BAYESIAN CASE STUDIES IN NONPARAMETRICS

by

Michael J. Symons

Department of Biostatistics, University of
North Carolina at Chapel Hill, NC.

Institute of Mimeo Series No. 1886
April 1991

Bayesian Case Studies in Nonparametrics

Michael J. Symons

Biostatistics, CB# 7400
SPH, UNC-CH
Chapel Hill, NC 27599-7400

Abstract

Elements of Bayesian nonparametric statistical thought are explored in a series of case studies. Interpretation of a measurement as continuous, ordered, polychotomous, or dichotomous provides a framework in which examples are presented. Bayesian analogues to frequentist nonparametrics and overt Bayesian techniques are employed. Examples included are as follows: (1) averaging over families of distributions, (2) estimation of a single distribution function, (3) comparing several distribution functions, (4) estimating the coefficient of a concomitant variable affecting a distribution function, (5) monitoring compliance with a dichotomous measurement, and (6) using the multinomial for a categorization of any measurement's range. Lindley (1972, §12.2) provides an intitial sketch. Hill's (1968) nonparametric Bayesian construct and Berliner and Hill's (1988) application to survival are also reviewed.

A commonality in the mechanics of these examples is the calculation of a marginal distribution over model parameters. Many are predictive distributions, resulting from an average over a likelihood and vague prior, and leaving observables for the calculations, as described by Roberts (1965) and advocated by Geisser (1971). Other specific observations from these efforts include the following points: (1) nonparametrics is typified by a proliferation of parameters; (2) biased estimation may provide smaller mean squared errors; (3) point null hypotheses are intrinsically unlikely; (4) the practical aspects of conditional inference with proportional hazards are embraced by Bayesian considerations; and (5) Fisher's Exact Test is a predictive distribution.

Introduction

Continuous, ordered, polychotomous, and dichotomous are four categories of measurement scale which provide a practical organization of the following case studies. Also, many measurements are continuous, but a strategy of classical nonparametrics is to view the observations in a less informative state, for example as ranked or divided into two or more categories, and thereby restricting the likelihood models appropriate for the data. Since the likelihood is the cornerstone of Bayesian inference, the formulation of these models, as they reflect limitations of the available data, is a point of focus. But, the handling of the parameters in the models illustrate Bayesian nonparametric methodology.

Continuous Measurements

A basic concern is that any parametric model, for example a normal distribution, is more of an approximation to, rather than being the exact representation of, the true distribution of the continuous measurement X. One approach then is to soften the specification of the distribution of X on the real line, say. Lindley (1972a, 66ff) views this as an impasse in its most general formulation, namely viewing the distribution of X over all possible continuous distributions. But a stride in this direction is possible by compounding a chosen distribution with a suitable weight function over the range of one or more of the chosen distribution's parameters. The result is a generalized distribution which includes the initial distributional choice as a limiting case.

Case Study #1: Compounding to Generalized Distributions

A series of specific instances illustrate the process:

(i) Negative Binomial: Poisson and Gamma Mixture

A textbook example of compounding that allows the intensity of the Poisson to vary over the population, see Greenwood and Yule (1920). A result is extra-variation in the negative binomial relative to the Poisson.

(ii) Pareto: Exponential and Gamma Mixture

Turnbull, Brown and Hu (1974) relax the constant hazard assumption of the exponential distribution by compounding with a gamma distribution. The one parameter exponential becomes a two parameter Pareto.

(iii) Student's t: Predictive Distribution for a Normal

Geisser (1971, p. 465ff) calculates the expected value of a normal density over the posterior for the mean and variance of the normal based upon a sample of size n with a Jeffreys (1961) vague

prior on the mean and variance. The two parameter normal becomes a three parameter Student's t: location, scale and degrees of freedom. The symmetry of the normal is retained, however.

(iv) Neyman Type B Distribution

Denote by A a small portion of a field F. Let the probability generating function (pgf) of the number of larvae found in A out of N=n hatchlings from a single egg mass in F have a binomial pgf, $(1-p+pz)^n$, where p is the probability of each larvae being in A. Then the pgf of the number of larvae observed in A from a randomly selected egg mass in F is

$$h(z) = \int_0^1 \int_0^\infty (1-p+pz)^n dF_N(n) dF_P(p) .$$

Suppose N is Poisson distributed with mean θ and a uniform distribution is appropriate for p. Then

$$h(z) = \int_0^1 \sum_{n=0}^{\infty} e^{-\theta}[\theta(1+p(z-1))]^n \Big/ n!\, dp = e^{-\theta} \int_0^1 e^{+\theta[1+p(z-1)]}\, dp$$

$$= \int_0^1 e^{\theta(z-1)p}\, dp = \frac{1}{\theta(z-1)}[e^{\theta(z-1)}-1].$$

Now if there are M=m egg masses in F, of which fraction π are in A, the number of larvae in A from m randomly selected egg masses has pgf $(1-\pi+\pi h(z))^m$. Letting $m\to\infty$ and $\pi\to 0$ while holding $m\pi=\lambda$, yields a generalized Poisson pgf, $\exp\{\lambda(h(z)-1)\}$. Substituting for h(z) from above yields the pgf of the Neyman Type B distribution:

$$G_B(z) = \exp\left\{\lambda\left[\frac{e^{\theta(z-1)}-1}{\theta(z-1)} - 1\right]\right\}.$$

See Feller (1943) for a comprehensive exposition on several related compound and generalized distributions.

These examples of compounding have nonparametric appeal in that the resulting distributions are often called generalized. That is, the original distribution whose parameters were averaged over can be identified in the parameter space of the resulting compound, usually in the limit of one parameter. But the compound has more parameters than the original model. These additional parameters are an inheritance from the mixing distribution(s). The process does yield more flexible distributional shapes, but can be viewed as more parametric than nonparametric. Statistical estimation of model parameters becomes more complicated, seemingly the expense of a more flexible model. Also, recognize that averaging over vague, unnormalized weight functions on the parameters of the distribution may not

yield a density as the compound.

In summary, advocating generalized distributions as models from a frequentist's perspective is justified by the process of compounding. For a Bayesian, the averaging process in compounding is mechanically the same as that of calculating a marginal distribution, or a predictive distribution as per Roberts (1965) and Geisser (1971, 1985). A variety of generalized distributions have been studied, including the Burr by Rodriquez (1982) and the generalized F by Kalbfleisch and Prentice (1980).

Ordered Measurements

Although a continuous measurement may be made, for example the time of failure in a follow-up study, nonparametric preference may be to retain only the ordered observations, i.e., their relative magnitudes, and to discard their absolute magnitudes. A key feature of the Bayesian approach in the next series of case studies is in the formulation of the likelihood of the data.

Case Study #2: Estimation of a Single Distribution Function

Censoring of observations is an inherent feature of survival applications with time to event measurements. Suppose $t_{(1)}, t_{(2)}, \ldots, t_{(j)}, \ldots, t_{(k)}$ are the ordered failure times corresponding to a random sample of n individuals placed on test, of which k ($\leq$ n) are observed to fail. Each individual has an observation t_i and indicator δ_i=1 or 0 as they failed at t_i or not. The special case where all individuals "fail" includes situations where the measurements are not time related, such as a body chemistry determination, and this approach of modelling the hazard becomes an alternative to density estimation. Let $d_1, d_2, \ldots, d_j, \ldots, d_k$ be the number of failures at $t_{(j)}$, j = 1, 2, .., k, respectively, and noting $\sum_{j=1}^{k} d_j = m \leq n$.

The discrete hazard has been proposed as an innocuous statistical model; see Cox (1972). It is a data based hazard model, postulating a positive spike λ_j in the hazard function at the observed failure time $t_{(j)}$. Or, the hazard $\lambda(t)=\lambda_j$ at $t_{(j)}$, and is zero otherwise, which corresponds to one parameter per distinct failure time, as shown in Figure 1. A likelihood for these parameters, $\underset{\sim}{\lambda}'=(\lambda_1, \lambda_2, \ldots, \lambda_j, \ldots, \lambda_k)$, of the hazard is

$$L(\underset{\sim}{\lambda}) = \prod_{i=1}^{n} \exp\left\{-\int_0^{t_i} \lambda(u)du\right\}\left[\lambda(t_i)\right]^{\delta_i}, \qquad (1)$$

since the n individuals from a random sample are independent of one another. Interpreting the integral in the argument of the exponential function in (1) as a Lebesque-Stieltjes integral yields

$$\int_0^{t_i} \lambda(u)du \equiv \int_0^{t_i} d\Lambda(u) = \sum_{j:t_i \geq t_{(j)}} \lambda_j , \tag{2}$$

with this discrete hazard model. The likelihood can be rewritten as follows:

$$L(\underline{\lambda}) = \prod_{j=1}^{k} e^{-R_j\lambda_j} \lambda_j^{d_j} , \tag{3}$$

where R_j is the number in the risk set, $\mathbf{R}_j$, for the jth hazard spike λ_j. Explicitly, $\mathbf{R}_j$ is the set of individuals with observation t_i equal to or exceeding the jth failure time $t_{(j)}$.

Estimation approaches follow directly from the likelihood. Now the log-likelihood from (3) is

$$\ln L(\underline{\lambda}) = \sum_{j=1}^{k} \left\{ -R_j\lambda_j + d_j \ln\lambda_j \right\} . \tag{4}$$

Usual procedures for maximum likelihood provide

$$\hat{\lambda}_j = d_j \Big/ R_j \tag{5}$$

and a variance of

$$\mathcal{V}(\hat{\lambda}_j) = \lambda_j^2/d_j . \tag{6}$$

Since the survival function is

$$S(t) = \exp\left\{ -\sum_{j:t \geq t_{(j)}} \lambda_j \right\} , \tag{7}$$

its maximum likelihood estimator (MLE) is obtained by replacing λ_j by its MLE, yielding

$$\hat{S}(t) = \exp\left\{ -\sum_{j:t \geq t_{(j)}} d_j/R_j \right\} = \prod_{j:t \geq t_{(j)}} e^{-d_j/R_j} \tag{8}$$

This is Nelson's (1972) estimator, especially with unique failures, i.e., $d_j \equiv 1$ for all $j = 1, 2, \ldots, k$. The variance of this estimator can be approximated by a linearized Taylor series approach, yielding

$$\hat{\mathrm{V}}(\hat{S}(t)) \doteq \left[\hat{S}(t)\right]^2 \sum_{j:t \geq t_{(j)}} d_j \Big/ R_j^2 . \tag{9}$$

A Bayesian approach focuses on the posterior distribution of $\underset{\sim}{\lambda}$, obtained as the normalized product of the likelihood and prior distribution of $\underset{\sim}{\lambda}$. As separate positive jumps, λ_j, in the cumulative hazard are provided by the discrete hazard model, a vague Jeffreys (1961) prior on each λ_j is taken:

$$p(\underset{\sim}{\lambda}) = \prod_{j=1}^{k} \lambda_j^{-1} . \tag{10}$$

Then the posterior for $\underset{\sim}{\lambda}$ is

$$p(\underset{\sim}{\lambda} \mid \underset{\sim}{d}, \underset{\sim}{R}) \propto \prod_{j=1}^{k} \left[\lambda_j^{d_j} \exp\left\{-R_j \lambda_j\right\} \right] \lambda_j^{-1} \tag{11}$$

and takes the product form

$$p(\underset{\sim}{\lambda} \mid \underset{\sim}{d}, \underset{\sim}{R}) = \prod_{j=1}^{k} p(\lambda_j \mid d_j, R_j) . \tag{12}$$

By inspection, each λ_j has an independent gamma posterior, specifically,

$$p(\lambda_j \mid d_j, R_j) = \frac{R_j^{d_j}}{\Gamma(d_j)} \lambda_j^{d_j - 1} \exp\left\{-R_j \lambda_j\right\}, \tag{13}$$

and reduces to an exponential when the failure at $t_{(j)}$ is unique, i.e., $d_j \equiv 1$.

With a squared error loss function, the posterior mean of the survival function (7) is the optimal Bayesian estimator. Direct computation and reduction yields

$$\mathcal{E}(S(t)) = \int_0^\infty \cdots \int_0^\infty \exp\left\{ -\sum_{j:t \geq t_j} \lambda_j \right\} \prod_{j=1}^{k} p(\lambda_j \mid d_j, R_j) d\underset{\sim}{\lambda} = \prod_{j:t \geq t_{(j)}} \left[\frac{R_j}{R_j + 1} \right]^{d_j} = S_B(t) \tag{14}$$

The variance of the probability of surviving to time t is with respect to the posterior distribution. Calculation by the computing form for variance yields

$$\mathcal{V}(S(t)) = \prod_{j:t \geq t_{(j)}} \left[\frac{R_j}{R_j+2}\right]^{d_j} - \left\{\prod_{j:t \geq t_{(j)}} \left[\frac{R_j}{R_j+1}\right]^{d_j}\right\}^2 . \tag{15}$$

The posterior variance reflects the uncertainty contained in the posterior distribution of $\underset{\sim}{\lambda}$, conditional on the multiplicity of failures, d_j, and the number R_j in each risk set, $\mathbf{R}_j$.

The product-limit form of (14) can also be viewed as a predictive survival distribution following Roberts (1965) and Geisser (1971). And, calculating the predictive distribution at $t_{(i+1)}$ generalizes Berliner and Hill's (1988) Theorem 2.3, allowing the multiplicity, d_j, of failures at times $t_{(j)}$ within the structure of the discrete hazard model. Berliner and Hill (1988) add a linear smoothing to their Theorem 2.3 estimator for applications. A piecewise exponential model provides similar smoothing and is detailed by Symons and Yuan (1987).

The Kaplan-Meier (1958) estimator follows from a binomial perspective on these data; see Cox and Oakes (1984, p. 48ff). Let λ_j be the probability of failure at $t_{(j)}$, i.e., the hazard for this discrete formulation. For R_j individuals at risk of failure at $t_{(j)}$, then

$$L(\underset{\sim}{\lambda}) = \prod_{j=1}^{k} \left\{\lambda_j^{d_j}(1-\lambda_j)^{R_j-d_j}\right\}, \tag{16}$$

conditioning on the R_j. From the log-likelihood, it follows that the maximum likelihood estimators, $\hat{\lambda}_j$, are the same as those in (5) with the survival likelihood, but the variances are

$$\mathcal{V}(\hat{\lambda}_j) = \lambda_j(1-\lambda_j) \Big/ R_j . \tag{17}$$

The survival probability is P[T > t], or

$$S(t) = \prod_{j:t \geq t_{(j)}} (1-\lambda_j) , \tag{18}$$

and its MLE is obtained by replacing the λ_j by their MLEs, yielding

$$\tilde{S}(t) = \prod_{j:t \geq t_{(j)}} (1-d_j/R_j) , \tag{19}$$

the Kaplan-Meier estimator, especially when $d_j \equiv 1$ for j=1, 2, ..., k. Cox and Oakes (1984, p. 175ff)

arrive at this same estimator using a multinomial presentation of the "atoms", λ_j, and self-consistency of the EM algorithm.

Using the linearized Taylor series approach, the variance is approximated by

$$\hat{V}(\tilde{S}(t)) \doteq [\tilde{S}(t)]^2 \sum_{j:t \geq t_{(j)}} \frac{d_j}{(R_j - d_j)R_j}, \tag{20}$$

recognized as Greenwood's Formula. Notice that the variance of the Kaplan-Meier and Nelson estimators condition upon the R_j, but treat d_j as random variables in a repeated sampling perspective.

There is also a connection between the Kaplan-Meier estimator and this Bayesian approach with the discrete hazard model. When all the failures are distinct, i.e., $d_j \equiv 1$ for all k failure times, and with a prior $\exp(\lambda_j)/\lambda_j$ on each spike in the hazard, the Bayes calculation in (14) yields the Kaplan-Meier estimator. This prior disproportionately favors ever larger values of λ_j and violates the common correspondence between Bayesian results with vague priors and standard frequentist procedures. However, in large samples the limit of the Kaplan-Meier estimator (19) and the Bayes estimator (14) are identical, so that comparative performance with only small and moderate sample sizes is of interest.

Another presentation of a Bayesian approach is to focus on the "atoms", λ_j, of the discrete hazard as being of the survival distribution, rather than of the hazard function. Hence, Cox and Oakes (1984, p. 57ff) present a Bayesian approach to nonparametric estimation building upon likelihood (16), rather than upon likelihood (1). Their approach leads naturally to a product of independent beta distributions as the posterior rather than the product of independent gamma distributions given in equation (12). However, their posterior distribution of the survival function does not take a simple form like that in equation (14).

The Nelson, Bayes and Kaplan-Meier estimators of the survival function were compared by Symons and Yuan (1987). The magnitude of the estimators are functionally ordered: Bayes greater than or equal to Nelson's, which in turn is greater than or equal to the Kaplan-Meier estimator. The Bayes and Nelson estimators are generally positively biased, but have smaller mean squared errors than Kaplan-Meier in the first half of the survival function. Kaplan-Meier has a smaller mean squared error in the second half of the survival function. An integrated mean squared error comparison was not definitive, but with smaller samples and greater censoring the Nelson and Bayes estimators had smaller integrated mean squared errors than the Kaplan-Meier estimator. Nelson's estimator may be preferable to the other two, overall, in a comparison utilizing frequentist criteria. It also is embraced by statistical theory of counting processes; see Aalen (1978) for example.

Case Study #3: Comparing Two or More Distribution Functions

Utilizing only the ranks of the observations and the development of the Case Study #2, suppose

there are two independently drawn random samples: $x_1, x_2, \ldots, x_m$ from Population #1 with distribution function $G(x \mid \underset{\sim}{\lambda})$ and $y_1, y_2, \ldots, y_n$ from Population #2 with distribution $H(y \mid \underset{\sim}{\theta})$. Or schematically,

Population	Sample	Distribution Function	Parameter Space
#1	$\underset{\sim}{x}' = (x_1, x_2, \ldots x_m)$	$G(x \mid \underset{\sim}{\lambda})$	$\underset{\sim}{\lambda} \in \Lambda$
#2	$\underset{\sim}{y}' = (y_1, y_2, \ldots y_n)$	$H(y \mid \underset{\sim}{\theta})$	$\underset{\sim}{\theta} \in \Theta$

The comparison of $G(\cdot)$ and $H(\cdot)$ is of primary importance; i.e., which of two hypotheses do the data favor,

$$H_0\colon G(\cdot) = H(\cdot) \quad \text{or} \quad H_a\colon G(\cdot) \neq H(\cdot)\,?$$

The Bayesian approach to testing is to compute the posterior odds in favor of one or the other hypothesis. Importantly, two hypotheses are needed. The objective is to revise, in light of the data $\underset{\sim}{x}$ and $\underset{\sim}{y}$, the prior odds, say of H_0,

$$\Omega(H_0) = p_0/(1-p_0)\,, \tag{21}$$

where $p_0 = P[H_0]$ is assessed before observing $\underset{\sim}{x}$ or $\underset{\sim}{y}$. The posterior odds favoring H_0 follows from the posterior probability that H_0 is true by Bayes theorem:

$$P[H_0 \mid \underset{\sim}{x},\underset{\sim}{y}] = P[\underset{\sim}{x},\underset{\sim}{y} \mid H_0]\, p_0/P(\underset{\sim}{x},\underset{\sim}{y}) \tag{22}$$

where

$$P(\underset{\sim}{x},\underset{\sim}{y}) = P[\underset{\sim}{x},\underset{\sim}{y} \mid H_0]\, p_0 + P[\underset{\sim}{x},\underset{\sim}{y} \mid H_a](1-p_0)\,.$$

Likewise, the posterior probability that H_a is true is given by

$$P[H_a \mid \underset{\sim}{x},\underset{\sim}{y}] = P[\underset{\sim}{x},\underset{\sim}{y} \mid H_a](1-p_0) \Big/ P(\underset{\sim}{x},\underset{\sim}{y})\,. \tag{23}$$

Now the posterior odds favoring, say H_0, is given by

$$\Omega(H_0 \mid \underset{\sim}{x},\underset{\sim}{y}) = \frac{P[H_0 \mid \underset{\sim}{x},\underset{\sim}{y}]}{1-P[H_0 \mid \underset{\sim}{x},\underset{\sim}{y}]} = \frac{P[\underset{\sim}{x},\underset{\sim}{y} \mid H_0]p_0}{P[\underset{\sim}{x},\underset{\sim}{y} \mid H_a](1-p_0)} = \frac{P[\underset{\sim}{x},\underset{\sim}{y} \mid H_0]}{P[\underset{\sim}{x},\underset{\sim}{y} \mid H_a]}\,\Omega(H_0) \tag{24}$$

Or, modifying the prior odds by a likelihood ratio factor produces the posterior odds.

Examining this likelihood ratio factor more closely, the numerator is the probability of the data given H_0 and is computed as the marginal probability of the data:

$$P[\underset{\sim}{x}, \underset{\sim}{y} \mid H_0] = \int_{\Lambda} P[\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{\lambda} \mid H_0]d\underset{\sim}{\lambda} = \int_{\Lambda} P[\underset{\sim}{x},\underset{\sim}{y} \mid \underset{\sim}{\lambda}, H_0]P[\underset{\sim}{\lambda} \mid H_0]d\underset{\sim}{\lambda}\,, \tag{25}$$

where $P(\underset{\sim}{x},\underset{\sim}{y} \mid \underset{\sim}{\lambda}, H_0)$ is the likelihood of the combined observations $\underset{\sim}{x}$ and $\underset{\sim}{y}$ from one distribution function, say $G(\cdot)$. The prior expression of knowledge about the parameters $\underset{\sim}{\lambda}$ is denoted by $P(\underset{\sim}{\lambda} \mid H_0)$, but does not depend upon H_0 since only a vague expression of information about these data dependent parameters seems appropriate.

Let $\underset{\sim}{z}$ denote the combined data set of $\underset{\sim}{x}$ and $\underset{\sim}{y}$. From Case Study #2, integrating equation (11) over the hazard jumbs λ_j yields

$$P[\underset{\sim}{z} \mid H_0] = \int_0^{\infty} \cdots \int_0^{\infty} \prod_{j=1}^{k_{m+n}} \exp\left\{-R(\underset{\sim}{z})_j\lambda_j\right\}\lambda_j^{d_j} \prod_{j=1}^{k_{m+n}} \lambda_j^{-1}d\underset{\sim}{\lambda} \tag{26}$$

where $R(\underset{\sim}{z})_j$ is the size of the risk set for the combined data set of $\underset{\sim}{x}$ and $\underset{\sim}{y}$ at λ_j, the hazard spike for the jth ordered observation in $\underset{\sim}{z}$. Again the vague prior form of Jeffreys (1961) is utilized. The calculation may be rewritten as a product of integrals and yields

$$P[\underset{\sim}{z} \mid H_0] = \prod_{j=1}^{k_{m+n}} \left[\Gamma(d_j)\, R(\underset{\sim}{z})_j^{-d_j}\right], \tag{27}$$

noting that k is no larger than m+n and achieves this value when there are no censored observations among $\underset{\sim}{z}$, and no ties among its ordered values.

The denominator in (24) is the marginal likelihood of the data when the distribution functions for $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are not the same and is calculated

$$P[\underset{\sim}{x},\underset{\sim}{y} \mid H_a] = \int_{\Lambda}\int_{\Theta} P(\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{\lambda},\underset{\sim}{\theta} \mid H_a)d\underset{\sim}{\theta}\, d\underset{\sim}{\lambda} = \int_{\Lambda} P(\underset{\sim}{x} \mid \underset{\sim}{\lambda})\, p(\underset{\sim}{\lambda})\, d\underset{\sim}{\lambda} \int_{\Theta} P(\underset{\sim}{y} \mid \underset{\sim}{\theta})\, p(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}\,, \tag{28}$$

by the independence of the two random samples. Each of these factors is computed as per under H_0 when $\Lambda \equiv \Theta$, yielding

$$P[\underset{\sim}{x},\underset{\sim}{y} \mid H_a] = \prod_{j=1}^{k_m}\left[\Gamma(d_j)R(\underset{\sim}{x})_j^{-d_j}\right] \prod_{i=1}^{k_n}\left[\Gamma(d_i)R(\underset{\sim}{y})_i^{-d_i}\right]. \tag{29}$$

To illustrate, consider independent random samples of five items obtained from each of three manufacturers on the waiting time to failure of a certain type of equipment. The time to failure of each of the 15 items was recorded as follows.

Manufacturer		
A	B	C
2.3	0.8	1.7
0.5	1.7	2.6
4.3	2.8	4.9
2.3	1.2	1.2
1.7	6.3	4.5

Source: Ericson (1966).

The posterior odds favoring a single distribution function is the likelihood ratio factor times the prior odds given in (24),

$$\Omega[H_0 \mid \underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z}] = \frac{P[\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z} \mid H_0]}{P[\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z} \mid H_a]}\,\Omega[H_0] \,. \tag{30}$$

The numerator of the likelihood ratio factor is from a three sample version of (27) and likewise for the denominator by (29), giving the likelihood ratio in (30) for the above data as follows:

$$\frac{[\tfrac{1}{16}][\tfrac{1}{14}][\tfrac{1}{13}]^2 2[\tfrac{1}{11}]^3[\tfrac{1}{8}]^2[\tfrac{1}{8}][\tfrac{1}{7}][\tfrac{1}{6}][\tfrac{1}{5}][\tfrac{1}{2}][\tfrac{1}{1}]}{[\tfrac{1}{5}][\tfrac{1}{4}][\tfrac{1}{3}]^2[\tfrac{1}{1}]\ [\tfrac{1}{5}][\tfrac{1}{4}][\tfrac{1}{3}][\tfrac{1}{2}][\tfrac{1}{1}]\ [\tfrac{1}{5}][\tfrac{1}{4}][\tfrac{1}{3}][\tfrac{1}{2}][\tfrac{1}{1}]} = 0.000002381 \,.$$

Two other likelihood ratio factors seem pertinent to interpreting the above factor of 2.38×10^{-6}. The least support for H_0 is 15 unique observations, which yields a likelihood ratio of 1.32×10^{-6}. The most support for H_0 is when the 15 observations are identical, yielding 4.39×10^{-1}. As the observed ratio is less than twice the worst case and a factor of 2×10^5 away from the best ratio, the null hypothesis is not supported. A Kruskal-Wallis test, by contrast, heavily favors the null hypothesis; see Conover (1971).

The generality of $G(\cdot)$ and $H(\cdot)$ both being discrete hazard models should be noted. A statement of not equal, especially with this model, does not distinguish between the smallest failure times, and the odd numbered failure times, all being from the same distribution. Additional structure on the model seems required, since the reasonableness of the discrete hazard model as a summary for data is questionable. One might readily prefer to smooth this estimator, as per Berliner and Hill

(1988), for example, or assume a smoother model, as per Cox and Oakes (1984, p. 53ff) or Symons and Yuan (1987) with a piecewise exponential. Further smoothing, e.g., using the roughness penalties in Good and Gaskin (1971), could also be considered. Notice that the implicit structure of the Kruskal-Wallis test tends in large samples toward an assumption of normality of the ranks with the alternative hypothesis being one of difference primarily in location. Bayes handling of a transformation to ranks, and their normality, is enabled by the central limit theorem with large samples, but for small samples research on a Bayesian analogue is needed.

Case Study #4: Estimating Proportional Hazards Regression Coefficients

Using covariates, applications involving several samples or having heterogeneity in the experimental units may be addressed. This additional information may be included presuming Cox's (1972) log-linear, proportional hazard formulation given by

$$\lambda(t \mid \underset{\sim}{z}) = \lim_{\Delta t \to 0} P\left\{t \leq T < t+\Delta t \mid T \geq t, \underset{\sim}{z}\right\} \Big/ \Delta t = \lambda_0(t) \exp(\underset{\sim}{z}'\underset{\sim}{\beta}) . \tag{31}$$

The positive failure time T is for an individual with p measured covariables $\underset{\sim}{z}' = (z_1, ..., z_p)$; $\underset{\sim}{\beta}$ is an associated column of p regression coefficients; and $\lambda_0(t)$ is an arbitrary, unspecified underlying hazard function. The inclusion of covariates with the discrete hazard model as $\lambda_0(t)$ in (31) is accomplished by replacing the hazard elements, λ_j, in (2) by $\lambda_j \exp(\underset{\sim}{z}'\underset{\sim}{\beta})$. Likelihood (3) is thereby expanded to include p regression coefficients $\underset{\sim}{\beta}$, as well as the k increments $\underset{\sim}{\lambda}$ of the underlying hazard, giving

$$L(\underset{\sim}{\lambda}, \underset{\sim}{\beta}) = \prod_{j=1}^{k} [\exp\{-\lambda_j \sum_{\ell \in R_j} \exp(\underset{\sim}{z}_\ell \underset{\sim}{\beta})\} \lambda_j^{d_j} \exp\{\underset{\sim}{s}(j)\underset{\sim}{\beta}\}], \tag{32}$$

where the sum of the covariates for those d_j individuals failing at t(j) is

$$\underset{\sim}{s}(j) = \sum_{\ell \in R_j} \delta_{j\ell}\, \underset{\sim}{z}_\ell , \tag{33}$$

and the δ are indicators of failure or not at specific times for individuals in R_j.

As inference on the vector of regression coefficients is of primary importance, the marginal posterior distribution of $\underset{\sim}{\beta}$ is obtained by integrating over the k nuisance parameters $\underset{\sim}{\lambda}$ in the joint posterior of $\underset{\sim}{\lambda}$ and $\underset{\sim}{\beta}$. An initially vague joint prior on $\underset{\sim}{\lambda}$ and $\underset{\sim}{\beta}$ is specified, again following Jeffreys (1961). Since the parameter range for each regression coefficient is the real line, a prior $p(\underset{\sim}{\lambda},\underset{\sim}{\beta})$ proportional to one being uniformly vague on each $\ln \lambda_j$ and on each regression coefficient is utilized. Then as per Lindley (1972), integrating the product of the likelihood (32) and prior (10) over each λ_j yields an unnormalized marginal posterior for $\underset{\sim}{\beta}$ as follows:

$$p(\underset{\sim}{\beta} \mid \underset{\sim}{t}, \underset{\sim}{\delta}, \mathbf{R}) \propto \prod_{j=1}^{k} \left\{ \frac{\exp[\underset{\sim}{s}(j)\underset{\sim}{\beta}]}{[\sum_{\ell \in \mathbf{R}_j} \exp(\underset{\sim}{z}_\ell \underset{\sim}{\beta})]^{d_j}} \right\}. \tag{34}$$

Several observations on this result are appropriate.

(a) With distinct failures, i.e., $d_j \equiv 1$ for all k failure times, the marginal posterior distribution of $\underset{\sim}{\beta}$ yields equivalent inference as Cox's (1972) approach in the sense that the mode of (34) corresponds to the maximum likelihood estimator with Cox's (1972, 1975) conditional, or partial, likelihood. Further, variance estimation based upon an approximated quadratic shape of the logarithm of this marginal posterior, or the observed information of Cox's (1972) likelihood, is also equivalent.

(b) With multiple failures at any $t_{(j)}$, i.e., $d \geq 2$ for at least one j, Bayesian inference is equivalent in mode, and the approximate quadratic shape of the logarithm of this marginal posterior is the same as that based upon the approximate likelihood described by Breslow (1974). Therefore, this Bayesian approach suggests that Breslow's (1974) approximation to Kalbfleisch and Prentice's (1973) marginal likelihood has a statistical basis as well as its value for computational practicality. Statistical practice relies on (34), e.g., as in the SAS procedure PHGLM. The essence of Breslow's (1972, 1974) approximation was also noted by Peto (1972), but more from a view of numerically approximating a probability than from likelihood considerations.

(c) Under a quadratic loss structure, the posterior mean of $\underset{\sim}{\beta}$ is preferred to the posterior mode as a point estimator. Even with a single covariate, numerical integration is needed for the required computations. However, symmetry of the posterior could be checked by an evaluation of the third partial of the natural logarithm of the posterior at the mode. Values close to zero would suggest symmetry in the posterior distribution; this is when the posterior mode should reasonably approximate the posterior mean.

(d) The above presentation is for baseline covariates, but results with time dependent covariates are immediately available when covariables $\underset{\sim}{z}_\ell$ are replaced by $\underset{\sim}{z}_{j\ell}$, i.e., by their value at the time of the jth failure.

Categorized Measurements

The information of any continuous measurement can be restricted by dividing the measurement range into exclusive and exhaustive parts. Rationales for such a division may be to address better the uncertainty of the measurement or to relate better to the clinical interpretations of the measurement. In forming a dichotomy for example, one portion may reflect normal responses and the other disease or

abnormality.

Case Study #5: Quality Control with Binomial

Suppose a particular product is shipped in lots of size N. Let the random variable X be the number of defectives among the N items to be shipped and X_0 be that number among the N_0 items selected for inspection. Then $X=X_0+X_r$, where X_r is the number of defectives among the $N_r=N-N_0$ remaining items. Quality control guidelines are directed at keeping the number of defectives in a lot to be shipped below some small number k. When $X_0=x_0$ equals or exceeds k, the entire lot is inspected and repacked. When x_0 is less than k, a decision to inspect and repack the entire lot can be based upon the probability of the number of defectives, X_r, among the N_r not inspected, being less than $k - x_0$. This development follows.

The probability of x_0 defectives among N_0 inspected items is

$$P[X_0 = x_0 \mid N_0, p] = \binom{N_0}{x_0} p^{x_0}(1-p)^{N-x_0}, \tag{35}$$

i.e., a binomial distribution with independence and equal probability of a defective item among the N_0 inspected items. However, p is unknown so the marginal probability of $X_0=x_0$ is computed

$$P[X_0 \mid N_0] = \int_0^1 P[x_0 \mid N_0,p]\, f(p)\, dp\,. \tag{36}$$

With a vague weight function for the unknown proportion defective, e.g., $f(p) \equiv 1$, the marginal calculation reduces to

$$P[X_0 \mid N_0] = (N_0+1)^{-1}\,. \tag{37}$$

Or, with the uniform distribution on the probability of a defective, every value of $X_0(= 0, 1, 2, ..., N_0)$ is equally likely before inspecting any of the lot. Jeffreys (1961) noted this as an argument for using $f(p) \equiv 1$ as a vague prior distribution for a binomial p. Also, form (37) is a predictive distribution; see Geisser (1985). If relevant, previous data are available, these could be incorporated through f(p), for example by a beta distribution in (36).

Consequently, the desired probability for shipping the lot without further inspection is

$$P[X_r < k - x_0 \mid N_r, N_0, X_0 = x_0]\,. \tag{38}$$

It is a marginal probability with respect to p and is a conditional distribution of X_r. By definition, this conditional distribution is the marginal joint distribution of X_r and X_0 divided by the marginal

distribution of X_0, which was computed in (36). Needed then is the marginal joint distribution of X_r and X_0,

$$P[X_r, X_0 \mid N_r, N_0] = \int_0^1 P[X_r, X_0 \mid N_r, N_0, p]\, f(p)\, dp\,. \tag{39}$$

Following the derivation of (37), with independence between X_r and X_0 due to random sampling and a constant but unknown probability of a defective, the integration yields the product of two binomial combination factors, one for each X_r and X_0, and the inverse of the normalization for a beta density. Recall that $p^{a-1}(1-p)^{b-1}$ is normalized by $\Gamma(a+b)/\Gamma(a)\Gamma(b)$. The result for (39) then is

$$P[X_r, X_0 \mid N_r, N_0] = \frac{1}{N_r+N_0+1}\binom{N_r}{x_r}\binom{N_0}{x_0}\Big/\binom{N_r+N_0}{x_r+x_0}. \tag{40}$$

Dividing the joint marginal distribution in (40) by the marginal distribution in (37) gives the desired conditional distribution that has been averaged over initially vague information on the probability of a defective for a single sample. Specifically,

$$P[X_r = x_r \mid N_r, N_0, x_0] = \frac{N_0+1}{N_r+N_0+1}\binom{N_r}{x_r}\binom{N_0}{x_0}\Big/\binom{N_r+N_0}{x_r+x_0}, \tag{41}$$

where the range of X_r is 0, 1, 2, ... N_r.

Compounding over vague weight functions on a binomial probability has been used in at least one other application. Documented by Gumbel (1958, 58ff) and attributed to H. A. Thomas, the distribution of exceedances was derived as sketched above. This discrete distribution used the weight function $[p(1-p)]^{-1}$, following Jeffreys (1961) guidelines for a vague prior distribution. Their focus was upon the probability of exceeding a previously observed extreme, say the m^{th} largest in n samples, some finite number of items in the next N samples. While extending the Bayes-Laplace theory to sampling a finite population, Jefferys (1961) derives the distribution of X, the total number of defects, given N, N_0, x_0 and a uniform prior on p. Geisser (1984) gives variations on the above mentioned result of Jeffreys (1961) for different vague priors on p.

Further, the hypergeometric distribution can be derived by the same process that leads to (41), but conditioning upon the total of X_r+X_0 rather than upon X_0. Part of Fisher's (1973, 96ff) description of his Exact Test for comparing two samples is compatible with the averaging in (36) and (39), but it is doubtful that he employed such a derivation. Using any prior f(p) for the proportion of defectives, and conditioning upon the total: X_r+X_0, X_r has the hypergeometric distribution.

In summary, the process of averaging over a vague distribution on the unknown probability of a

defective yields results like those in classical nonparametric statistics. Also, the resulting predictive distributions can have a repeated sampling orientation, based entirely on observable quantitites. For the above quality control scenario, the probability to ship the lot, given x_0 defectives among N_0 observed samples, is the probability of seeing no more than $k-x_0-1$ defectives among the N_r remaining samples. The resulting discrete distribution is a predictive one in the spirit of Roberts (1965) and Geisser (1970, 1985) and is a cousin to Fisher's Exact Test.

Case Study #6: Polychotomized Measurements

Continuous measurements are measured only to some finite degree of precision in practice. Time of death, as recorded on a death certificate, is available only to the day. Following Lindley (1965, p. 162), the sampling of any random variable "can be converted to the multinomial situation by grouping" on its range. Further, Lindley (1972a) points out that if a measurement "X is finite, then the class of *all* probability distributions over X is described by points in the simplex (p_1, p_2, ..., p_n: $p_i \geq 0$, $\Sigma p_i=1$) and (prior) distributions can be described and developed." The conjugate family is Dirichlet, and when appropriate, its use substantially reduces the difficulties of the general case.

Following the quality control scenario of the previous section for illustration, suppose defects are classified as one of s exclusive types. The probability of observing a defect of the ith type is p_i, $i = 1$, ..., s. The probability of not observing any defect is $1-\sum_{i=1}^{s} p_i$, where $\sum_{i=1}^{s} p_i \leq 1$. Let the random variable X_i be the number of defects of the ith type among a lot of size N and let X_{0i} be that number among the N_0 items inspected. Then $X_i = X_{0i} + X_{ri}$, where X_{ri} is the number of type i defectives among the $N_r = N - N_0$ remaining items. Quality control guidelines are directed at keeping the number of type i defectives in a lot to be shipped below some small number k_i for each $i = 1$, ..., s. When $X_{0i} = x_{0i}$ equals or exceeds k_i, for any $i = 1$, ..., s, the entire lot is inspected and repacked. When x_{0i} is less than k_i for all $i = 1$, ..., s, a decision to inspect and repack the lot can be based upon the probability of the number of type i defectives, x_{ri}, among the N_r not inspected, being less than $k_i - x_{0i}$ for each $i = 1$, ..., s. The statistical development follows.

The probability of x_{01}, x_{02}, ..., x_{0s} defectives among N_0 inspected items is

$$P[X_{0i} = x_{0i},\ i = 1, \ldots, s \mid N_0, p_i,\ i = 1, \ldots, s]$$

$$= \frac{N!}{x_1! \ldots x_{s+1}!}\ p_1^{x_1} \cdots p_s^{x_s} \left(1 - \sum_{i=1}^{s} p_i\right)^{x_{s+1}}, \qquad (42)$$

i.e., an s-variate multinomial distribution with independence and constant probability p_i of a type i defect among the N_0 inspected items. However, p_i is unknown so the marginal probability of $X_{0i} = x_{0i}$, $i = 1$, ..., s, is computed following an integration like that in (36), but over the simplex

region $S_s = \{(p_1, \ldots, p_s), p_k \geq 0, i = 1, \ldots, k \text{ and } \sum_{i=1}^{s} p_i \leq 1\}$. See Wilks (1962, p. 177ff). With vague prior information available for the unknown proportions of type i defects, i.e., a uniform weight function on S_s, the marginal calculation reduces to

$$P[x_{01}, \ldots, x_{0s} \mid N_0] = \frac{N_0!}{(N_0+s)!} . \tag{43}$$

With the generalized Haldane prior in Jeffreys (1962), viz., $[p_1 \cdots p_{s+1}]^{-1}$, the marginal distribution in (43) is $N_0[x_1 \cdots x_{s+1}]^{-1}$.

The probability for shipping the lot without further inspection is the probability of simultaneously acceptable numbers of defectives for each of the s types, specifically

$$P[X_{ri} < k_i - x_{0i}, i = 1, \ldots, s \mid N_r, N_0, X_{0i} = x_{0i}, i = 1, \ldots, s] . \tag{44}$$

This is a marginal distribution with respect to the s type i defective probabilities, p_i, and is a conditional distribution of the numbers of defects of each type, X_{ri} for $i = 1, \ldots, s$. By definition, (44) is determined as the ratio of the joint distribution of X_{ri} and X_{0i}, $i = 1, \ldots, s$, and the marginal distribution of X_{0i}, $i = 1, \ldots, s$, as computed in (43). The marginal joint distribution of the X_{ri} and X_{0i}

$$P[X_{ri}, X_{si}, i = 1, \ldots, s \mid N_r, N_0]$$

$$= \int_{S_s} P[X_{ri}, X_{0i}, i = 1, \ldots, s \mid N_r, N_0, p_i, i = 1, \ldots, s] f(p_1, \ldots, p_s)\, dp_1 \ldots dp_2 . \tag{45}$$

With independence between the X_{ri} and X_{0i}, $i = 1, \ldots, s$, due to random sampling of the inspected subset, and the constant but unknown probability of a type i defective, this integration yields the product of two multinomial factors from the inspected and remaining samples and the inverse of the normalization of a Dirichlet density. Dividing the result from (45) by that in (43), yields the desired conditional distribution that has been averaged over vague information on the probability of the i defective types, $i = 1, \ldots, s$. Using a uniform weight function for $f(p_1, \ldots, p_s)$ gives

$$P[X_{ri} = x_{ri}, i = 1, \ldots, s \mid N_r, N_0, x_{0i}, i = 1, \ldots, s]$$

$$= \frac{(N_0+s)!\, N_r!}{(N_r+N_0+s)!} \; \frac{(x_{r1}+x_{01})! \cdots (x_{r(s+1)}+x_{0(s+1)})!}{x_{r1}! \cdots x_{r(s+1)}!\, x_{01}! \cdots x_{0(2+1)}!} , \tag{46}$$

where the range of the X_{ri} is the non-negative intergers, subject to the constraint that $x_{r1} + \cdots + x_{r(s+1)} = N_r$.

The flexibility afforded by the s-variate multinomial is self-evident. Lindley's (1965) indication of its generality goes beyond the above quality control scenario and handles, for example, the distribution of any random variable by an s-variate exhaustive, but exclusive, grouping on the variable's range. Attention may be focused on selected categories of outcome. This reduces the dimension of the multinomial by the appropriate marginal distribution calculation. For example, interest may be in controlling the total number of defectives, regardless of type. Then the random variables $x_{0+} = \sum_{i=1}^{s} X_{0i}$ and $X_{r+} = \sum_{i=1}^{s} X_{ri}$ are considered and the s-variate multinomial reduces to a binomial, the subject of Case Study #5.

Discussion and Summary

Hill's (1968) work under a finite population model of N elements having M distinct values provides a benchmark of introduction and generality to Bayes nonparametrics. In the absence of ties ($M \equiv N$) the predictive posterior distribution from n sample observations is to assign probability $(n+1)^{-1}$ to each interval prescribed by the n order statistics, analogus to (37). Berliner and Hill (1988) accommodate censoring for survival applications. The product limit form (14) is deduced and they employ a linear smoothing between unique failures for applications. Within the structure of the discrete hazard model, tied observations and covariates are accommodated in Case Study #4.

The hazard function is emphasized in Case Study #2. A piecewise uniform development on the density, as contrasted with that on the hazard, is inspired by Berliner and Hill (1988), however, the inclusion of covariates with their perspective is not clear. A piecewise exponential model is detailed by Symons and Yuan (1987) achieving a smoother estimator than (14) and also accounts for the full information of each censored observation. Covariates can also be included with the piecewise exponential development.

Connections with frequentist nonparametric procedures are also noted. First, the comparison of distribution functions in Case Study #3 highlights the importance of specifying an alternative hypothesis and suggests that this may not be explicit in a frequentist procedure unless power assessments are incorporated. The Bayesian calculations with the discrete hazard model yielded a poor comparison with the Kruskal-Wallis test due to the generality of the discrete hazard model and the location sensitivity of the Kruskal-Wallis procedure. Second, the binomial portrayal of a quality control scenario in Case Study #5 led to a predictive distribution related to Fisher's Exact test. And following Geisser (1984), the hypergeometric distribution can be derived as a predictive distribution. Third, the nonparametric appeal of Cox's conditional likelihood is embraced by the Bayesians

procedures in Case Study #4. Specifically, the likelihood employed for practical estimation of the proportional hazards regression coefficients is a marginal one obtained by integrating out the nuisance parameters of the hazard. Frequentist approaches use this likelihood only as an approximation to procedures that require examining all possible orderings of tied observations. The Bayesian procedures incorporate the data at their available level of precision, including possible ties in measurements.

The terminology of nonparametric or distribution free is unsatisfactory, as per Noether (1984). Geisser's (1988) comment to Berliner and Hill (1988) suggests another term: "low structure". Whatever the language one point seems clear: parameter spaces of these case studies and constructs are typified by dimensions commensurate with the sample size. Low dimensional parameter spaces of parametric models are not included, as even Case Study #1 pointed to generalized distributions and higher dimensional parameter spaces.

For the basic ideas and applications of predictive distributions and probabilities, see Jeffreys (1961) for early examples, Roberts (1965) for an early sketch, and Geisser (1971ff) for continued development and advocacy. A commonality in the mechanics of these examples is the calculation of a marginal distribution over the model parameters. Many are predictive distributions, resulting from an average over a likelihood and vague prior that leaves only observables for the calculations. Highlights include:

- another example where biased estimation provides smaller mean squared error.
- clearly identified alternative hypotheses are a necessity of Bayesian inference, in contrast to that of Classical methods.
- practical decisions with ties in proportional hazards regression are retrieved by a marginal posterior of the regression coefficients.
- predictive distributions may be a Bayesian means to achieve nonparametric ends. For example, Fisher's Exact test can be derived as a predictive distribution.

Selected References

Berliner, M. L. and B. M. Hill (1988). Bayesian Nonparametric Survival Analysis. JASA, 83: 772-784 (with discussion).

Breslow, N. E. (1972). Contribution to the discussion of paper by D. R. Cox. Journal of the royal Statistical Society, Series B, 34: 216-7.

Breslow, N. E. (1974). Covariance analysis of censored survival data. Biometrics, 30: 89-99.

Conover, W. J. (1971). Pratical Nonparametric Statistics. John Wiley & Sons, New York.

Cox, D. R. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B, 34: 187-220.

Cox, D. R. (1975). Partial likelihood. Biometrika, 62: 269-276.

Cox, D. R. and D. Oakes (1984). Analysis of Survival Data. Chapman and Hall, New York.

Ericson, W. A. (1966). Personal Communication: Exam Problem. The University of Michigan, Ann Arbor Michigan.

Feller, W. (1943). On a general class of "contagious" distributions. Annals of Mathematical Statistics, 14: 389-400.

Geisser, S. (1971). The inferential use of predictive distributions. Foundations of Statistical Inference. Edited by V. P. Godambe and D. A. Sprott. Holt, Rinehart and Winston, Toronto, pp. 456-469.

Geisser, S. (1984). On prior distributions for binomial trials. The American Statistician, 38: 244-251.

Geisser, S. (1985). On the prediction of observables: A selective update. Bayesian Statistics 2, Edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith. Elsevier Science Publishers (North-Holland), pp. 203-230.

Good, I. J. and Gaskins, J. A. (1971). Nonparametric roughness penalties for probability densities. Biometrika, 58: 255-277.

Gumbel, E. J. (1958). Statistics of Extremes. Columbia University Press, New York.

Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. JASA, 63: 677-691.

Jeffreys, H. (1961). Theory of Probability. Third Edition. Oxford Press, London.

Kalbfleisch, J. D. and R. L. Prentice (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika, 60: 267-278.

Kalbfleisch, J. D. and R. L. Prentice (1980). The Statistical Analysis of Failure Time Data. Wiley, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53: 457-481.

Lindley, D. V. (1972a). Bayesian Statistics. A Review. Society for Industrial and Applied

Mathematics.

Lindley, D. V. (1972b). Contribution to discussion of paper by D. R. Cox. Journal of the Royal Statistical Society, Series B, 34: 208-9.

Nelson, W. B. (1972). Theory and applications of hazard plotting for censored failure data. Technometrics, 14: 945-965.

Noether, G. E. (1984). Nonparametrics: The Early Years - Impressions and Recollections. The American Statistician, 38: 173-178.

Peto, R. (1972). Contribution to discussion of paper by D. R. Cox. Journal of the Royal Statistical Society, Series B, 34: 205-7.

PHGLM. (1980). SAS Supplemental Library User's Guide. SAS Institute, North Carolina.

Rodriguez, R. N. (1982). Burr distributions. Encyclopedia of Statistical Sciences. Edited by S. Kotz, N. L. Johnson and C. B. Read. Volume 1: 335-340.

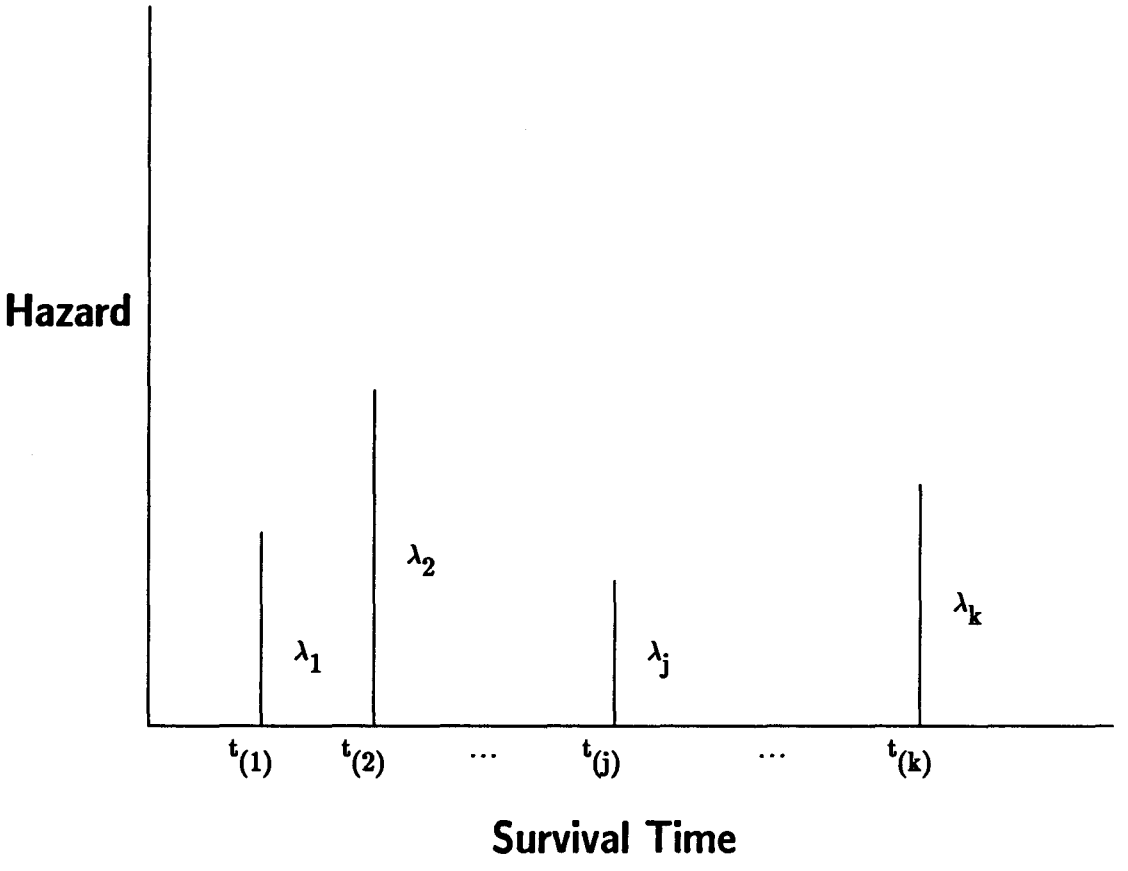
Roberts, H. V. (1965). Probabilistic prediction. Journal of the American Statistical Association, 60: 50-62.

Symons, M. J. and Y. C. Yuan (1987). Bayesian inference for survival data with nonparametric hazards and vague priors. Institute of Statistics Mimeo Series, No. 1491.

Turnbull, B. W., B. W. Brown, Jr. and M. Hu (1974). Survivorship analysis of heart transplant data. J. American Statistical Assoc. 69: 74-80.

Wilks, S. S. (1962). Mathematical Statistics. Wiley, New York.

Figure 1. Discrete Hazard Model*



* $\lambda_j = P[T=t_{(j)}|T\geq t_{(j)}]$

$t_{(j)}$ = Time of j^{th} failure