The Annals of Statistics
1989, Vol. 17, No. 4, 1509–1531

THE ADMISSIBILITY OF THE KAPLAN–MEIER AND OTHER MAXIMUM LIKELIHOOD ESTIMATORS IN THE PRESENCE OF CENSORING

BY G. MEEDEN,[1] M. GHOSH,[2] C. SRINIVASAN[3] AND S. VARDEMAN

Iowa State University, University of Florida, University of Kentucky and Iowa State University

For the nonparametric estimation of a survival function when censoring is present, the Kaplan–Meier estimator is often used. The admissibility of this estimator and other related maximum likelihood estimators is demonstrated. This is done by reducing the problem to one involving just the multinomial distribution and then using the stepwise Bayes technique to prove admissibility.

1. Introduction. The admissibility of various well-known nonparametric estimators was recently demonstrated in Meeden, Ghosh and Vardeman (1985). The argument proving admissibility involved two essential steps. The first was to note that to prove admissibility for the nonparametric problem it was enough to prove admissibility when the family of possible distributions was just assumed to be some multinomial family of distributions, rather than the large nonparametric family. The second step then used the stepwise Bayes technique to prove admissibility for the multinomial problem. Here we will use the same argument to prove the admissibility of the Kaplan–Meier estimator for the nonparametric estimation of the survival function when censoring is present. In fact, the admissibility of a whole class of maximum likelihood estimators, which includes the Kaplan–Meier estimator, is demonstrated. The only additional complication is that the censoring mechanism must be modeled and then taken into account. This description and modeling of the censoring is carried out in Section 3. In Section 4, the stepwise Bayes technique is used to demonstrate the admissibility of the maximum likelihood estimator for the multinomial problem. In Section 5, the admissibility results for the nonparametric problem are given.

If we were just proving the admissibility of the Kaplan–Meier estimator this paper would be considerably shorter. Although the proofs for the Kaplan–Meier estimator and the more general case are essentially the same, the general case requires considerably more notation. For someone who is just interested in the Kaplan–Meier estimator the argument can be simplified using the notation of Section 3.4. We believe, however, that the more general result is important.

Received July 1986; revised December 1988.

[1]Research partially supported by NSF Grant DMS-84-01740.

[2]Research partially supported by NSF Grant DMS-86-00066 and partially supported by the Army Research Office under Contract DAAG29-85-K-0189.

[3]Research partially supported by NSF Grant DMS-85-05774.

AMS 1980 *subject classifications*. Primary 62G05, 62C15.

Key words and phrases. Kaplan–Meier estimator, admissibility, maximum likelihood, censoring, multinomial distribution, stepwise Bayes.

Multinomial problems involving the more general type of censoring described here are not uncommon in statistical practice. For example, consider a two-dimensional contingency table where for some of the data points the value of the first variable is missing. For a recent discussion of such problems and other examples from a Bayesian point of view, see Dickey, Jeong and Kadane (1987).

2. Modeling the censoring mechanism.

2.1. *Introduction.* In this and the next section we will always assume that we are collecting observations from some multinomial population. As was noted above, once the multinomial situation is understood the nonparametric problem will follow easily.

For definiteness suppose we are to observe a random sample of size n from a population with five possible values, say $\alpha_1 < \alpha_2 < \alpha_3 < \alpha_4 < \alpha_5$. However, because of censoring on any given trial we may not learn the actual value of the observation but only that it belonged to a certain subset of the α_i's. Many such censoring schemes are possible. For example, suppose the following is what actually happens. On any given trial which results in α_1 we always get to observe α_1. On any trial which results in an α_2 sometimes we get to observe α_2 and at other times we only learn that the observation was at least as large as α_2. Similarly, on any trial which results in an α_i for $i = 3$, 4 or 5 we either learn that the actual observation was α_i or only that it was at least as large as α_i. On a trial which resulted in an α_i with $i > 1$, what we actually observe is the result of the random censoring mechanism which we have no control over. Hence, in actuality for this experiment, because of the censoring there are really eight possible outcomes. They are the five outcomes corresponding to each of the α_i's and the three censored outcomes at least as large as α_2, α_3 and α_4. Schematically this censoring mechanism is represented in Figure 1.

In the figure the point labeled (2) represents the outcome at least as large as α_2, the point labeled (2, 2) represents the outcome at least as large as α_3 and the

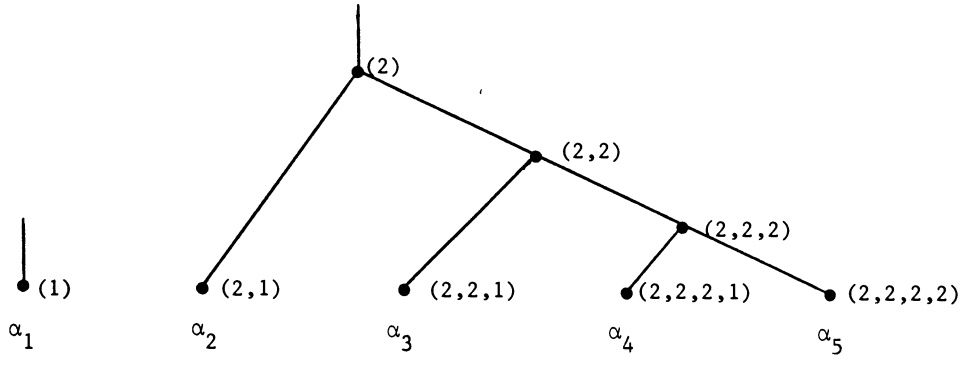

FIG. 1.

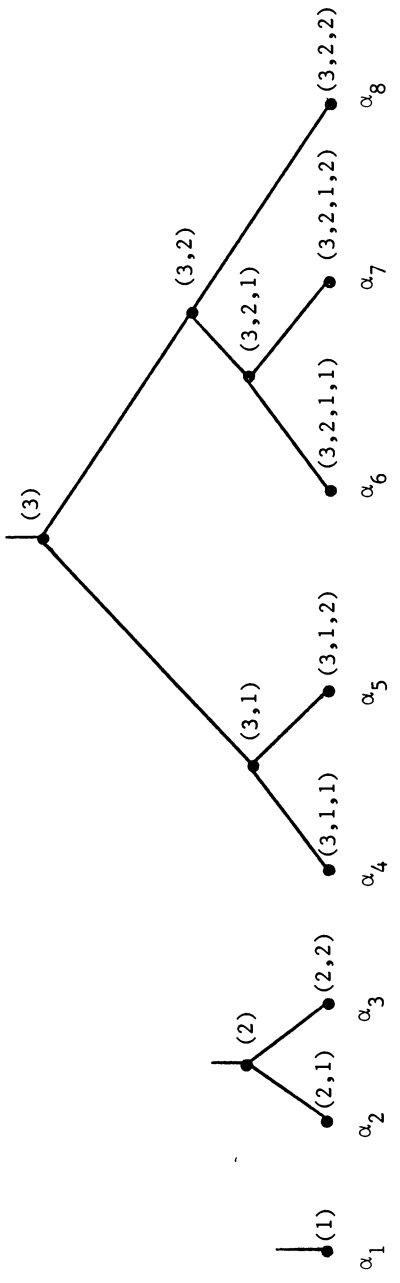
(1)
(2)
(2,1)
(2,2)
(3)
(3,1)
(3,1,1)
(3,1,2)
(3,2)
(3,2,1)
(3,2,1,1)
(3,2,1,2)
(3,2,2)
α_1
α_2
α_3
α_4
α_5
α_6
α_7
α_8

FIG. 2.

point labeled $(2, 2, 1)$ represents the outcome α_3. This method of labeling will soon be explained in greater detail.

We note in passing that the censoring in this example is of the type that is present when the Kaplan–Meier estimator is applicable, i.e., for a given observation you either learn the actual value or just that it is larger than some number.

Before studying the general situation we consider one other special case. Suppose a random sample of size n is to be taken from a population with eight possible values, say $\alpha_1 < \alpha_2 < \cdots < \alpha_8$. Because of censoring on any given trial, in addition to the eight basic outcomes, we sometimes only learn that the outcome belongs to one of the following sets: $\{\alpha_2, \alpha_3\}$, $\{\alpha_4, \alpha_5\}$, $\{\alpha_6, \alpha_7\}$, $\{\alpha_6, \alpha_7, \alpha_8\}$ or $\{\alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8\}$. Hence, on each trial there are in fact 13 possible outcomes for the experiment. For example, when α_7 is the result of a particular trial, what we actually observe can be either α_7, $\{\alpha_6, \alpha_7\}$, $\{\alpha_6, \alpha_7, \alpha_8\}$ or $\{\alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8\}$. Each of the 13 possible outcomes will be called a node and they are represented in Figure 2.

Note that $(3, 2, 1)$ is the node corresponding to the outcome $\{\alpha_6, \alpha_7\}$. The logic behind the labeling of the nodes will be given when the general situation is discussed.

We now consider the general problem. We assume that a random sample of size n is to be taken from a population with k possible values $\alpha_1, \alpha_2, \ldots, \alpha_k$. Because of censoring on any particular trial, rather than observing the actual value, we may only learn that it belongs to a certain subset of $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$. Let $\mathscr{U}$ denote the class of possible subsets that can be observed for a given, fixed censoring mechanism. We assume the censoring mechanism is such that $\mathscr{U}$ satisfies the following three properties:

(2.1)
(i) For $i = 1, \ldots, k$, $\{\alpha_i\} \in \mathscr{U}$.
(ii) For any two distinct sets belonging to $\mathscr{U}$ they are either disjoint or one is contained in the other.
(iii) The set $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ does not belong to $\mathscr{U}$.

Note that both censoring mechanisms in the two earlier examples satisfy these three conditions. Although there are many censoring mechanisms which satisfy the three conditions of (2.1) there are many others that do not. For example, any censoring mechanism which permits overlapping subsets of possible values fails to satisfy these conditions. The last condition eliminates the possibility that on a particular trial one learns only that an observation was taken, but absolutely nothing about the actual value. Although in many cases such an assumption is reasonable, it is not needed in what follows. However, for ease of exposition we will assume for now that it holds and show how it can be eliminated later.

From now on we will assume that a fixed censoring mechanism is given by some collection of subsets $\mathscr{U}$ which satisfies the conditions of (2.1). Then on any particular trial the statistician can observe any member of $\mathscr{U}$. We will now show how these outcomes can be labeled in a convenient fashion.

First, for each α_i find the largest member of $\mathscr{U}$ which contains it. Call this set u_i. Note that it is possible that some of the sets $u_1, \ldots, u_k$ are identical. However, the collection of distinct sets among $u_1, \ldots, u_k$ forms a partition of

$\{\alpha_1, \ldots, \alpha_k\}$. If there are m such distinct sets we denote (in some order) these distinct sets by (1), (2), ..., (m).

If the set (1) is a singleton, i.e., contains exactly one of the α_i's, we do nothing more with it. If (1) is not a singleton we then find for each α_i belonging to (1) the largest member of $\mathcal{U}$ which contains α_i and is a proper subset of (1). As in the above, the distinct members of this collection of sets forms a partition of the set (1). If this partition of (1) consists of m_1 distinct sets we denote these sets (in some order) by $(1,1), (1,2), \ldots, (1, m_1)$. Now, in turn, we consider each of the sets $(1, j)$. If for a particular j, $(1, j)$ is a singleton we do nothing more with it. If $(1, j)$ is not a singleton, then for each α_i belonging to $(1, j)$ we find the largest member of $\mathcal{U}$ which contains α_i and is a proper subset $(1, j)$. As before this gives us a partition of $(1, j)$ for each $(1, j)$ which is not a singleton. Now, in turn, we consider every set which is a member of the partition of $(1, j)$ for some j. If such a set is a singleton then we do nothing more with it. If it is not a singleton, then just as before we partition it. We continue in this way until we are just left with singletons and there are no further sets to partition.

We next repeat this process on the set (2) and so on through the set (m). This process allows us to denote the subsets in the collection $\mathcal{U}$ in a convenient way. A typical member of $\mathcal{U}$ is denoted by $(i, j, \ldots, r, s)$ and is called a node. These nodes, or members of $\mathcal{U}$, are exactly the set of all possible outcomes on any particular trial of the experiment. Note that if the collection $\mathcal{U}$ is just all the singletons $\{\alpha_i\}$ for $i = i, \ldots, k$ (i.e., there is no censoring), then node (i) is just α_i. It was this method that led to the labels attached to the nodes in Figures 1 and 2. For example, in Figure 2 the node $(3,2,1)$ corresponds to the outcome $\{\alpha_6, \alpha_7\}$ while the node $(3,2,1,1)$ corresponds to the outcome $\{\alpha_6\}$.

2.2. *Probability structure for Figure* 2. For the censoring experiments we are considering, one needs to specify two different probability structures. The first models how the α_i's are generated on a given trial while the second models the censoring mechanism that determines which subset of $\mathcal{U}$, i.e., which node, is actually observed by the statistician. To see how this will be done we return to the special case given in Figure 2.

Recall that in Figure 2 there are eight possible values for the underlying process. The most natural probability model is to let $q(i)$ denote the probability that α_i occurs for $i = 1, \ldots, 8$ where $\sum_{i=1}^{8} q(i) = 1$. Then the probability of any set of α_i's can be expressed using the $q(i)$'s. For example, the probability that the underlying outcome is in the set $\{\alpha_6, \alpha_7\}$ is $q(6) + q(7)$. However, because one needs to model the censoring mechanism as well, there is a more convenient way to model the underlying probability structure. We will now describe this model.

Let $\lambda(1)$, $\lambda(2)$ and $\lambda(3)$ be a probability distribution over the three nodes (1), (2) and (3), i.e., $\lambda(i) \geq 0$ and $\sum_{i=1}^{3} \lambda(i) = 1$. Next let $\lambda(1|2)$ and $\lambda(2|2)$ be the conditional probability distribution over the nodes $(2,1)$ and $(2,2)$ given node (2). Similarly we have the following conditional probability distributions: $\lambda(1|3)$ and $\lambda(2|3)$ over the nodes $(3,1)$ and $(3,2)$ given node (3), $\lambda(1|3,1)$ and $\lambda(2|3,1)$ over the nodes $(3,1,1)$ and $(3,1,2)$ given node $(3,1)$, $\lambda(1|3,2)$ and $\lambda(2|3,2)$ over the

nodes $(3, 2, 1)$ and $(3, 2, 2)$ given node $(3, 2)$ and $\lambda(1|3, 2, 1)$ and $\lambda(2|3, 2, 1)$ over the nodes $(3, 2, 1, 1)$ and $(3, 2, 1, 2)$ given node $(3, 2, 1)$. Under this parameterization the probability that underlying process generates α_2, which is the node $(2, 1)$, is $\lambda(2)\lambda(1|2)$ and the probability of generating α_6, the node $(3, 2, 1, 1)$, is $\lambda(3)\lambda(2|3)\lambda(1|3, 2)\lambda(1|3, 2, 1)$. Note that among the distributions defined by the λ's there are only seven independent parameters and this probability structure is just a reparameterization of $q(1), \ldots, q(8)$.

Next we describe how the censoring mechanism will be modeled. We see from Figure 2 that if the underlying probability structures generate α_1, then the statistician always observes α_1. On the other hand, if α_4 is generated the statistician may observe one of three sets $\{\alpha_4\}$, $\{\alpha_4, \alpha_5\}$ or $\{\alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8\}$. What the statistician actually observes is determined by a probabilistic censoring device. The probability structure of this censoring device will now be specified.

If the singleton node $(2, 1)$, i.e., α_2, is generated by the probability structure given by the λ's, then we observe node (2) with probability $p(2)$ and node $(2, 1)$ with probability $p(2, 1)$. Note that $p(2) + p(2, 1) = 1$. Similarly if node $(2, 2)$ is generated we observe nodes (2) and $(2, 2)$ with probabilities $p(2)$ and $p(2, 2)$. If node $(3, 1, 2)$ is generated we observe the nodes (3), $(3, 1)$ and $(3, 1, 2)$ with respective probabilities $p(3)$, $p(3, 1)$ and $p(3, 1, 2)$ where these three probabilities sum to 1. Or, if node $(3, 2, 1, 1)$ is generated we observe the nodes (3), $(3, 2)$, $(3, 2, 1)$ and $(3, 2, 1, 1)$ with respective probabilities $p(3)$, $p(3, 2)$, $p(3, 2, 1)$ and $p(3, 2, 1, 1)$ where these probabilities also sum to 1. Note that the censoring probabilities are defined in similar fashion for the other singleton nodes.

We can now calculate the probability of observing the various nodes on a particular trial. For example, the probability we observe node $(2, 1)$ is $p(2, 1)\lambda(2)\lambda(1|2)$. The probability we observe node (2) is

$$p(2)\lambda(2)\lambda(1|2) + p(2)\lambda(2)\lambda(2|2) = p(2)\lambda(2)$$

and the probability we observe node $(3, 2)$ is

$$\lambda(3)\lambda(2|3)\lambda(1|3, 2)[\lambda(1|3, 2, 1)p(3, 2) + \lambda(2|3, 2, 1)p(3, 2)] + \lambda(3)\lambda(2|3)\lambda(2|3, 2)p(3, 2) = p(3, 2)\lambda(3)\lambda(2|3),$$

and so on.

These calculations make it clear why the parameterization of the underlying probability distribution involving the λ's is more convenient for this problem than the more natural one given by the q's. Now in what follows both the probability model for generating the data, given by the λ's, and the probability model for the censoring, given by the p's, will be considered to be unknown. It will be for this combined parameter space that the admissibility results will be proved for the "multinomial" problem with censoring.

2.3. *Probability structure for the general model.* We now return to the general situation where the possible values for the underlying probability distribution are $\alpha_1, \ldots, \alpha_k$ and the censoring mechanism is given by a collection of sets $\mathscr{U}$ satisfying (2.1). As before a member of $\mathscr{U}$ is denoted by $(i, j, \ldots, r, s)$ and is called a node. The length of a node is just the number of indices in the vector

which names the node. If A is a node of length l_1 and B is a node of length l_2 we say that node A precedes node B if $l_1 < l_2$ and if the first l_1 coordinates which name B are exactly the coordinates which name A. For example, in Figure 2 the node $(3, 2)$ precedes the node $(3, 2, 1, 2)$. As was demonstrated in the special case, this labeling of the nodes can be used in specifying the underlying probability structure.

First we let $\lambda(\cdot)$ be a probability function defined on all the nodes of length 1. The rest of the distribution is defined conditionally. Given the node $(i, j, \ldots, r)$, we let $\lambda(s|i, j, \ldots, r)$ be the conditional probability of the node $(i, j, \ldots, r, s)$. If $(i, j, \ldots, r)$ is of length l, then $\lambda(\cdot|i, j, \ldots, r)$ concentrates all its mass on the nodes of length $l + 1$ which are immediately preceded by $(i, j, \ldots, r)$. Hence, the unconditional probability assigned to the node $(i, j, \ldots, r, s)$ is

$$\lambda((i, j, \ldots, r, s)) = \lambda(i)\lambda(j|i) \cdots \lambda(s|i, j, \ldots, r). \tag{2.2}$$

Next we will give the probability distribution which describes the censoring mechanism. Before this can be done, however, we need some additional notation. If $(i, j, \ldots, r, s)$ is a node which is not a singleton, then we let $\Lambda(i, j, \ldots, r, s)$ denote the set of all singleton nodes which are preceded by the node $(i, j, \ldots, r, s)$. If on a given trial of the experiment, because of censoring, we observe the node $(i, j, \ldots, r, s)$, then we know that the singleton node generated by the underlying probability structure defined by λ's belongs to the set $\Lambda(i, j, \ldots, r, s)$. Whenever any of the singleton nodes belonging to this set are generated, we assume that the censoring mechanism yields the node $(i, j, \ldots, r, s)$ with probability $p(i, j, \ldots, r, s)$. Hence, whenever the node $(i, j, \ldots, r, s)$ is observed our model assumes that for all the singleton nodes, which could have given rise to the observation, the censoring probability is the same. This type of censoring is called noninformative [see, for example, Berger and Wolpert (1984), page 96].

Hence, with noninformative censoring it is easy to check that on a particular trial we observe the node $(i, j, \ldots, r, s)$ with probability

$$p(i, j, \ldots, r)\lambda((i, j, \ldots, r, s)). \tag{2.3}$$

For a given generic node $(i, j, \ldots, r, s)$ which is not a singleton, let $\Lambda = \Lambda(i, j, \ldots, r, s)$ and let α denote a generic singleton belonging to Λ. Then on a particular trial, for a given p and λ,

$$P_{\lambda, p}(\alpha \text{ was generated}|(i, j, \ldots, r, s) \text{ observed}) = \frac{p(i, j, \ldots, r, s)\lambda(\alpha)}{\sum_{\beta \in \Lambda} p(i, j, \ldots, r, s)\lambda(\beta)}$$

$$= \frac{\lambda(\alpha)}{\sum_{\beta \in \Lambda}\lambda(\beta)} = P_{\lambda}(\alpha|\Lambda).$$

Conversely, it can be seen that if the previous equation is true then the censoring mechanism is of the form previously described. This explains why this censoring mechanism is called noninformative.

Suppose on a particular trial the outcome α_i is generated by the underlying probability distribution. For each α_i let $V(\alpha_i)$ be the set which contains the node

α_i and all the other nodes which precede it. So given the value α_i the node that is actually observed must belong to the set $V(\alpha_i)$. So as we saw in the discussion of the example given in Figure 2, each α_i introduces a constraint over the censoring probabilities. That is, for each α_i the sum of the censoring probabilities for all the nodes belonging to $V(\alpha_i)$ must be 1.

We have now described the parameter space for this problem. Because of the censoring the parameter space has two components. The first is θ_λ which consists of all possible choices for the various λ's, which defines the underlying probability structure. The second is θ_p which consists of all possible choices for the various p's which yield noninformative censoring. Hence, other than the assumption of noninformative censoring we are following standard frequentist practice and assuming the possible distributions are completely unknown. As was noted in (2.3) we are also assuming that the underlying probability structure and the censoring mechanism are independent of one another. Under these assumptions the appropriate parameter space is $\theta = \theta_\lambda \times \theta_p$.

λ denotes a typical member of θ_λ. The labeling of the nodes gives a convenient way to order the members of λ. That is, we will take

$$\begin{aligned}\lambda = (&\lambda(1), \lambda(2), \ldots,\\ &\lambda(1|1), \lambda(2|1), \ldots, \lambda(1|2), \lambda(2|2), \ldots, \ldots, \ldots,\\ &\lambda(1|1,1), \lambda(2|1,1), \ldots, \lambda(1|1,2), \lambda(2|1,2), \ldots, \ldots,\\ &\ldots, \ldots, \ldots, \ldots, \qquad \ldots,\\ &\ldots, \ldots, \ldots, \qquad \ldots, \ldots,),\end{aligned}$$

i.e., we have taken the λ's in lexicographic order. When it is convenient to think of the λ's in this order we will write $\lambda = (\lambda_{[1]}, \ldots, \lambda_{[\gamma]})$ when γ is the total number of λ's that we have.

Suppose we now observe n independent trials of the experiment. Let $y(i, j, \ldots, r, s)$ be the number of times the node $(i, j, \ldots, r, s)$ is observed in the sample. Then, given the data, the likelihood function defined over θ is given by

$$\begin{aligned}&L(\lambda\text{'s}, p\text{'s}|\text{data})\\ &= \Pi_{(i,j,\ldots,r,s)}\{p(i, j, \ldots, r, s)\\ &\qquad \times \lambda(i)\lambda(j|i) \cdots \lambda(s|i, j, \ldots, r)\}^{y(i,j,\ldots,r,s)},\end{aligned} \tag{2.4}$$

where $\Pi_{(i,j,\ldots,r,s)}$ denotes the product over all possible nodes. Let $\Pi_{(i)}$ denote the product over all nodes of length (1), $\Pi_{(i,j)}$ denote the product over all nodes of length (2) and so on until finally $\Pi^*_{(i,j,\ldots,r,s)}$ denotes the product of the set of nodes of maximum length. Let $\bar{y}(i, j, \ldots, r, s)$ be the total number of times $(i, j, \ldots, r, s)$ or a node which it precedes is observed. Then the likelihood function can be rewritten as

$$\begin{aligned}L(\lambda\text{'s}, p\text{'s}|\text{data}) &= \Pi_{(i,j,\ldots,r,s)}[p(i, j, \ldots, r, s)]^{y(i,j,\ldots,r,s)}\\ &\quad \times \Pi_{(i)}[\lambda(i)]^{\bar{y}(i)}\Pi_{(i,j)}[\lambda(j|i)]^{\bar{y}(i,j)} \cdots\\ &\quad \times \Pi^*_{(i,j,\ldots,r,s)}[\lambda(s|i, j, \ldots, r)]^{\bar{y}(i,j,\ldots,r,s)}.\end{aligned} \tag{2.5}$$

In most problems the censoring probabilities will be considered nuisance parameters. What one is really interested in estimating is the λ's or various functions of the λ's. If the sample is such that every node was observed at least once, then the maximum likelihood estimates (MLEs) for the λ's are easily found. In this case it follows immediately that

$$\text{MLE of } \lambda(i_0) = \frac{\bar{y}(i_0)}{\Sigma_{(i)} y(i)} = \frac{\bar{y}(i_0)}{n},$$

$$\text{MLE of } \lambda(j_0|i) = \frac{\bar{y}(i, j_0)}{\Sigma_j \bar{y}(i, j)} = \frac{\bar{y}(i, j_0)}{\bar{y}(i) - y(i)}$$

and

$$\text{MLE of } \lambda(s_0|i, j, \ldots, r) = \frac{\bar{y}(i, j, \ldots, r, s_0)}{\Sigma_s \bar{y}(i, j, \ldots, r, s)}.$$

From this we see that the MLE of (2.2), the probability that the λ's assign to node $(i_0, j_0, \ldots, r_0, s_0)$, is given by

$$\frac{\bar{y}(i_0)}{n} \frac{\bar{y}(i, j_0)}{\Sigma_j \bar{y}(i, j)} \times \cdots \times \frac{\bar{y}(i_0, j_0, \ldots, r_0, s_0)}{\Sigma_s \bar{y}(i_0, j_0, \ldots, r_0, s)}. \tag{2.6}$$

On the other hand, it is possible because of censoring that the sample is such that not every node was observed. In this case, it is possible that the MLE for some of the λ's is undefined. In the next section it will be pointed out how this difficulty is easily overcome and the admissibility of various MLEs will be demonstrated.

3. Admissibility in the multinomial case.

3.1. *Introduction.* We will now use the stepwise Bayes technique to prove the admissibility of the MLE for estimating λ with the loss function being the total sum of all the squared error losses. A typical stepwise Bayes argument proceeds as follows. A prior distribution is defined over a subset of the parameter space. The Bayes estimate for all points in the sample space where it is uniquely determined is computed. If this is the entire space the estimator is determined and the process is over.

If this is not the case let S_1 denote the set of points where the first estimator is not uniquely determined. Then a new problem is considered having as its sample space S_1 and the family of possible distributions as the original distributions conditioned on S_1. Then another prior is defined over the part of the parameter space for which these conditional distributions are well-defined and whose support is disjoint from the support of the first prior. The points in the sample space which did not have a unique Bayes estimator under the first prior but do under the second prior are found and the estimates are computed. If any

points in the sample space still do not have a uniquely defined estimate the above process is repeated and a third prior is chosen, disjoint from the first two. We continue in this way until we have a stepwise Bayes estimator defined for each point in the sample space. It is possible using this argument to prove the admissibility of the MLE in the usual multinomial problem with no censoring [see, for example, Brown (1981)]. The argument which we give below is essentially this earlier one with some additional complexities because of the problems caused by censoring.

One minor difficulty is that the MLE of the λ's need not be defined for all possible outcomes. For example, consider the model in Figure 2 and suppose that a sample of size $n = 3$ was taken and all three observations were the node $(3, 1)$, i.e., $y(3, 1) = 3$. From our earlier remarks it is clear that the MLE of $\lambda(3) = 1$ and of $\lambda(1|3) = 1$. In addition, with the exception of $\lambda(1|3, 1)$ and $\lambda(2|3, 1)$ the MLEs of all the other λ's are 0 or defined to be 0. For example, the MLE of $\lambda(2)$ is 0 and if we adopt the convention that $0/0 = 0$, then the MLEs of $\lambda(1|2)$ and $\lambda(2|2)$ are both 0 as well. Since $\lambda(1|3, 1)$ and $\lambda(2|3, 2)$ do not appear in the likelihood function [see (2.5)], the MLE for each of them is undefined. From a frequentist point of view these parameters cannot be estimated. This would cause no problem for a Bayesian, however, if he had a prior distribution over $\lambda(1|3, 1)$ and $\lambda(2|3, 1)$. He then could treat this as a no data problem, which it really is, and estimate $\lambda(1|3, 1)$ with his prior mean. In what follows we will use this Bayesian solution when the MLE is undefined.

Returning to the general model we will define a prior distribution over θ_λ. Let h be a prior distribution over all nodes of length 1. Proceeding inductively, if $(i, j, \ldots, r)$ is a node of length $l \geq 2$ and is not a singleton let $w(i, j, \ldots, r) = w$ denote the number of nodes of length $l + 1$ which it immediately precedes. Let $h_{(i, j, \ldots, r)}$ be a prior density over the simplex $\{(\lambda(1|i, j, \ldots, r), \ldots, \lambda(w|i, j, \ldots, r)): \lambda(m|i, j, \ldots, r) \geq 0 \text{ for } m = 1, \ldots, w \text{ and } \sum_{m=1}^{w} \lambda(m|i, j, \ldots, r) = 1\}$. We will assume that all the densities are independent and hence the product, denoted by H, is a probability density over θ_λ. If, in addition, one defined a prior over θ_p and assumed it was independent of H, then a straightforward Bayesian analysis could, in principle, be carried out. We will not do that however. The prior H, or more properly factors of H, will be appealed to only for those outcomes where all the nodes were not observed and some of the λ's do not appear in the likelihood.

A second and more important difficulty in using the stepwise Bayes technique is choosing the priors in the proper order so that the resulting estimator is indeed the MLE; or equivalently the sets of sample points for which the Bayes estimators are computed at each step must come in the correct order. We will now introduce some additional notation that will help in the ordering.

Given a sample of size n let Γ be the set of nodes which were observed at least once, i.e., the set of $(i, j, \ldots, r, s)$'s for which $y(i, j, \ldots, r, s) > 0$. So Γ is just a subcollection of $\mathscr{U}$.

We say that a set of distinct nodes $A_1, \ldots, A_m$ forms a chain of length m if A_m precedes A_{m-1} which precedes $A_{m-2}, \ldots,$ which precedes A_1.

Let Γ^* be the subset of Γ which contains all the nodes in Γ which do not precede any other node belonging to Γ. Let A_1 be a node belonging to Γ^*. Consider the subset of Γ which consists of all the nodes which precede A_1. If this subset is empty we are done and A_1 is a chain. If this subset is not empty let A_2 be the member of this set which precedes no other member of the set. Now repeat this process using A_2 and the remaining members of the set. This yields a maximal chain $A_1, A_2, \ldots, A_u$ of nodes of Γ dominating A_1.

Now let $B_1 \in \Gamma^*$ with $B_1 \neq A_1$ and find the corresponding maximal chain generated by B_1, say $B_1, \ldots, B_u$. Note $A_1 \neq B_1$ but after that the A_i's and B_j's can have nodes in common. Hence, given a set Γ, we can construct a unique chain using each member of Γ^*. Note that every member of Γ will belong to at least one chain. Given a Γ which yields m chains let $l(1) \leq l(2) \leq \cdots \leq l(m)$ be the lengths of the m chains in nondecreasing order. So a set Γ yields a unique vector of numbers $(l(1), \ldots, l(m))$, the lengths of its chains.

The stepwise Bayes argument will proceed by choosing priors which first take care of all Γ's which yield just one chain. Then priors are chosen which take care of Γ's yielding two chains and so on. In addition, within each group we will have to proceed in the proper order. Because our loss is the sum of the individual squared error losses, any Bayes estimate of a parameter is just its posterior expectation.

3.2. *The stepwise argument for Figure* 2. In the first step of the argument we consider all those samples Γ which yield just one chain. Within this class we first consider all the chains of length 1, then all chains of length 2 and so on. Within each group the order in which we consider the chains does not matter.

When we observe one chain of length 1 that means all n of our observations were identical. For example, suppose $y(1) = n$, i.e., α_1, the node labeled (1), was observed all n times. Consider a prior on θ_λ which puts mass 1 on the point $\lambda(1) = 1$ and under this prior the Bayes estimate of $\lambda(1)$ is 1 and for all the other λ's is 0 which is the MLE.

Next suppose the sample is such that $y(3,2,1,1) = n$, i.e., the node α_6, the node labeled $(3,2,1,1)$, was observed each time. On θ_λ consider the prior which puts mass 1 on the point $\{\lambda(3) = 1, \lambda(2|3) = 1, \lambda(1|3,2) = 1$ and $\lambda(1|3,2,1) = 1\}$. On θ_p consider the prior which puts mass 1 on the point $\{p(3,2,1,1) = 1\}$. Under this prior the Bayes estimate of $\lambda(3)$, $\lambda(2|3)$, $\lambda(1|3,2)$ and $\lambda(1|3,2,1)$ are all 1 and the Bayes estimates of the rest of the λ's are all 0. Again this is just the MLE. Note that under this prior no nodes, other than $(3,2,1,1)$ have positive probability. It was for this reason that it was necessary to introduce a prior over θ_p even though we are only interested in estimating the λ's. Without this prior over θ_p the stepwise Bayes argument would not yield the admissibility of the MLE. As the argument proceeds, at each step we will specify a prior over θ_p and θ_λ.

Note that all the other singleton nodes can be taken care of in a similar way. Next we consider a node which is not a singleton. Suppose, for example, $y(3,2,1) = n$. In this case we need an estimate of $\lambda(1|3,2,1)$ and $\lambda(2|3,2,1)$ even

though they do not appear in the likelihood. For this reason our prior over θ_λ is the product of two priors. The first puts mass 1 on the point $\{\lambda(3) = 1, \lambda(2|3) = 1 \text{ and } \lambda(1|3,2) = 1\}$ and the second is $h_{(3,2,1)}$ the factor of H, which concentrates its mass on the simplex $\lambda(1|3,2,1) + \lambda(2|3,2,1) = 1$. Over θ_p we take the prior which puts mass 1 on the set $\{p(3,2,1) = 1\}$. For this prior the Bayes estimates of $\lambda(3)$, $\lambda(2|3)$ and $\lambda(1|3,2)$ are each 1, which are the MLEs, and the Bayes estimates of $\lambda(1|3,2,1)$ and $\lambda(2|3,2,1)$ are their respective prior expectations. As before no other nodes have positive probability under this prior.

It is easy to check that all other Γ's which yield just one chain of length 1 can be handled in a similar fashion. Next we will consider Γ's which yield just one chain of length 2. For example, consider a Γ for which

$$\begin{aligned} &y(2) > 0 \quad \text{and} \quad y(2,1) > 0, \\ &\text{and } y(2) + y(2,1) = n. \end{aligned} \tag{3.1}$$

On θ_λ we take the prior which puts mass 1 on the point $\{\lambda(2) = 1 \text{ and } \lambda(1|2) = 1\}$. On θ_p we take any prior which is concentrated on the set $\{p(2) > 0, p(2,1) > 0 \text{ and } p(2) + p(2,1) = 1\}$ independent of the prior over θ_λ. The Bayes estimates of $\lambda(2)$ and $\lambda(1|2)$ are each 1. The only samples which have positive probability under this prior, other than those which satisfy (3.1), are the two Γ's with $y(2) = n$ and $y(2,1) = n$, respectively. These two samples have already been taken care of when we considered chains of length 1.

Another possibility for a sample with one chain of length 2 is when $y(3) > 0$ and $y(3,2,1) > 0$ with $y(3) + y(3,2,1) = n$. The prior over θ_λ is the product of two priors. The first puts mass 1 on the point $\{\lambda(3) = \lambda(2|3) = \lambda(1|3,2) = 1\}$ and the second is $h_{(3,2)}$, the factor of H, which concentrates its mass on the simplex $\lambda(1|3,2,1) + \lambda(2|3,2,1) = 1$. On θ_p we take any prior which is concentrated on the set $\{p(3) > 0, p(3,2,1) > 0 \text{ and } p(3) + p(3,2,1) = 1\}$ and independent of the prior over θ_λ.

The only Γ's, other than the type being considered, which have positive probability under this prior, are the two with $y(3) = n$ and $y(3,2,1) = n$ which we have already taken care of. Again the Bayes estimates are clear and are the MLEs.

It is easy to check that all single chains of length 2 can be handled in the same way. Then all single chains of length 3 can be taken care of, and then all single chains of length 4 and so on until all single chains of any length have been considered.

The next major step is to consider all Γ's which yield two chains. The order in which we proceed is determined as follows. Suppose Γ_1 yields $(l_1(1), l_1(2))$ and Γ_2 yields $(l_2(1), l_2(2))$. If $l_1(1) < l_2(1)$ or if $l_1(1) = l_2(1)$ and $l_1(2) < l_2(2)$, then Γ_1 is considered first. If $l_1(1) = l_2(1)$ and $l_1(2) = l_2(2)$, then the order in which they are considered does not matter.

We first consider Γ's which yield two chains, each of length 1. For example, suppose that $y(1) > 0$, $y(2,1) > 0$ and $y(1) + y(2,1) = n$. The prior over θ_λ is the product of two priors. The first puts mass 1 on the point $\{\lambda(1|2) = 1\}$. The second prior is concentrated on the set $\{\lambda(1) > 0, \lambda(2) > 0 \text{ and } \lambda(1) + \lambda(2) = 1\}$ and is proportional to $(1 - (\lambda(1))^n - (\lambda(2))^n)/(\lambda(1)\lambda(2))$. Over θ_p we take the

prior which puts mass 1 on the point $\{p(1) = p(1,2) = 1 \text{ and } p(2) = 0\}$. For the restricted problem [i.e., ignoring the samples $y(1) = n$ and $y(2,1) = n$ which we have already taken care of] the resulting Bayes estimates of $\lambda(1)$ and $\lambda(2)$ are just the MLEs.

In the previous example the second prior over θ_λ is the product of two factors

$$[1 - (\lambda(1))^n - (\lambda(2))^n] \tag{3.2}$$

and

$$1/\lambda(1)\lambda(2). \tag{3.3}$$

It is the second factor (3.3) which is the important one. The first factor (3.2) just ensures that the prior is proper, i.e., ensures that it integrates to 1, for the restricted problem. It has no influence on the value of the stepwise Bayes estimator which depends only on factor (3.3). Under the prior defined in the previous paragraph the only samples that are possible are those n samples with $y(1) \geq 0$, $y(2,1) \geq 0$ and $y(1) + y(2,1) = n$. Recall that the earlier stages of the stepwise argument have already taken care of the two samples with $y(1) = n$ and $y(2,1) = n$. Hence at this stage for the restricted problem with the remaining $(n - 2)$ possible samples with $y(1) > 0$, $y(2,1) > 0$ and $y(1) + y(2,1) = n$, the new likelihood function is the original likelihood divided by (3.2). Hence when we compute the stepwise Bayes estimator for the restricted problem at this stage the factor (3.2) cancels and the value of the estimator depends only on (3.3). So we see that we did not need to find explicitly the factor (3.2). In what follows we will usually omit giving explicitly the factor in the prior which comes from normalizing the likelihood function.

Next we consider samples Γ where $y(1) > 0$, $y(2) > 0$ and $y(1) + y(2) = n$. The prior is the product of four factors. On the set $\{\lambda(1) > 0, \lambda(2) > 0 \text{ and } \lambda(1) + \lambda(2) = 1\}$ we have the factor $1/(\lambda(1)\lambda(2))$. The second factor is $h_{(2)}$, the factor of H_1 which concentrates its mass on the simplex $\lambda(1|2) + \lambda(2|2) = 1$. Over θ_p we have the factor which puts mass 1 on the point $\{p(2) = 1\}$. The final factor is the normalizing factor for the likelihood function for this restricted problem. Under this prior the stepwise Bayes estimate of $\lambda(1)$ is $y(1)/n$ and of $\lambda(2)$ is $y(2)/n$ and of $\lambda(1|2)$ and $\lambda(2|2)$ are their respective prior expectations under $h_{(2)}$. It is easy to check that all samples Γ yielding just two chains each of length 1 can be taken care of in the same way.

Next we consider samples of the type, where $y(1) > 0$, $y(2) > 0$, $y(2,1) > 0$ and $y(1) + y(2) + y(2,1) = n$. The prior over $\theta_\lambda \times \theta_p$ will be the product of four factors. The first factor is $1/\lambda(1)\lambda(2)$ and is concentrated on the set $\{\lambda(1) > 0, \lambda(2) > 0 \text{ and } \lambda(1) + \lambda(2) = 1\}$. The second factor is 1 on the set $\{\lambda(1|2) = 1\}$. The third factor is some probability density function on the set $\{0 < p(2) < 1\}$. The actual choice is not important. The final factor is

$$\sum (\lambda(1))^{y(1)}[p(2)\lambda(2)]^{y(2)}[p(2,1)\lambda(2)\lambda(1|2)]^{y(2,1)}, \tag{3.4}$$

which is the normalizing factor for the likelihood function, for the restricted problem at this stage, where the summation is over all samples with $y(1) > 0$,

$y(2) > 0$, $y(2,1) > 0$ and $y(1) + y(2) + y(2,1) = n$. Under this prior the stepwise Bayes estimate of $\lambda(1)$ is $y(1)/n$ and of $\lambda(2)$ is $[y(2) + y(2,1)]/n$.

It is easy to check that all samples Γ yielding just two chains can be taken care of in the same way when the ordering given above is used.

Please note that since the prior used in the previous example contains the factor (3.4) the prior over $\theta_\lambda \times \theta_p$ cannot be factored into two independent parts, one on θ_λ and one on θ_p. However, the factor (3.4) does not enter in the computation of the stepwise Bayes estimate and the remaining factors are in fact "independent." This is why the stepwise Bayes estimates depend only on the factors involving the λ's alone.

The next step is to consider all samples Γ which have three chains. After those are completed the next step is to consider all samples Γ which yield four chains and so on until all possible samples have been taken care of. Within each step the order in which the samples are considered is determined as follows. Let Γ_1 and Γ_2 be two samples each yielding m chains. Let $(l_i(1), \ldots, l_i(m))$ be the vector lengths of the sample Γ_i for $i = 1$ and 2. We will consider Γ_1 before Γ_2 if for some integer $2 \leq j \leq m$,

$$\begin{aligned} &l_1(i) = l_2(i) \quad \text{for } i = 1, \ldots, j-1 \\ &\text{and } l_1(j) < l_2(j). \end{aligned} \tag{3.5}$$

3.3. *The stepwise argument for the general case.* Rather than continue the proof in the special case we will return to the general problem. As we have noted for a given sample Γ not all members of the vector $(\lambda_{[1]}, \lambda_{[2]}, \ldots, \lambda_{[\gamma]})$ need appear in the likelihood function given in (3.5). We need a convenient notation to denote which λ's appear and which do not appear in the likelihood function for Γ.

For a given Γ let $S(\Gamma)$ denote all the nodes of length 1 which belong to Γ and $S'(\Gamma)$ all the nodes of length 1 which do not belong to Γ. Proceeding inductively given a node of the form $(i, j, \ldots, r)$, let

$$\begin{aligned} S'(\Gamma)_{(i,j,\ldots,r)} &= \left\{ \begin{array}{l} (i, j, \ldots, r, s)\colon (i, j, \ldots, r, s) \notin \Gamma \text{ and it} \\ \qquad \text{does not precede any member of } \Gamma \end{array} \right\}, \\ S(\Gamma)_{(i,j,\ldots,r)} &= \left\{ (i, j, \ldots, r, s)\colon (i, j, \ldots, r, s) \notin \underset{(i,j,\ldots,r)}{S'(\Gamma)} \right\}. \end{aligned} \tag{3.6}$$

Note that $(i, j, \ldots, r, s) \in S'(\Gamma)_{(i,j,\ldots,r)}$ if and only if $\bar{y}(i, j, \ldots, r, s) = 0$. Hence the union of the $S'(\Gamma)$'s identifies all of the λ's which do not appear in the likelihood function when the sample is Γ.

We are now ready to prove the admissibility of the MLE discussed earlier.

THEOREM 1. *Assume a random sample of size n is taken from a population with k possible values, $\alpha_1, \alpha_2, \ldots, \alpha_k$. Assume that what is actually observed on each trial is the result of a noninformative censoring mechanism which is*

described by a collection of sets $\mathcal{U}$ satisfying conditions (i) *and* (ii) *of* (2.1). *Let the collection $\mathcal{U}$ be used to parameterize the underlying probability structure and censoring mechanism, resulting in the parameter space $\theta = \theta_\lambda \times \theta_p$.*

Consider the problem of estimating the vector λ with the sum of the individual squared error losses as the loss function.

Let H be a probability density over θ_λ which is a product of individual densities over the factor spaces making up the product space θ_λ.

Given a sample Γ we define the estimator $\delta_H(\Gamma) = (\delta_{[1]}, \ldots, \delta_{[\gamma]})$ of $\lambda = (\lambda_{[1]}, \ldots, \lambda_{[\gamma]})$ as follows:

> *For the node (i) of length 1 the estimate of $\lambda(i)$ is $\bar{y}(i)/n$. Given the node $(i, j, \ldots, r)$ the estimate of $\lambda(s_0|i, j, \ldots, r)$ when $s_0 \in S(\Gamma)_{(i, j, \ldots, r)}$ is*
>
> $$\frac{\bar{y}(i, j, \ldots, r, s_0)}{\Sigma_s \bar{y}(i, j, \ldots, r, s)}$$
>
> *and when $s_0 \in S'(\Gamma)_{(i, j, \ldots, r)}$ is the expectation of $\lambda(s_0|i, Bj, \ldots, r)$ under the prior density $h_{(i, j, \ldots, r)}$. δ_H is called the H-augmented MLE of λ and is admissible.*

PROOF. We first prove the result in the special case where condition (iii) of (2.1) holds as well. Once this is done the more general case will follow easily. The argument uses the stepwise Bayes technique and proceeds just as we described in the special case. We first consider, in the proper order, all samples Γ^* which yield just one chain and then those that yield two chains and so on. Suppose that we are at the stage in the argument where we are considering a sample Γ which yields m chains with lengths $(l(1), \ldots, l(m))$.

The prior over $\theta = \theta_\lambda \times \theta_p$ will consist of three types of factors: ones involving just the λ's, others involving just the p's and a final factor involving both λ's and p's which is the normalizing factor for the likelihood function for the restricted problem. This last factor need not be found explicitly since it plays no role in the calculation of the stepwise Bayes estimate.

The factors over θ_λ which must be used at this stage are found as follows. For any node (i) of length 1 which belongs to $S'(\Gamma)$ the point $\{\lambda(i) = 0\}$ is assigned probability 1. If $S(\Gamma)$ contains just one element, say (i_0) we assign mass 1 to the point $\{\lambda(i_0) = 1\}$. If $S(\Gamma)$ contains v members, say nodes $(i_1), \ldots, (i_v)$, the prior is concentrated on the set $\{(\lambda(i_1), \ldots, \lambda(i_v))$: $\lambda(i_a) > 0$ for $a = 1, \ldots, v$ and $\Sigma_{a=1}^{v} \lambda(i_a) = 1\}$; over this set the factor is

$$1 \Big/ \prod_{a=1}^{v} \lambda(i_a).$$

Proceeding inductively we assume that we have taken care of all the λ's which assign probability to all nodes of the form $(i, j, \ldots, r)$. For a given $(i, j, \ldots, r)$

we now define a factor over the probability distribution $\lambda(\cdot|i, j, \ldots, r)$. There are three different possibilities. The first is that $S(\Gamma)_{(i, j, \ldots, r)}$ is empty. In this case none of the $\lambda(\cdot|i, j, \ldots, r)$'s appear in the likelihood function for Γ and we take as the factor for the $\lambda(\cdot|i, j, \ldots, r)$'s the prior $h_{(i, j, \ldots, r)}$ belonging to H. The second possibility is that $S(\Gamma)_{(i, j, \ldots, r)}$ contains just one point, say $(i, j, \ldots, r, s_0)$. Here we take the factor which puts mass 1 on the point $\{\lambda(s_0|i, j, \ldots, r) = 1\}$. The final possibility is that $S(\Gamma)_{(i, j, \ldots, r)}$ contains $v > 1$ members, say $(i, j, \ldots, r, s_1), \ldots, (i, j, \ldots, r, s_v)$. In this case the factor is concentrated on the set $\{(\lambda(s_1|i, j, \ldots, r), \ldots, \lambda(s_v|i, j, \ldots, r)): \lambda(s_a|i, j, \ldots, r) > 0$ for $a = 1, \ldots, v$ and $\sum_{a=1}^{v} \lambda(s_a|i, j, \ldots, r) = 1\}$; over this set the factor is

$$1 \Big/ \prod_{a=1}^{v} \lambda(s_a|i, j, \ldots, r).$$

It remains now to specify a factor over θ_p, the parameter space for the censoring probabilities. We begin by setting to 0 any censoring probability associated with any node that was not observed in Γ. For the nodes that were observed, i.e., those belonging to Γ, all the associated censoring probabilities are unrestricted beyond the obvious restrictions that certain sets of the censoring probabilities must sum to 1. Over this set we choose any continuous distribution which assigns positive mass to every open subset. The actual choice of this factor is not important since it will not affect the estimates of the various λ's. However, the first step of setting some of the censoring probabilities at 0 is important, because without this, the stepwise Bayes argument would not work.

It is now easy to check that for this Γ and for the prior over $\theta_\lambda \times \theta_p$ the resulting Bayes estimate is δ_H and that the admissibility of δ_H follows from the stepwise Bayes technique. This completes the proof in the special case when condition (iii) of (2.1) holds as well. □

We will now show how the case when $\{\alpha_1, \ldots, \alpha_k\}$ belongs to $\mathscr{U}$ follows from the special case just considered.

First of all, if we remove the set $\{\alpha_1, \ldots, \alpha_k\}$ from $\mathscr{U}$, then we can define θ_λ and θ_p as described in the preceding argument. Next we define the singleton node α_ϕ which denotes the outcome where we only learn that X was observed but nothing about its actual value. This node will neither precede nor be preceded by any other node. We will label it (1). Next we will introduce a node, labeled (0), which will precede all the other nodes except α_ϕ or (1). Furthermore, we will let $\lambda(0)$ be the probability we observe node (0) and $\lambda(\phi) = 1 - \lambda(0)$ be the probability we observe α_ϕ or (1). For example, in the model given in Figure 2 if we allow for the possibility that $\{\alpha_1, \ldots, \alpha_8\}$ is an outcome, then $\lambda(0)\lambda(2|0)\lambda(1|2, 0)$ is the probability that the model will generate the value α_2. We let θ_λ^* denote the parameter space for this new scheme with the two added nodes, i.e., it is just θ_λ augmented by $\lambda(0)$. In addition, under this new scheme the assumptions of the earlier special case are satisfied since $\{\alpha_\phi, \alpha_1, \ldots, \alpha_k\}$ is not a possible observation.

One possible objection to the above scheme is that the nodes (0) and (1) are really the same. This can be overcome if we take θ_p to be the parameter space for the censoring distributions. That is, we never allow the node (0) to be observed. Therefore, the proof of the theorem for the case when $\{\alpha_1, \ldots, \alpha_k\} \in \mathscr{U}$ follows from the special case if we introduce the above two special nodes and take as our parameter space $\theta_\lambda^* \times \theta_p$. Operationally, this means when computing the maximum likelihood estimator of a parameter of interest, other than $\lambda(0)$, we can just ignore the observations that are completely missing.

As far as we know the stepwise Bayes argument was first used in Johnson (1971) and named in Hsuan (1979). The argument given above is similar to one given in Alam (1979) and Brown (1981) for estimating multinomial probabilities without any censoring.

Rather than just estimating the vector λ we might wish to estimate some function of λ. One case of particular interest is to estimate for a given node $(i, j, \ldots, r, s)$ the probability assigned to this node under λ. This is given in (2.2) and we will denote it by $P_\lambda((i, j, \ldots, r, s))$.

If in the theorem we let

$$\delta_H(\Gamma)_{(i)} \text{ denote the estimate of } \lambda(i)$$

and

$$\delta_H(\Gamma)_{(s|i, j, \ldots, r)} \text{ denote the estimate of } \lambda(s|i, j, \ldots, r),$$

then a natural estimate of $P_\lambda((i, j, \ldots, r, s))$ is

$$\delta_H(\Gamma)_{(i, j, \ldots, r, s)} = \delta_H(\Gamma)_{(i)}\delta_H(\Gamma)_{(j|i)} \cdots \delta_H(\Gamma)_{(s|i, j, \ldots, r)}. \tag{3.7}$$

Assuming squared error loss, it is easy to check, using the same argument that proved Theorem 1, that $\delta_H(\Gamma)_{(i, j, \ldots, r, s)}$ is an admissible estimator. This follows from the independence in the prior distributions and the convenient form of the likelihood function in (2.5). Hence we have proved the following corollary.

COROLLARY. (i) *Let* $(i, j, \ldots, r, s)$ *be a given node. For estimating* $P_\lambda((i, j, \ldots, r, s))$, *as defined in* (2.2), *with squared error loss the estimator* $\delta_H(\Gamma)_{(i, j, \ldots, r, s)}$ *defined in* (3.7) *is admissible.*

(ii) *More generally, suppose we wish to estimate simultaneously* $P_\lambda((i, j, \ldots, r, s))$ *for* v *different nodes when the loss function is the sum of the individual squared error losses. Then the estimator that is the vector of the corresponding estimators from* (i) *is admissible.*

On any particular trial what the statistician observes depends on two things. The first is the underlying probability distribution described by λ and the second is the noninformative censoring mechanism described by p. Until now we have assumed that the censoring mechanism is the same for each trial. However,

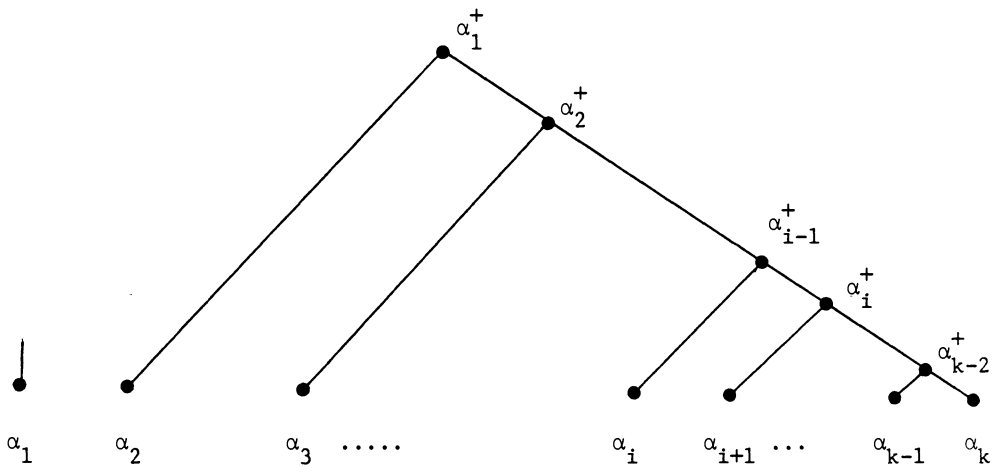

FIG. 3.

this is not necessary. One may let the censoring mechanism vary from trial to trial as long as for each trial it is assumed to be noninformative. The same argument proves the admissibility of the stepwise Bayes estimator considered above for this more general censoring scheme.

3.4. *A Kaplan–Meier type example.* The previous argument proves the admissibility of the MLE for various censoring mechanisms. It also follows, using the standard technique, that a unique Bayes estimator for this problem is admissible. We now wish to consider other estimators which incorporate prior information but are not actually Bayes estimators. A slight modification of the previous stepwise Bayes argument proves their admissibility as well.

What follows is generally true for any censoring mechanism satisfying (2.1). However, we will consider only one special case which is a generalization of Figure 1. As we will see in the next section this special case is closely related to the Kaplan–Meier estimator.

Let $\alpha_1 < \alpha_2 < \cdots < \alpha_k$ be given and consider the censoring mechanism described in Figure 3.

Rather than use the general notation developed earlier, we have named the nodes in an obvious way. For example, the node α_i^+ corresponds to a censored observation which is greater than or equal to α_{i+1}. (Note $\alpha_{k-1}^+ = \alpha_k$.) Let λ_1 denote the probability of α_1 and $1 - \lambda_1$ the probability of α_1^+. For $2 \leq i \leq k - 1$ let λ_i and $1 - \lambda_i$ be the conditional probabilities of α_i and α_i^+ given the node α_{i-1}^+.

Let $1 \leq u \leq k - 1$ be fixed and consider the problem of estimating $P_\lambda(\alpha_u^+)$, the probability that an observation from the underlying distribution is strictly greater than α_u. Given a Γ of size n it follows that the value of the estimator

given in the corollary for estimating $P_\lambda(\alpha_u^+)$ is

$$\prod_{i=1}^{u} \left[\bar{y}(\alpha_i^+)/(y(\alpha_i) + \bar{y}(\alpha_i^+)) \right]. \tag{3.8}$$

We will now modify this estimator by incorporating some prior information, by choosing for each i a weight $m_i \geq 0$. Let $m = (m_1, \ldots, m_k)$ denote this vector of weights, at least one of which is assumed to be nonzero. We can think of m_i as representing our prior beliefs about how likely an observation at α_i is. The larger we choose m_i the more probable we believe α_i to be. Another way to interpret the m_i's is to assume that the following imaginary experiment has been carried out. We assume that $\sum_{i=1}^{k} m_i$ independent observations were taken from the distribution parameterized by the λ's and that m_i of these turn out to be α_i. (Note that the m_i's need not be integers.) The results of this imaginary experiment should then be combined with the outcomes we observed when the n trials of the real experiment were observed. This combined "sample" suggests that

$$\prod_{i=1}^{u} \left[\bar{y}_m(\alpha_i^+)/(y_m(\alpha_i) + \bar{y}_m(\alpha_i^+)) \right] \tag{3.9}$$

is an admissible estimator of $P_\lambda(\alpha_u^+)$, where

$$y_m(\alpha_i) = y(\alpha_i) + m_i$$

and

$$\bar{y}_m(\alpha_i^+) = \bar{y}(\alpha_i^+) + \sum_{j=i+1}^{k} m_i.$$

The proof of the admissibility of the estimator in (3.9) is very similar to that of the estimator in (3.8) with one exception. That is, as the stepwise Bayes argument works through the possible samples one always uses the combined "sample," i.e., the actual outcomes combined with the imaginary ones. This is a straightforward modification which causes no difficulties.

For example, consider the censoring mechanism described in Figure 1 where the nodes, however, are labeled as in Figure 3. Suppose that $m_3 > 0$ is the only m_i for $i = 1, \ldots, 5$ which is greater than 0. In addition, suppose we have a sample of size 4 which resulted in $y(\alpha_1) = y(\alpha_2) = y(\alpha_1^+) = y(\alpha_2^+) = 1$. Now combining this with the imaginary sample of m_3 observations at α_3 we can proceed just as before.

4. Admissibility of the Kaplan–Meier estimator. We consider the following model. Let $X_1, X_2, \ldots, X_n$ be the true survival times of n individuals. We assume that $X_1, X_2, \ldots, X_n$ are iid with some common distribution function F on the interval $(0, \infty)$. These observations can be censored on the right however

by n follow-up times $W_1, \ldots, W_n$. The W_i's are assumed to be mutually independent and independent of the X_i's but they need not be identically distributed. In addition, we assume a value $T^* > 0$ such that the experiment must stop when time T^* is reached. Hence, each W_i has a distribution concentrated on $(0, T^*]$. When $X_i \leq W_i$ we observe the death time for the ith individual. When $X_i > W_i$ this individual is censored or lost to us and we only learn that the individual survived past time W_i.

F is assumed to be unknown and belong to the family of all cumulative distribution functions on $(0, \infty)$. For a given F we wish to estimate the survival function

$$S_F(u) = 1 - F(u) = P_F(X > u) \tag{4.1}$$

with loss function

$$L(\hat{S}, S_F) = \int_0^\infty [\hat{S}(u) - S_F(u)]^2 \, dw(u), \tag{4.2}$$

where w is some weight function on $(0, \infty)$.

For this problem the nonparametric product limit estimator was introduced in Kaplan and Meier (1958). We will show that this estimator is admissible. This, incidentally, disproves a suggestion made by Cornfield and Detre (1977) that the Kaplan–Meier estimator may indeed be inadmissible under squared error loss. However, since the Kaplan–Meier estimator of $S(u)$ is undefined for large values of u and when all the observations are censored we will have to first define it for all possible situations.

First, we choose r numbers $a_1 < a_2 < \cdots < a_r$ where $a_1 > T^*$. For a distribution function F which concentrates its mass on $(0, T^*]$ and the set $\{a_1, a_2, \ldots, a_r\}$ we define the parameters $\gamma_i(F) = \gamma_i$ for $i = 1, \ldots, r-1$. If X is distributed according to F, then we let

$$\gamma_i = \gamma_i(F) = P_F(X = a_i | X \geq a_i).$$

Since F is not known, for each i let G_i be a prior distribution for γ_i and $\tilde{\gamma}_i = \int_0^1 \gamma \, dG_i(\gamma)$. Hence, the statistician should choose the a_i's and G_i's to reflect his or her prior beliefs about the unknown $F(u)$ for $u > T^*$.

Now let Γ denote a typical set of outcomes for the experiment. For each individual i, Γ will contain either a death time or a censoring time. Assume that Γ contains at least one death time and at least one censoring time. Let $\alpha_1 < \alpha_2 < \cdots < \alpha_k$ be the k distinct death times that appear in Γ. Let $y(\alpha_i)$ be the number of deaths at time α_i. Let $\tau_1 < \tau_2 < \cdots < \tau_l$ denote the distinct censoring times that appear in Γ. Let $y(\alpha_i^+)$ be the number of censored observations with censoring times in the interval $[\alpha_i, \alpha_{i+1})$ for $i = 1, \ldots, k-1$ and $y(\alpha_k^+)$ be the number of censored observations falling in the interval $[\alpha_k, T^*]$. Note that $\sum_{i=1}^k y(\alpha_i) + \sum_{i=1}^k y(\alpha_i^+) = n$ if $\alpha_1 \leq \tau_1$. Finally, as before, let $\bar{y}(\alpha_i^+) =$

$\Sigma_{j=i+1}^{k} y(\alpha_j) + \Sigma_{j=i}^{k} y(\alpha_j^+)$. For such a Γ the estimate of $S(u)$ is given by

$$
\begin{aligned}
\delta_a(\Gamma) &= 1 \quad \text{for } 0 < u < \alpha_1, \\
&= \prod_{i=1}^{j} \frac{\bar{y}(\alpha_i^+)}{y(\alpha_i) + \bar{y}(\alpha_i^+)} \quad \text{for } \alpha_j \le u < \alpha_{j+1} \quad \text{and for } j = 1, \ldots, k-1, \\
&= \prod_{i=1}^{k} \frac{\bar{y}(\alpha_i^+)}{y(\alpha_i) + \bar{y}(\alpha_i^+)} \quad \text{for } \alpha_k \le u < T^*, \qquad (4.3) \\
&= \prod_{i=1}^{k} \frac{\bar{y}(\alpha_i^+)}{y(\alpha_i) + \bar{y}(\alpha_i^+)} \prod_{i=1}^{j} (1 - \tilde{\gamma}_i) \quad \text{for } a_j \le u < a_{j+1} \\
&\qquad\qquad \text{and for } j = 1, \ldots, r-1, \\
&= 0 \quad \text{for } u \ge a_r.
\end{aligned}
$$

There remain two special cases for which $\delta_a(\cdot)$ must be defined. The first is when Γ contains no censored observations. For this Γ, δ_a is defined as in (4.3) for $0 < u < \alpha_k$ and is defined to be 0 for $u \ge \alpha_k$. The second is when Γ contains no death times, i.e., all the observations were censored. For this Γ, δ_a is defined to be 1 for $0 < u < a_1$ and is defined as in (4.3) for $u > a_1$. The estimator is subscripted by a to indicate that it depends on the prior information incorporated in the a_i's and will be called the (a)-augmented Kaplan–Meier estimator.

THEOREM 2. *For estimating the survival function, given in* (4.1), *with the loss function given in* (4.2) *when* $F \in \mathcal{F}$, *the family of all distributions on* $(0,\infty)$, *the* (a)*-augmented Kaplan–Meier estimator given in* (3.3), *is admissible.*

The proof will be omitted since it follows easily from the results of the previous section and the argument given in Meeden, Ghosh and Vardeman (1985).

For some, an unappealing property of the Kaplan–Meier estimator is that it estimates a constant survival probability between successive death times. But it is exactly this property that allows the previous stepwise Bayes argument to prove its admissibility. In Susarla and Van Ryzin (1976) a family of nonparametric Bayesian estimators of the survival function was introduced. They were derived from a Dirichlet process prior and the estimate of the survival function was not constant between death times. In addition, the Kaplan–Meier estimator is a limit of such estimators.

We believe that these estimators of Susarla and Van Ryzin are admissible, but cannot prove that this is so. However, the estimator given in (3.9) is essentially a "discrete" version of the type considered by Susarla and Van Ryzin. In addition, the admissibility of the resulting estimator follows using the same argument that proved Theorem 2.

We will end by briefly indicating how the estimator in (3.9) is adapted to the problem of estimating a survival function. To do this, the statistician's prior

beliefs about the distribution function F both above and below the value T^* must be utilized. As with the Kaplan–Meier estimator, values $a_1 < a_2 < \cdots < a_r$, with $a_1 > T^*$, must be chosen along with the prior distributions G_i for the parameters γ_i which result in the values $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_{r-1}$. To represent his prior beliefs below T^* the statistician must choose values $0 < b_1 < b_2 < \cdots < b_s < T^*$ along with weights $m_i > 0$ for $i = 1, \ldots, s$. As in Section 3.4, the m_i's can be interpreted as the outcomes of an imaginary sample.

Consider now a sample Γ, for which at least one death time and at least one censoring time was observed for the n individuals. Let $\alpha_1' < \alpha_2' < \cdots < \alpha_{k'}'$, be the k' distinct death times which appear in Γ. Let $y(\alpha_i')$ be the number of deaths at time α_i'. Let $k = k' + s$ and assume for notational convenience that the set of α_i''s and b_i's are distinct. Let $\alpha_1 < \alpha_2 < \cdots < \alpha_k$ denote the collection of values of the α_i''s and b_i's arranged in increasing order. We define $y_m(\alpha_i)$ for $i = 1, \ldots, k$ as follows. If α_i is α_j', then $y_m(\alpha_i) = y(\alpha_j')$ and if α_i is b_j, then $y_m(\alpha_i) = m_j$. Let $\tau_1 < \tau_2 < \cdots < \tau_l$ denote the distinct censoring times that appear in Γ. Let $y_m(\alpha_i^+)$ be the number of censored observations with censoring times in the interval $[\alpha_i, \alpha_{i+1})$ for $i = 1, \ldots, k-1$ and $y_m(\alpha_k^+)$ be the number of censored observations falling in the interval $[\alpha_k, T^*]$. Just as before we let $\bar{y}_m(\alpha_i^+) = \sum_{j=i+1}^{k} y_m(\alpha_j) + \sum_{j=1}^{k} y_m(\alpha_j^+)$. Now if in (4.3) we replace $y(\alpha_i)$ and $\bar{y}(\alpha_i^+)$ with $y_m(\alpha_i)$ and $\bar{y}_m(\alpha_i^+)$ we have defined another estimator which we will denote by $\delta_{m,a}$. We will call this the (m, a)-augmented Kaplan–Meier estimator to indicate its dependence on the prior information incorporated through the use of the m_i's and the a_i's. It is easy to check that this is a discretized version of the estimator considered by Susarla and Van Ryzin and its admissibility follows from the previous section just as the admissibility of the (a)-augmented Kaplan–Meier estimator did.

When all the observations are uncensored the Kaplan–Meier estimator reduces to the sample distribution function. For this situation Cohen and Kuo (1985) have demonstrated admissibility under a more general loss function. Brown (1988) gives some additional admissibility and inadmissibility results when estimating a distribution function.

One advantage of giving the more general proof, rather than just concentrating on the Kaplan–Meier estimator, is that it highlights the special structure of the Kaplan–Meier estimator which allows admissibility to be proved. In particular, we have assumed that the censoring mechanism is independent of the underlying probability structure. It was noted by a referee that the Kaplan–Meier estimator can sometimes be used even when the censoring mechanism and the underlying probability structure are dependent. See, for example, Jacobsen (1986). It would be of interest to settle the admissibility question for these situations as well.

REFERENCES

ALAM, K. (1979). Estimation of multinomial probabilities. *Ann. Statist.* **7** 282–283.

BERGER, J. and WOLPERT, R. (1984). *The Likelihood Principle*. IMS, Hayward, Calif.

BROWN, L. D. (1981). A complete class theorem for statistical problems with finite sample space. *Ann. Statist.* **9** 1289–1300.

BROWN, L. D. (1988). Admissibility in discrete and continuous invariant nonparametric estimation problems and in their multinomial analogs. *Ann. Statist.* **16** 1567–1593.

COHEN, M. and KUO, L. (1985). The admissibility of the empirical distribution function. *Ann. Statist.* **13** 262–271.

CORNFIELD, J. and DETRE, K. (1977). Bayesian life table analysis. *J. Roy. Statist. Soc. Ser. B* **39** 86–94.

DICKEY, J., JEONG, J. and KADANE, J. (1987). Bayesian methods for censored categorical data. *J. Amer. Statist. Assoc.* **82** 773–781.

HSUAN, F. C. Y. (1979). A stepwise Bayesian procedure. *Ann. Statist.* **7** 860–868.

JACOBSEN, M. (1986). Right censoring and the Kaplan–Meier and Nelson–Aalen estimators. Technical Report, Inst. Math. Statist., Univ. Copenhagen.

JOHNSON, B. M. (1971). On the admissible estimator for certain fixed sample binomial problems. *Ann. Math. Statist.* **42** 1579–1587.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

MEEDEN, G. and GHOSH, M. (1981). Admissibility in finite problems. *Ann. Statist.* **9** 237–240.

MEEDEN, G., GHOSH, M. and VARDEMAN, S. (1985). Some admissible nonparametric and related finite population sampling estimators. *Ann. Statist.* **13** 811–817.

SUSARLA, V. and VAN RYZIN, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71** 897–902.

G. MEEDEN
DEPARTMENT OF STATISTICS
270 VINCENT HALL
UNIVERSITY OF MINNESOTA
206 CHURCH STREET, S.E.
MINNEAPOLIS, MINNESOTA 55455

M. GHOSH
DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32601

C. SRINIVASAN
DEPARTMENT OF STATISTICS
UNIVERSITY OF KENTUCKY
LEXINGTON, KENTUCKY 40506

S. VARDEMAN
DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
AMES, IOWA 50011