (a / b) | | |

Table 1: Factors in the brake bore experiment

# 5 Applications

In this section, examples from industrial experiments will be used to illustrate the application of the Bayesian approach. Three examples are used, including two robust design experiments and one fractional factorial.

## 5.1 A Brake Bore Experiment

The data considered arise from an experiment for optimizing the quality of piston bores in grey iron brake housings. The data and original analysis are given in Bostelman, Buck and Henry (1987). The response is a binary classification (good/bad) of brake bore surface texture.

The experimental design is a product array design, with seven control factors at two levels each and two noise factors, one with two levels, and the other with three. Factor names and levels are given in Table 1. The noise factors are arranged in a full factorial design, the control factors in an eight run fractional factorial. Three observations are taken at each level of the control and noise combinations, making a total of 8x6x3=144 observations.

The two degrees of freedom for noise factor H are represented by orthogonal contrasts $H_l$ and $H_q$. Despite the qualitative nature of factor H, the linear contrast is interpretable (comparison of types 1 and 3), and the quadratic contrast has some meaning (comparison of type 2 with mean of the other two types). This facilitates a scored ANOVA analysis which is done for comparison with the Bayesian approach. The design matrix $X$ was scaled so that $X^TX/n = I$, putting the posteriors of the regression coefficients roughly on a comparable scale.

The full model has 33 coefficients, as well as a cutpoint. These 33 coefficients include seven control main effects, three noise main effects, the noise by noise interactions $IH_l$ and $IH_q$, and 21 interactions between the seven control main effects and the three noise main effects.

The MLEs for a GLM with these 34 parameters do not exist, suggesting that a Bayesian approach be taken. The prior for the cutpoint is chosen to be diffuse, since interest is focused on the regression coefficients. The prior for the 33 coefficients is taken to be multivariate normal, with zero mean vector and covariance matrix $I_{33}\sigma_\beta^2$, where $I_{33}$ is the 33x33 identity matrix. Several values in the range (0,1) are used for the variance $\sigma_\beta^2$. The effect of varying the prior variance can be seen in Figure 1. These plots represent two characteristics of the posteriors - the posterior median, and the "zero p-values" described in Section 4. We see that for the most part, the same

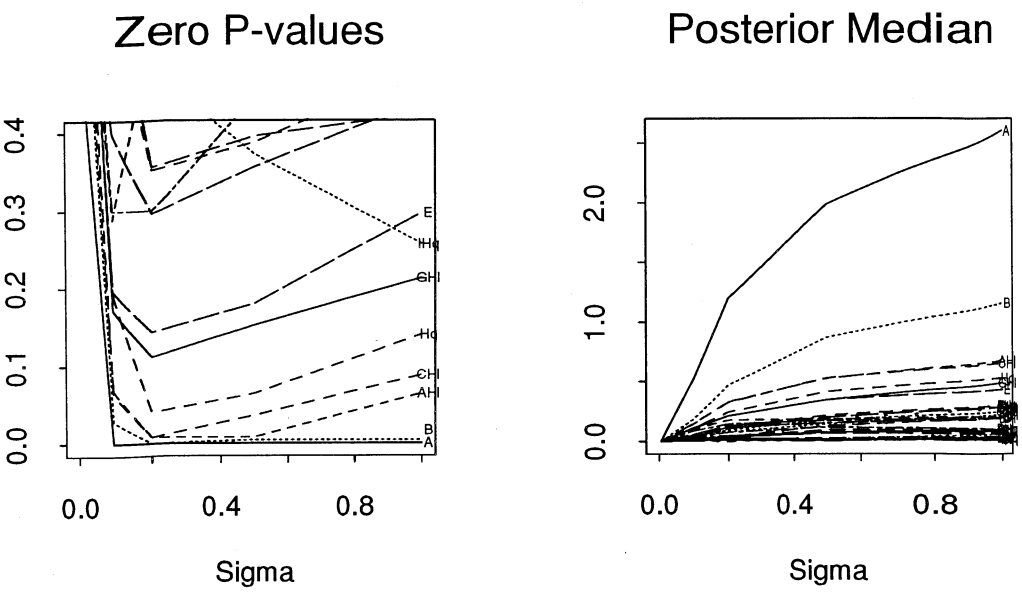


Figure 1: Posterior medians and zero p-values with $\sigma_\beta \in (0.005, 1)$ for the brake bore experiment

factors remain important across varying priors.

A final model is selected using the zero p-values as a criterion, with prior $\sigma_\beta = 1$. All models considered had the hierarchical (or "hereditary") property that for every interaction between two main effects, the model also included those main effects. Coefficients whose posterior medians were close to zero were removed from the model in three stages. The first reduction was to a model with terms A, B, C, E, F, G, $H_l$, $H_q$, I, $AH_l$, $CH_l$, $CH_q$, $FH_q$, $GH_l$, $GH_q$, and $IH_q$. Terms F, I, $CH_q$, $FH_q$, and $IH_q$ were dropped to obtain the second reduced model. At this stage, all terms involving factor G appeared insignificant, and were also removed from the model.

The final model contained the terms A, B, C, E, $H_l$, $H_q$, $AH_l$, $CH_l$. Of these, C and $H_l$ were not significantly different from zero, but were included to maintain the hierarchical model structure. Histograms of their posteriors for $\sigma_\beta = 1$ (a reasonably diffuse prior) are given in Figure 2, including superimposed prior densities. The "o" represent central regions of 95% probability. From these plots, we see the magnitude of the effects, as well as our certainty about their values.

The fitted model may now be used to draw inference about the process. We are interested in identifying factor settings that produce a high proportion of good parts over differing levels of noise factor H and uncertainty of effects. We look at the percentage of good parts at all possible (+/-) levels of important factors. These percentages are first averaged over the noise factor H, and over the posteriors of the parameters $\beta_1, ..., \beta_8, \gamma_1$ so that promising factor levels may be identified. Since H is a qualitative variable representing types of parts, we assume that the types are produced in equal proportions; its distribution consists of point masses at $\{1, 2, 3\}$ with probabilities $\frac{1}{3}$.

In Table 2, we present mean posterior probabilities of obtaining a good part at each of the 16 factor settings. We see that the settings $(A, B, C, E) = (1, 1, 1, -1)$ and $(1, 1, 1, 1)$ appear to have

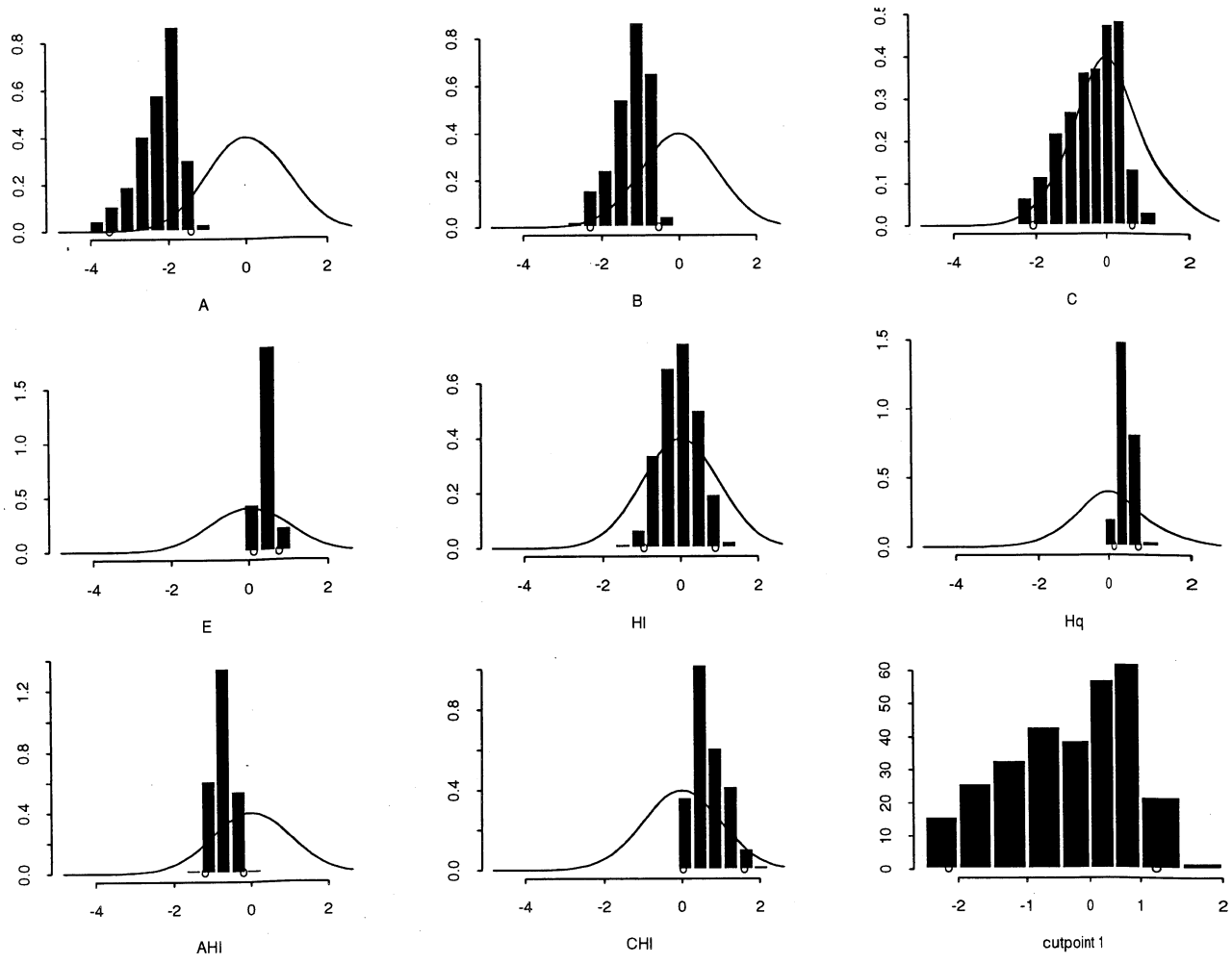


Figure 2: Posteriors for distributions and superimposed priors with $\sigma_\beta = 1$ for the final model of the brake bore experiment

| Setting | A | B | C | E | Pr(good) | Score |
|---|---|---|---|---|---|---|
| 9 | -1 | -1 | -1 | 1 | 0.015 | 2.06 |
| 13 | -1 | -1 | 1 | 1 | 0.016 | 1.92 |
| 1 | -1 | -1 | -1 | -1 | 0.063 | 1.92 |
| 5 | -1 | -1 | 1 | -1 | 0.074 | 1.79 |
| 11 | -1 | 1 | -1 | 1 | 0.197 | 1.84 |
| 15 | -1 | 1 | 1 | 1 | 0.290 | 1.69 |
| 3 | -1 | 1 | -1 | -1 | 0.329 | 1.70 |
| 10 | 1 | -1 | -1 | 1 | 0.415 | 1.41 |
| 7 | -1 | 1 | 1 | -1 | 0.486 | 1.56 |
| 2 | 1 | -1 | -1 | -1 | 0.551 | 1.25 |
| 14 | 1 | -1 | 1 | 1 | 0.716 | 1.26 |
| 12 | 1 | 1 | -1 | 1 | 0.854 | 1.17 |
| 6 | 1 | -1 | 1 | -1 | 0.898 | 1.12 |
| 4 | 1 | 1 | -1 | -1 | 0.945 | 1.03 |
| 16 | 1 | 1 | 1 | 1 | 0.989 | 1.04 |
| 8 | 1 | 1 | 1 | -1 | 0.998 | 0.90 |

Table 2: Bayesian analysis and scored ANOVA evaluation of factor settings for the brake bore experiment

the best chance of producing a good part. The former setting is only slightly better which reflects the marginal importance of E.

Based on this table, we may recommend the setting (A,B,C)=(1,1,1) as the optimum value. It is interesting that if the factor C is omitted from the model, the conclusions vary slightly, with settings 8, 4, 12, 16 becoming the four most important settings. The binary nature of the data suggests that criteria other than averaged probabilities will yield similar recommendations. Plots of posteriors for the proportion of good parts in Figure 3 confirm the claim that a mean measure is adequate, since settings 8 and 16 have the best distributions.

The Bayesian analysis technique used here may be compared with an ANOVA of a response scored as 1 for good, 2 for bad. The ANOVA identifies all of the same factors found above as significant, and in addition identifies G and $GH_l$ as important (possibly because the variation in the binomial observations is less than normally distributed ones as assumed by ANOVA). For comparison, we consider the same model used in the Bayesian analysis above. We then average over the noise to obtain a score representing the average quality of the produced parts. These scores for the 16 possible settings of A, B, C, and E are displayed in Table 2.

The same four settings 8, 4, 16, 6, offer the best quality product, although the order has changed slightly; 4 appears better than 16. Some of the scores fall outside the original range, and the two best combinations appear different. Our analysis suggests that settings 8 and 16 are nearly the same, however. Moreover, the interpretation of the scores is difficult. Since the scores are 1 and 2, it is not clear what is meant by an estimated score of 0.90 or 1.26. The interpretation offered by the Bayesian analysis seems far more intuitive, since it gives estimates for the proportion of good parts at a given factor setting.

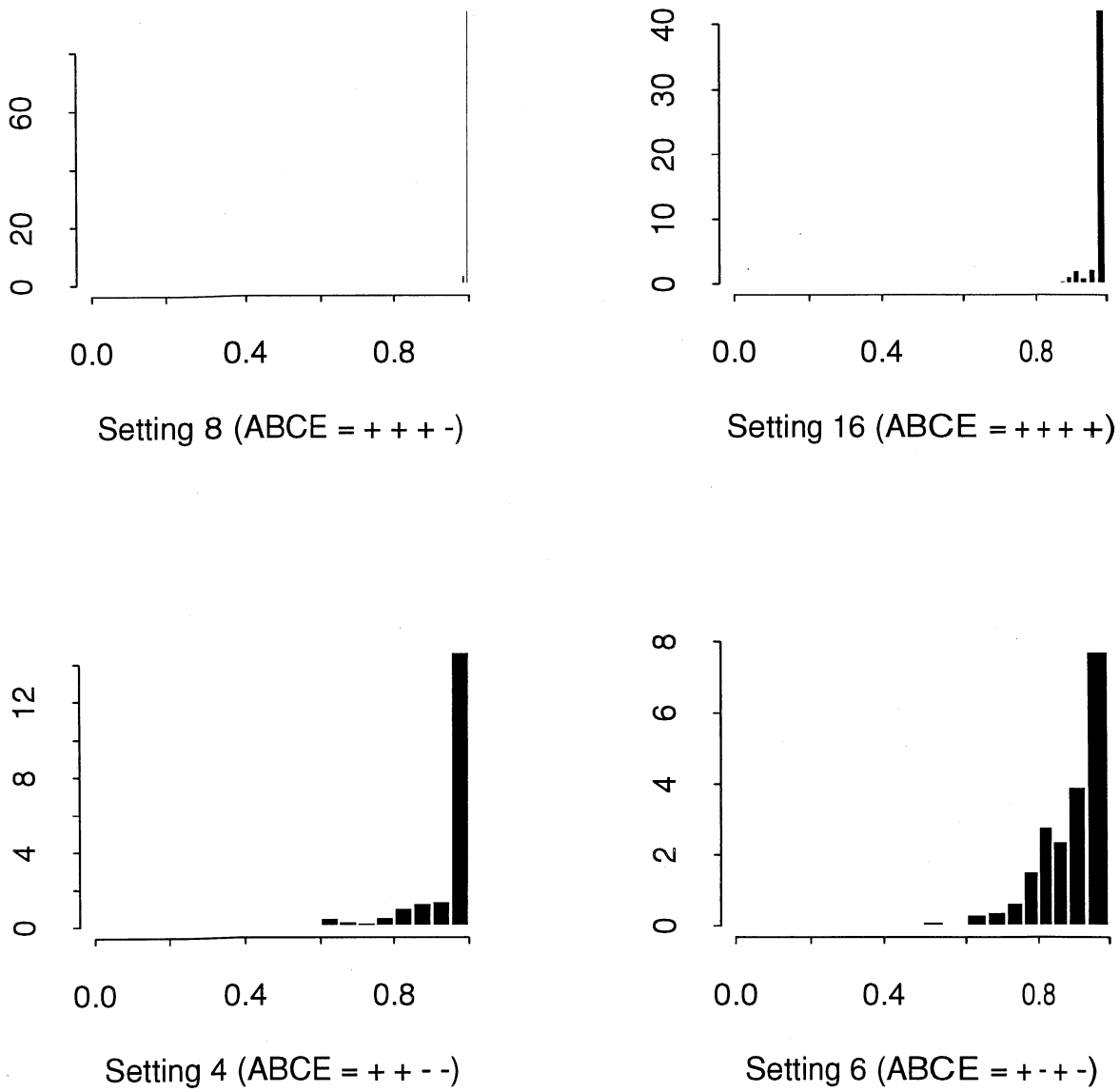


Figure 3: Posterior for proportion of good parts at factor settings 8,16,4,6 for the brake bore experiment

| | Control Factors | | Noise Factors |
|---|---|---|---|
| A | Shot Weight (185 / 250) | H | Shift (second / third) |
| B | Mold Temperature (70° F / 120° F) | I | Shell quality (good / bad) |
| C | Foam Block (use / do not use) | | |
| D | RTV Insert (use / do not use) | | |
| E | Vent Shell (vented / unvented) | | |
| F | Spray Wax Viscosity (2:1 / 4:1) | | |
| G | Tool Elevation (level / elevated) | | |

Table 3: Factors in the foam molding experiment

## 5.2 A Foam Molding Experiment

The data, which was originally analyzed by Jinks (1987), arise from an experiment designed to reduce voids in a urethane-foam product. The response consists of three levels (very good, acceptable, needs repair), while all the design variables are at two levels. The design is an fractionated eight run control array crossed with a four run noise array. The factors are shown in Table 3. At each level of the control and noise factors, ten parts are classified into one of the three categories, yielding a total of 8x4x10=320 observations.

The full model consists of main effects for factors A-I, an HI interaction, and control by noise interactions between factors A-G and H, I. In total, there are 24 effects plus two cutpoints which define the three ordered categories. As in the brake bore example, the Bayesian analysis uses a diffuse prior for the two cutpoints, and an independent normal prior with zero mean vector for the coefficients.

The relative importance of regression coefficients seems insensitive to the choice of prior variance, as illustrated in Figure 4. The effects A, B, C, E, F, G, H, I, HI, AH, and EI are identified as the most important by the zero p-value criterion. A full ranking of the factors by their "zero p-value" is given in Table 4.

This reduced hierarchical model is fit and all of the factors remain important. Posteriors are given in Figure 5. A scored ANOVA (with the scores 1,2,3) lends support to the the significance of these results, since the same factors are identified as influential, in roughly the same order.

Next, we assess the behavior of the process under uncontrolled noise factors H and I. Since no information is available about the behavior of the noise factors under production, they are assumed to be uncorrelated normal random variates, with mean zero and variance one. The model is then used to calculate the mean proportion of observations in each of the response categories, averaging over both the noise factors, and the posteriors of the coefficients.

We first consider the expected proportions at all 64 settings of factors A, B, C, E, F, G, averaged over posteriors and noise. In Figure 6 we see that the optimal setting is number 10, since it has the highest proportion of very good parts, and the majority of the remaining parts are acceptable. This plot allows us to identify settings 14, 26, and 2 as other possible choices. The averaged probabilities of the three outcomes are displayed for these four settings along with the next six best settings in Table 5. From the table, it seems clear that the best setting is number 10, since it produces 87% good parts as compared to 74% for the next best setting.

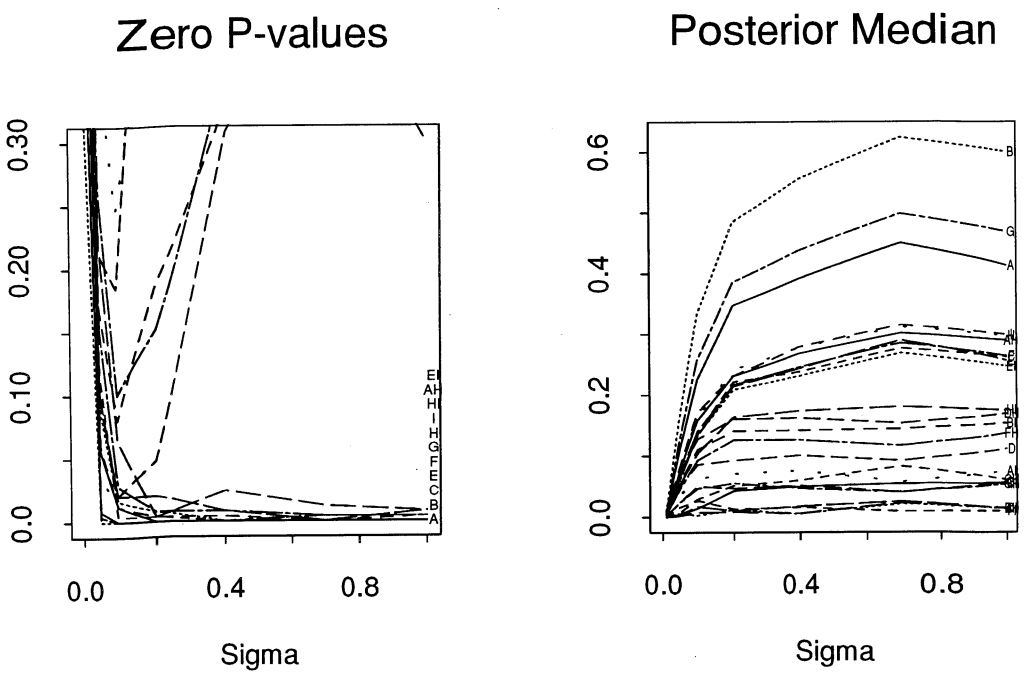


Figure 4: Posterior zero p-values and medians for the foam experiment

| Factor | Zero P-value |
|---|---|
| A | 0.00 |
| B | 0.00 |
| G | 0.00 |
| H | 0.00 |
| AH | 0.00 |
| C | 0.00 |
| F | 0.00 |
| I | 0.00 |
| EI | 0.00 |
| HI | 0.01 |
| E | 0.01 |
| DH | 0.30 |
| ⋮ | ⋮ |
| FI | 0.92 |

Table 4: Zero p-values for the full model in the foam experiment

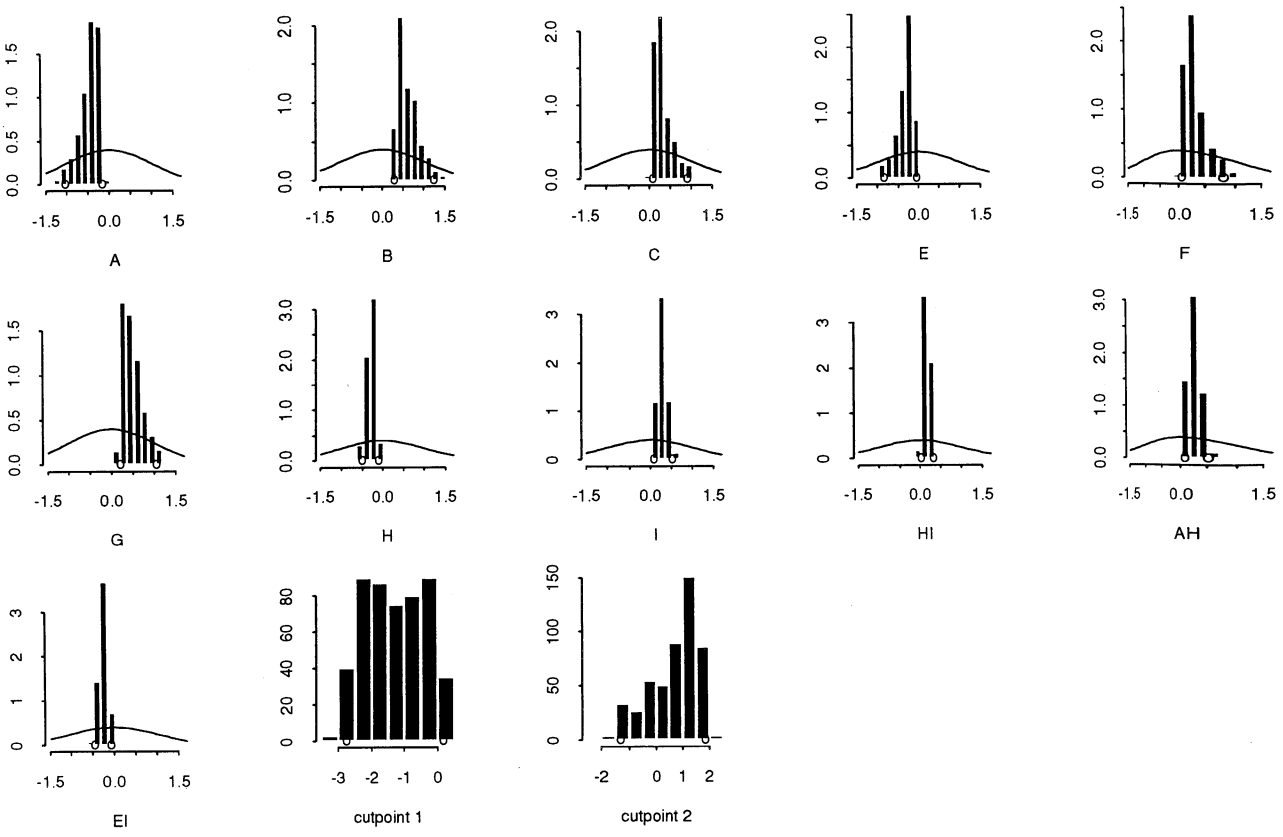


Figure 5: Posterior distributions and superimposed priors with $\sigma_\beta = 1$ for final model of the **foam** experiment

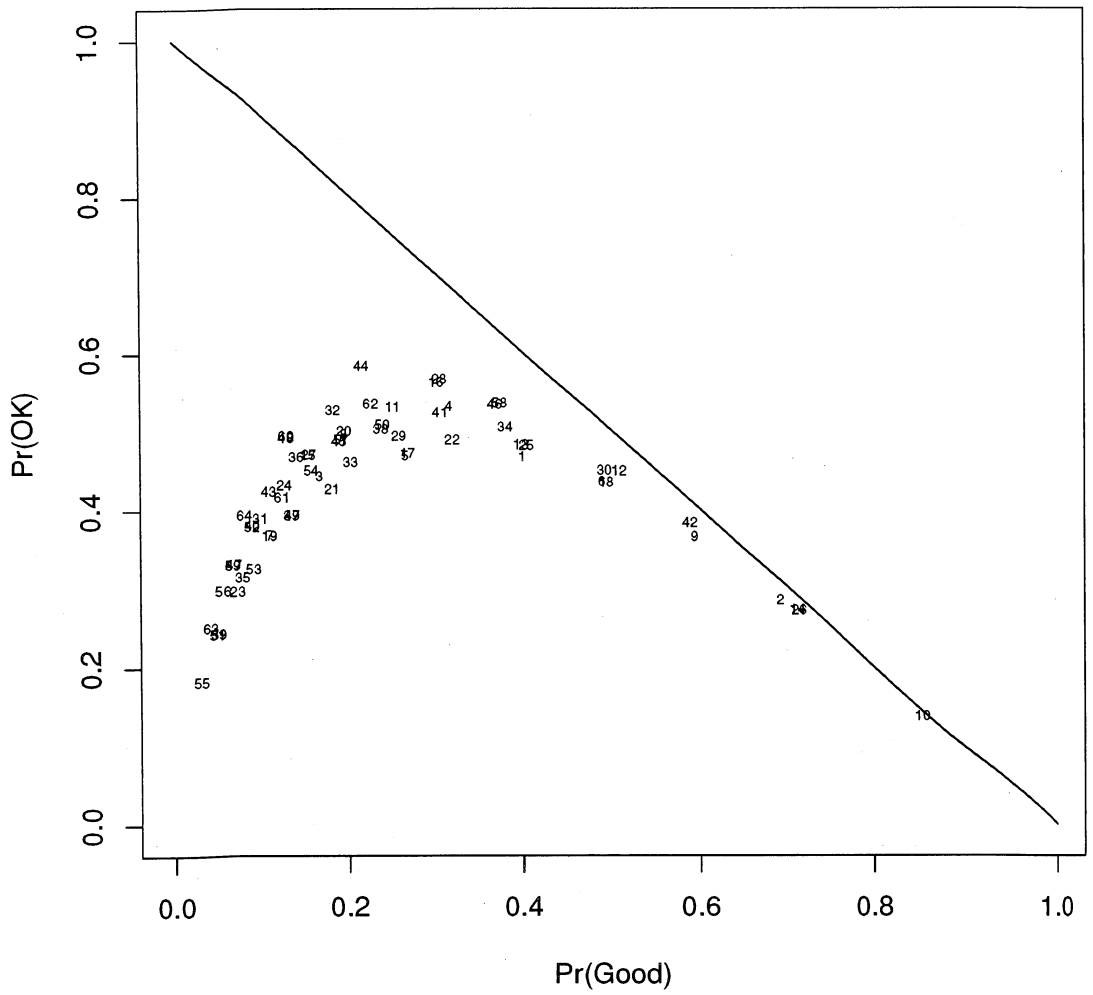


Figure 6: Average proportions at 64 factor settings for the foam experiment

| | | | | | | | Probabilities | | | Scores | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Setting | A | B | C | E | F | G | Good | OK | Poor | (3,2,1) | (3,2,0) | (3,3,1) |
| 6 | 1 | -1 | 1 | -1 | -1 | -1 | 0.49 | 0.44 | 0.08 | 2.41 | 2.33 | 2.85 |
| 18 | 1 | -1 | -1 | -1 | 1 | -1 | 0.49 | 0.44 | 0.07 | 2.42 | 2.35 | 2.86 |
| 30 | 1 | -1 | 1 | 1 | 1 | -1 | 0.49 | 0.45 | 0.06 | 2.43 | 2.37 | 2.88 |
| 12 | 1 | 1 | -1 | 1 | -1 | -1 | 0.51 | 0.45 | 0.04 | 2.47 | 2.42 | 2.91 |
| 9 | -1 | -1 | -1 | 1 | -1 | -1 | 0.59 | 0.37 | 0.04 | 2.55 | 2.51 | 2.92 |
| 42 | 1 | -1 | -1 | 1 | -1 | 1 | 0.59 | 0.38 | 0.03 | 2.56 | 2.53 | 2.94 |
| 2 | 1 | -1 | -1 | -1 | -1 | -1 | 0.69 | 0.29 | 0.02 | 2.67 | 2.64 | 2.95 |
| 14 | 1 | -1 | 1 | 1 | -1 | -1 | 0.71 | 0.27 | 0.02 | 2.69 | 2.68 | 2.97 |
| 26 | 1 | -1 | -1 | 1 | 1 | -1 | 0.71 | 0.27 | 0.02 | 2.70 | 2.68 | 2.97 |
| 10 | 1 | -1 | -1 | 1 | -1 | -1 | 0.86 | 0.14 | 0.01 | 2.85 | 2.84 | 2.99 |

Table 5: The best settings according to various averaged criteria for the foam experiment

If we consider applying scores to these probabilities, many of the results will agree, since most of the optimal settings found in the previous paragraph consist of very few poor parts. The performance criterion for three scorings (again averaged over posteriors and noise) are given in Table 5. The first scoring (3,2,1) reflects a linear ordering of the three categories, while the second assumes that poor parts are worse. The third scoring (3,3,1) puts good and ok parts on an equal footing, rating them higher than poor parts. By all three criteria, settings 10 is the best, and by the first two criteria, it is a clear winner. Using the (3,3,1) criterion, there are a number of other settings almost as good.

The analysis above uses only the mean levels of the cell probabilities. Although the main goal is to obtain all parts in the "very good" category, additional information about the behavior of the process may be obtained by looking at the distribution of proportions over noise factors H and I, and uncertainty in $\beta$ and $\gamma$. Instead of examining the joint posterior of two outcome proportions, we might instead consider the distribution of the univariate performance measures described in Table 5.

To illustrate this technique, consider the scoring (3,2,1). From Table 5 we identify setting 10, 26, 2, and 14 as candidates for the best setting. The distribution of the performance measure 3*Pr(very good) + 2*Pr(acceptable) + 1*Pr(poor) is displayed in Figure 7 for each of these settings. Not only does setting 10 have the best average performance measure, but the performance distribution is favorable. Run 2 has a more skewed performance distribution, and is likely to be less optimal than either settings 14 or 26, which appear to have comparable score distributions.

## 5.3 An Injection Molding Experiment

This experiment, which is analyzed by Steinberg and Bursztyn (1993), was conducted to improve the quality of injection molded plastic handles. The goal of the experiment was to produce "on target" parts, in contrast to the "higher the better" and "lower the better" experiments previously discussed. The response variable is the "amount of material", ranging from 1 (too little) to 7 (too much), with a target value of 4. No values of 7 were actually recorded from the data, so the analysis

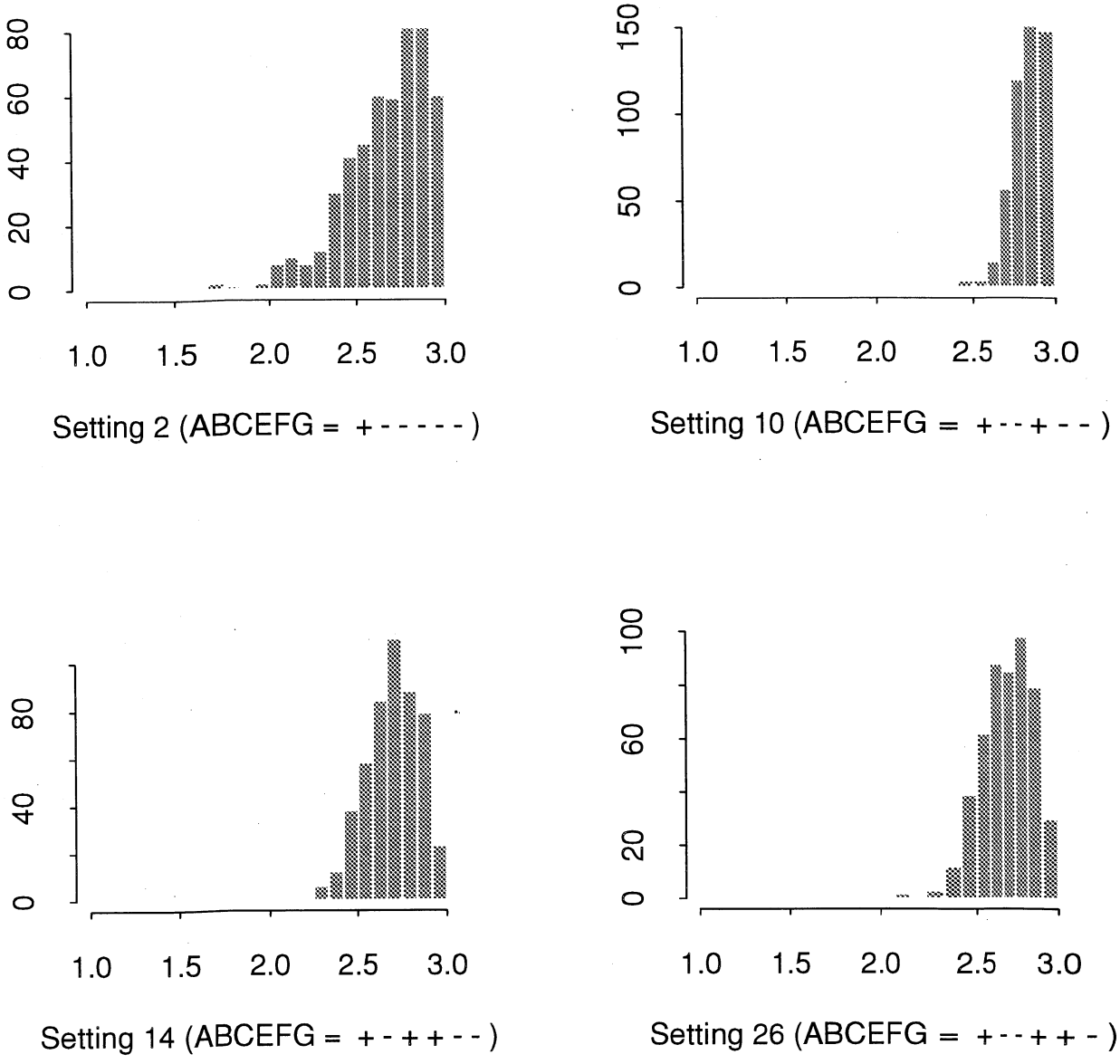


Figure 7: Distributions of scores over noise and posteriors of $\beta$ and $\gamma$ for the foam experiment

is for the six observed response categories. The original data set consisted of 17 control factors and two noise factors. For illustrative purposes, we analyze a quarter fraction of the experiment with no noise effects (the noise factors were set as follows: mold = 2, water temperature = 1). The quarter fraction is a $2^{(17-12)}$ fractional factorial design, with 32 observations. The factors are labeled A-Q, and are grouped as follows: A-D are process temperatures, E-G stroke counts, H-L pressures, and N-Q are process times.

The model fitting strategy was the same as that used in the previous examples; a brief summary is given since we wish to focus on optimization. A saturated model with 31 effects was first considered, and all insignificant terms dropped in groups. The final model chosen is the one with main effects C, E, F, I, L, O, and a FL interaction; the main effect of factor F was not significant, but was included to simplify interpretation.

It is the optimization stage of this example that is interesting, since the on-target problem presents challenges not present in the other data sets. In "higher the better" experiments it is often enough to determine the direction of each effect, and then set that factor to a corresponding high or low level, so as to push the probability mass to the high end of the scale. In an on-target problem, things are not that simple. Since the target is a middle category, the setting of each factor will depend on the settings of the others. For example, if one effect is positive, and the other negative, we might wish to cancel them out to get on target. There are two ways to do this (if we consider settings of only $\pm 1$ levels of factors), and as the number of terms in the model increases, there are many different ways to accomplish similar goals. Thus, for a single on-target goal, we are likely to have several "optimal" factor settings. This is in fact an advantage, since we may either choose the cheapest settings, or use secondary criteria to distinguish between the settings.

These concepts are more clearly illustrated in the optimization stage for this experiment. Since we have little knowledge of the process involved, we consider optimizing over high and low levels ($\pm 1$) of the factors. This means that we have $2^6 = 64$ settings to consider. All optimization is relative to the posterior distributions for the proportions of parts in each response category. We seek factor settings that produce on-target posteriors, namely ones with most parts in response group four. While there may be other favorable properties, we assume that it is crucial that settings produce a high proportion of on-target parts. Thus to identify promising settings, we consider the average posterior proportion of parts in response category four. The ten most promising settings are displayed in Table 6. There are seven settings which produce virtually the same average proportion; we shall concentrate on these seven in the following.

If our only goal is to have as many parts on target as possible then we may choose between these settings on a cost basis. In most situations, some settings will be less expensive. If we have other priorities, we may use them to choose between the seven settings. Possible secondary criteria might include:

- **Low variability in the proportion of on-target parts.** We may examine histograms of the posterior proportions of on-target parts, as shown in Figure 8. In this case, there are only minor differences between the distributions.

- **Other outcome groups.** Slight underfills or overfills (categories 3 and 5) may have a smaller loss associated with them than more extreme outcomes. Hence we may wish to examine the proportions of parts in all six categories. We can look either at the mean proportions

| Setting | C | E | F | I | L | O | FL | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 61 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | 0.00 | 0.02 | 0.05 | 0.82 | 0.11 | 0.00 |
| 36 | 1 | -1 | -1 | -1 | 1 | 1 | -1 | 0.00 | 0.01 | 0.03 | 0.81 | 0.14 | 0.00 |
| 1 | -1 | -1 | -1 | -1 | -1 | -1 | 1 | 0.00 | 0.03 | 0.07 | 0.81 | 0.09 | 0.00 |
| 34 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | 0.01 | 0.05 | 0.10 | 0.80 | 0.04 | 0.00 |
| 20 | -1 | 1 | -1 | -1 | 1 | 1 | -1 | 0.00 | 0.01 | 0.04 | 0.80 | 0.14 | 0.00 |
| 18 | -1 | 1 | -1 | -1 | -1 | 1 | 1 | 0.01 | 0.05 | 0.11 | 0.80 | 0.04 | 0.00 |
| 6 | -1 | -1 | -1 | 1 | -1 | 1 | 1 | 0.00 | 0.03 | 0.08 | 0.80 | 0.09 | 0.00 |
| 50 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 0.00 | 0.00 | 0.01 | 0.72 | 0.24 | 0.02 |
| 8 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | 0.00 | 0.01 | 0.03 | 0.72 | 0.24 | 0.00 |
| 3 | -1 | -1 | -1 | -1 | 1 | -1 | -1 | 0.00 | 0.01 | 0.03 | 0.71 | 0.25 | 0.00 |

Table 6: The best settings according to average on-target proportions for the injection molding experiment

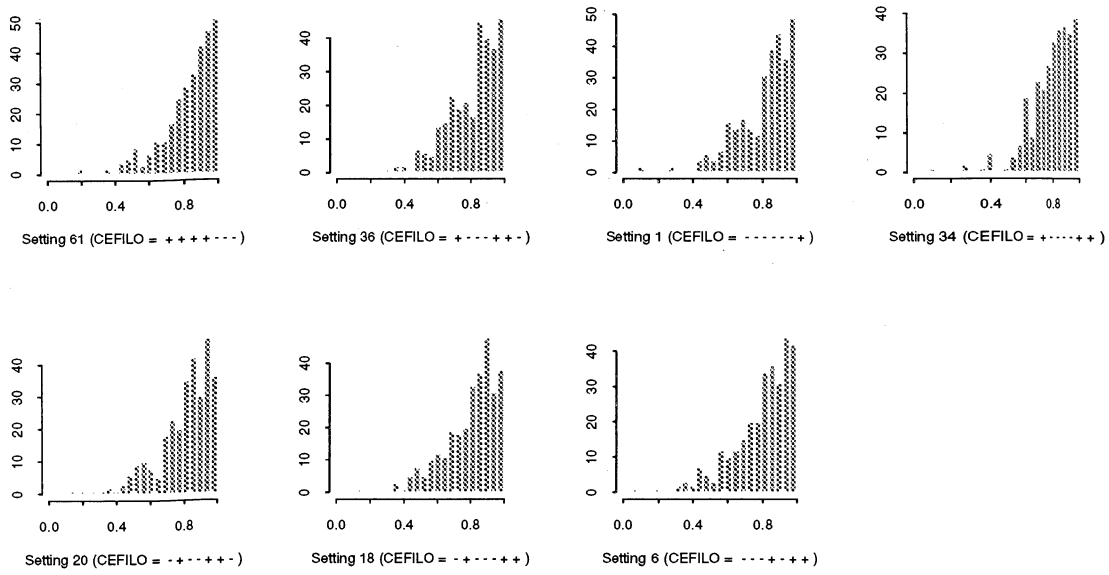


Figure 8: Posterior proportions of on-target parts for seven best settings in the injection molding experiment

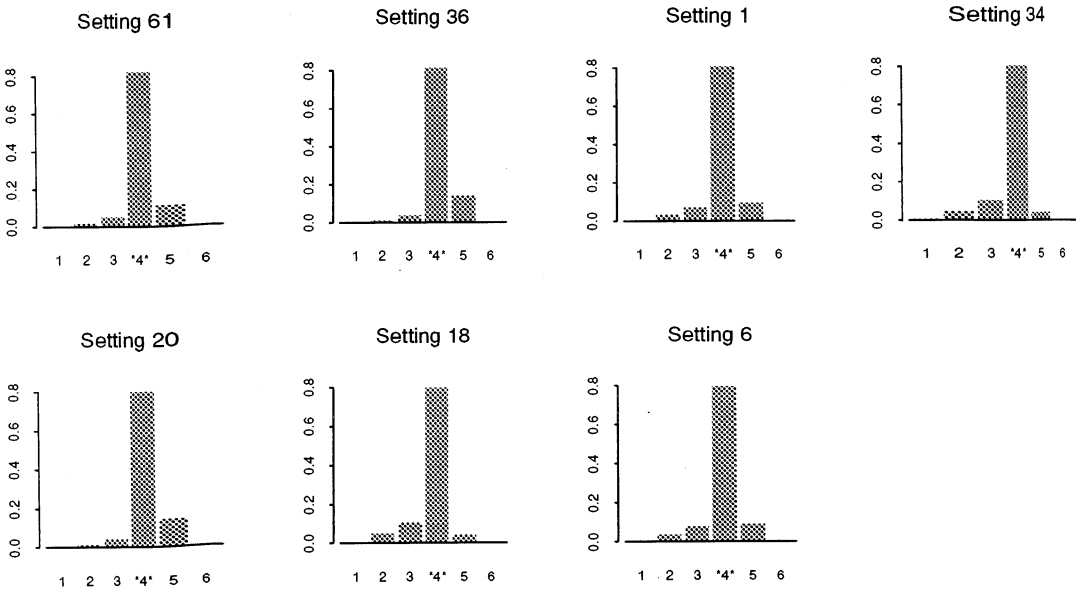


Figure 9: Average posterior proportions for all six response categories at various factor settings for the injection molding experiment

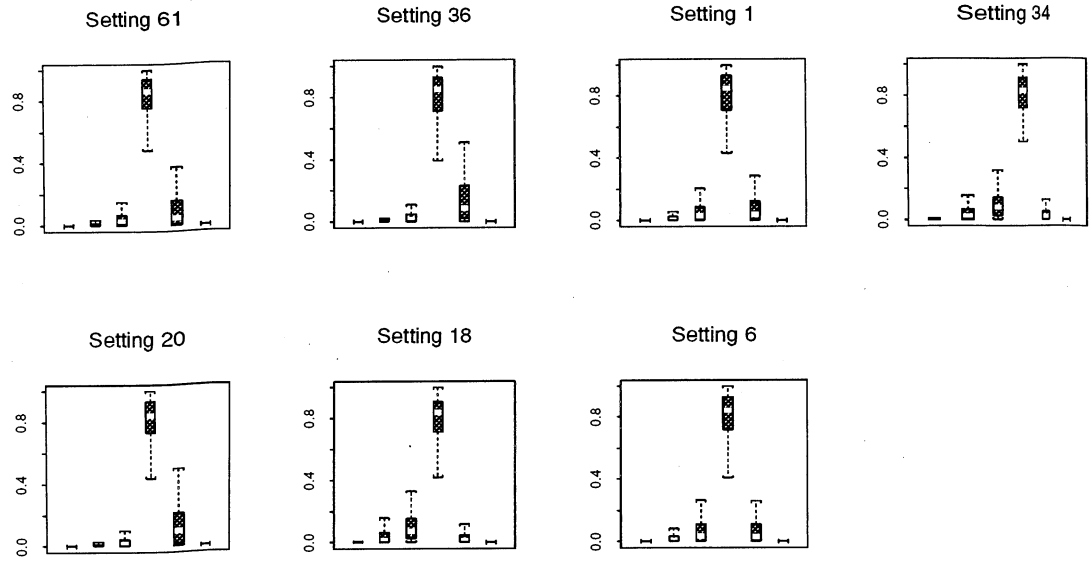


Figure 10: Average posterior proportions for all six response categories at various factor settings for the injection molding experiment

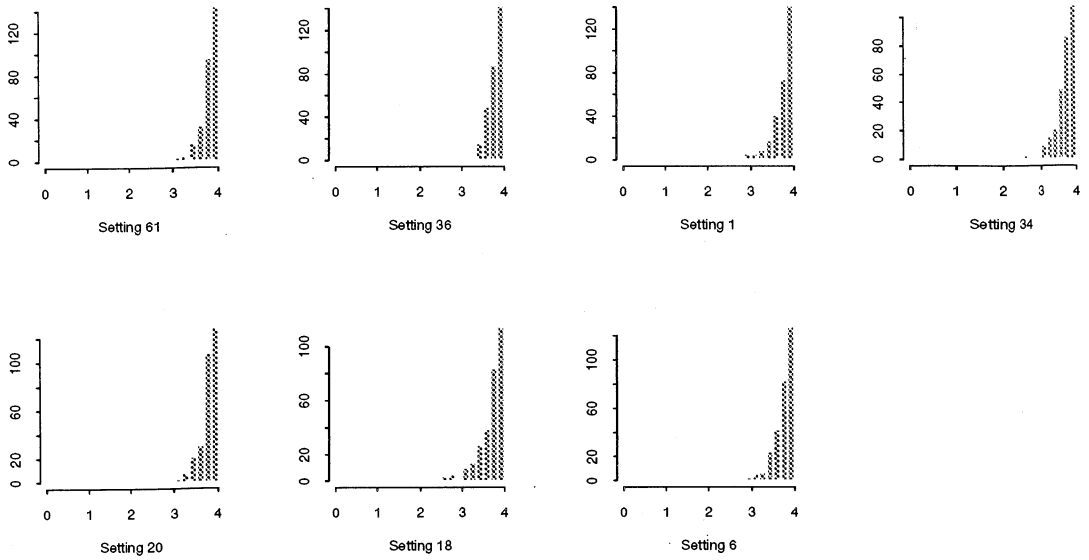


Figure 11: On/Near-target score for seven factor settings in the injection molding experiment

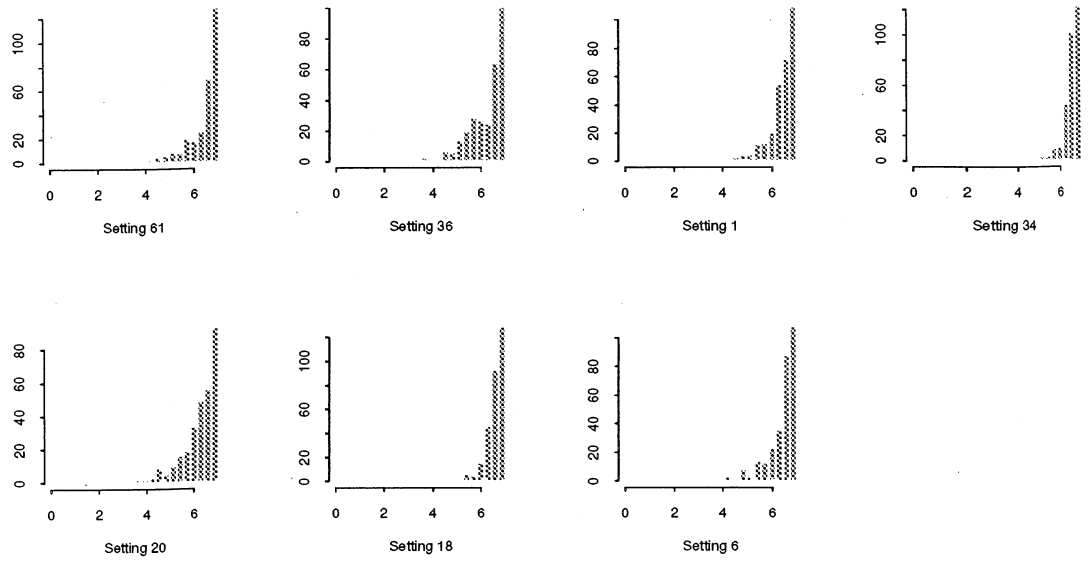


Figure 12: Acceptable underfills score for seven factor settings in the injection molding experiment

(represented by barplots in Figure 9), or representations of the posteriors for each outcome group, as represented in the boxplots in Figure 10. In these plots, we see that settings 1, 34, 18, and 6 produce more outcomes outside the range (3,5). If only parts outside this range must be scrapped, these four settings would be less preferable.

- **A specific criterion based on scoring.** We may wish to summarize several requirements with a single scoring. For example, if we want as many parts on target as possible, and parts just off target remain acceptable, we could use the scores (0,1,3,4,3,1). The distribution of this score calculated from the posteriors is displayed in Figure 11. We see that there are only minor differences between these scores. On the other hand, slight underfills might be acceptable, but overfills less useful. We might then choose the scoring (2,5,6,7,2,0) to reflect this. The posteriors for this score are given in Figure 12. These scores produce distributions that are more distinct. Figure 9 suggested this, since it shows that some settings produce more underfills and some more overfills. On the basis of Figure 12, we might select settings 34 and 18 as most favorable.

We will not try to choose the overall best one or two settings from these criteria, since we have no knowledge about the real goals of the experiment, other than to produce on-target parts. In conclusion, note that for the on-target problem, we are likely to get several settings that produce comparable results. This may be seen as an advantage, since we may either choose the cheapest settings if we have no secondary goals, or if there are secondary goals, use them to identify the best factor level settings.

# 6 Summary and Conclusions

For one reason or another, most of the current methods for analyzing ordered categorical data are problematic. The Bayesian approach outlined in this paper is an attempt to overcome these difficulties. By combining a Generalized Linear Model with Bayesian estimation techniques, we arrive at a model that appears appropriate for the data at hand. While the model itself has been considered previously, the area of application suggests new uses for the model, and advantages therein.

One of the most significant advantages is the ability of Bayesian models to account for uncertainty in the estimation of the parameters. By acknowledging the uncertainty rather than using point estimates, we will hopefully get a more honest choice of optimal factor settings. The variability in the parameter estimates fits in nicely in situations such as robust design experiments, where we assume additional variation in the process is induced by uncontrollable noise factors. Our approach provides an unified way of looking at these two sources of variability.

Other advantages of this approach include numerous graphical techniques for both model selection and process optimization, methods for checking the sensitivity of the results to the prior assumptions, and more meaningful conclusions, namely, an estimate of the proportion of parts in each category for a given factor level setting.

Of special interest is the on-target problem, since many different settings often produce similar results. In such situations, we have the option of either choosing the least expensive factor settings,

or using other model-based criteria for distinguishing between competing settings. The number of possibilities is thereby reduced, which then need to be confirmed by making a few additional runs.

While this approach deals with many difficulties previously encountered, the model considered is only one of many possible models that might be fit to such data. In some situations, it might be interesting to consider other Bayesian models for describing ordered categorical data, as well as adapting this model for a number of different experimental situations involving this type of data.

## Acknowledgments

This research was supported by General Motors of Canada Limited, the Manufacturing Research Corporation of Ontario, and the Natural Sciences and Engineering Research Council of Canada. The authors thank C. F. J. Wu, J. Lawless, and G. Bennett for their insightful comments.

## References

Albert, J. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Bostelman, M. A., Buck, D. K. and Henry, J. E. (1987), "Optimization of Design and Process Parameters for Piston Bores in Grey Iron Brake Housings," In *Fifth Symposium on Taguchi Methods*, American Suppliers Institute, pp 43–66. .

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Hamada, M. and Wu, C. F. J. (1990), "A Critical Look at Accumulation Analysis and Related Methods," *Technometrics*, 32, 119–162.

Hamada, M. and Wu, C. F. J. (1992), "Analysis of Censored Data from Fractionated Experiments: a Bayesian Approach," The Institute for Improvement in Quality and Productivity Research Report RR-92-11, University of Waterloo.

Jinks, J. (1987), "Reduction of Voids in a Urethane-Foam Product," In *Fifth Symposium on Taguchi Methods*, American Suppliers Institute, 135–148.

Lawless, J. F. and Singhal, K. (1978), "Efficient Screening of Nonnormal Regression Models," *Biometrics*, 34, 318–327.

McCullagh, P. (1980), "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society, Series B*, 42, 109–142.

Nair, V. N. (1986), "Testing in Industrial Experiments with Ordered Categorical Data," *Technometrics*, 28, 283–291.

Shoemaker, A. C., Tsui, K. L. and Wu, C. F. J. (1991), "Economical Experimentation Methods for Robust Design," *Technometrics*, 33, 415–427.

Steinberg, D. M., and Bursztyn, D. (1993), "Confounded Dispersion Effects in Robust Design Experiments with Noise Factors," Center for Quality and Productivity Improvement Report 93, University of Wisconsin-Madison.

Taguchi, G. (1974), "A New Statistical Analysis for Clinical Data, the Accumulating Analysis, in Contrast with the Chi-Square Test," *Saishin Igaku*, 29, 806–813.

Tanner, M. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–550.

Tse, S. K. (1986), "On the Existence and Uniqueness of Maximum Likelihood Estimates in Polytomous Response Models," *Journal of Statistical Planning and Inference*, 14, 269–273.

Welch, W. J., Yu, T. K., Kang, S. M. and Sacks, J. (1990), "Computer Experiments for Quality Control by Parameter Design," *Journal of Quality Technology*, 22, 15–22.