The Annals of Statistics
1985, Vol. 13, No. 2, 552–563

GOODNESS-OF-FIT TESTS FOR CENSORED SURVIVAL DATA[1]

BY ROBERT J. GRAY AND DONALD A. PIERCE

Dana-Farber Cancer Institute and Oregon State University

The problem of testing the underlying distribution in a general regression model is considered when the data are right censored. It is proposed that tests be based on efficient scores from a certain class of parametric alternatives. The connection between tests based on the scores and more traditional approaches using empirical distribution functions is explored. Neyman's smooth test is extended to allow for censored data.

1. Introduction. The problem considered here is that of testing the assumed form of the underlying probability distribution in generalized regression models for censored data. The goal is to provide a method which is applicable in a wide variety of particular models, and which avoids the difficult distributional problems usually encountered. This is accomplished, following Neyman (1937), by embedding the model in a larger parametric family and using score tests.

We assume the hypothesized model is of the following form. $T_1, \cdots, T_n$ are independent continuous failure times, with the distribution function of T_i given by $F(t; \lambda_i, \delta)$, where $\lambda_i = x_i'\beta$, the x_i are vectors of known covariates, and $\gamma = (\beta, \delta)$ is a vector of unknown parameters. We write $F_i(t; \gamma)$ for $F(t; \lambda_i, \delta)$, $\bar{F}_i$ for the survivor function $1 - F_i$, and f_i for the density. We will also assume that the censoring times V_i have distribution functions C_i. Generally only $Y_i = \min(T_i, V_i)$ and the indicator $Z_i = I(T_i \leq V_i)$ are observed.

Although it is difficult in practice to separate the issues, our interest is in testing the adequacy of the form of F, rather than in aspects relating to the adequacy of the covariables. For example, one may want to test the exponential model $F_i(t; \beta) = 1 - \exp\{-\exp(x_i'\beta)t\}$ under the *assumption* that scale factors of the form $\exp(x_i'\beta)$ represent adequately the systematic part of the regression. In the general setting, this problem can be specified as follows: the hypothesis to be tested is that the $U_i = F_i(T_i; \gamma)$ are independent and identically distributed with uniform distributions on (0, 1), with the alternative of interest being that the U_i are independent and identically distributed with some other distribution.

We consider alternative densities for the U_i of the form

$$\exp\{\textstyle\sum_{\ell=1}^{m} \theta_\ell \psi_\ell(u) - K(\theta)\},$$

where the ψ_ℓ are known functions and $K(\theta)$ is a normalizing constant. This gives

Received January 1983; revised January 1985.

[1] This work was conducted while the first author was at Oregon State University and was supported by PHS grants CA-27532 from the National Cancer Institute, and ES-0021 from the National Institute of Environmental Health Sciences, DHHS.

AMS 1980 *subject classifications.* Primary 62F03; secondary 62G30.

Key words and phrases. Goodness-of-fit tests, censored data, survival analysis, Neyman's smooth test, efficient scores, generalized residuals, empirical distribution functions.

a family of densities for T_i of the form

$$f_i(t; \gamma)\exp\{\textstyle\sum_{\ell=1}^{m} \theta_\ell \psi_\ell(F_i(t; \gamma)) - K(\theta)\}. \tag{1.1}$$

We emphasize that the family arising from (1.1), by taking different functions for the ψ_ℓ, is very general. The approach adopted here is to test the null hypothesis $\theta = 0$ using score tests. A primary point here is that applying the standard score test to (1.1) leads to a principle for handling censoring which is somewhat independent of the choice of ψ_ℓ. As discussed in the next section, this principle leads to methods which are substantially different from commonly used approaches.

For an omnibus goodness-of-fit test the idea is to focus on a general purpose choice of the ψ_ℓ, selected independently of the null hypothesis family, which would be suitable for a fairly rich class of alternatives. One such possibility is to take $\psi_\ell(u) = u^\ell$, $\ell = 1, \cdots, m$. In the uncensored case with no unknown parameters γ, the resulting score test is identical to the test introduced by Neyman (1937), which was designed to have good power against "smooth" alternatives. Thomas and Pierce (1979) used the score test approach to extend Neyman's test to allow for estimation of γ. In Section 3, we consider extending this test to the censored data case.

As another possibility, Thomas and Pierce noted that if the ψ_ℓ functions are taken to be indicator functions of specified intervals forming a partition of (0, 1), then the resulting score test is identical with the generalization of the classical Pearson chi-square test given by Rao and Robson (1974). Of course, more specialized possibilities for the ψ_ℓ functions could also be considered. For example, the score for the Weibull ($\bar{F}(t; \alpha, \sigma) = \exp\{-(t/\alpha)^\sigma\}$) alternative to the exponential ($\sigma = 1$) can be found from (1.1) with $m = 1$ and

$$\psi_1(u) = 1 + \log\{-\log(1 - u)\}\{1 + \log(1 - u)\}.$$

The major problem with applying more traditional approaches for testing the fit of the underlying distribution in the censored regression case is that the asymptotic distributions of test statistics will depend on the true values of the parameters, the covariates, and the generally unknown censoring distributions, even when the regression models are of the location-scale type. There seem to be virtually no general results on goodness-of-fit tests for this case available in the literature. It is possible, of course, to fit a more general parametric model containing the hypothesized model and use a likelihood ratio test; the generalized gamma model employed by Farewell and Prentice (1977) is notably flexible for this purpose. Even for the identically distributed case with censored data, there are only a few results available, mostly for specialized censoring models. See Smith and Bain (1976), Pettit (1976, 1977), Turnbull and Weiss (1978), Bargal (1981), Chen (1981), Habib (1981) and Mihalko and Moore (1980). Generally these results again reflect the dependence of the distributions on the true values of the parameters and the censoring distribution.

It is seen below that the distribution of the scores from (1.1) do not avoid this dependence. However, the asymptotic distribution of the scores is provided by

standard results, and instead of the complex distributional problems associated with other methods, we are only faced with the problem of carrying out the information calculations, which is complicated by their dependence on the generally unknown censoring distributions. In Section 3 this problem is considered in detail for the Neyman smooth test.

2. Score tests and their relationship to approaches based on empirical distribution functions. For independent and noninformative censoring the log-likelihood of the data under the alternative (1.1) is

$$\ell(\gamma, \theta) = \sum_{i=1}^{n} \Big\{ z_i(\log f_i(y_i; \gamma) + \sum_{\ell=1}^{m} \theta_\ell \psi_\ell\{F_i(y_i; \gamma)\}) + (1 - z_i)\log \int_{F_i(y_i;\gamma)}^{1} \exp(\sum_{\ell=1}^{m} \theta_\ell \psi_\ell(u))\, du - K(\theta) \Big\} \tag{2.1}$$

(see Kalbfleisch and Prentice, 1980, Section 5.2). An additional term involving the censoring distributions, which does not depend on the parameters, has been omitted from (2.1). We emphasize that this is an ordinary likelihood from a fully parametric model, unlike the partial likelihoods arising from the partially nonparametric models currently popular for survival analysis, so that classical likelihood results apply here. The scores from (2.1) are given by

$$\begin{aligned} &\partial\ell(\hat\gamma_0, 0)/\partial\theta_\ell \\ &= \sum_{i=1}^{n} \Big\{ z_i\psi_\ell(\hat u_i) + (1 - z_i) \int_{\hat u_i}^{1} \psi_\ell(u)\, du/(1 - \hat u_i) - \partial K(0)/\partial\theta_\ell \Big\} \qquad (2.2) \\ &= \sum_{i=1}^{n} \{z_i\psi_\ell(\hat u_i) + (1 - z_i)E(\psi_\ell(U) \mid U > \hat u_i) - E(\psi_\ell(U))\}, \end{aligned}$$

where U has a uniform distribution on (0, 1), $\hat\gamma_0$ is the maximum likelihood estimate for γ when $\theta = 0$, and $\hat u_i = F_i(y_i; \hat\gamma_0)$.

Uncensored observations contribute a term of $\psi_\ell(F_i(t_i; \hat\gamma_0))$ to (2.2). For censored observations, this term is replaced by its conditional expectation

$$E_{H_0}\{\psi_\ell(F_i(T_i; \gamma)) \mid T_i > y_i\}\big|_{\gamma=\hat\gamma_0}.$$

Thus (2.2) is the expectation of the score for uncensored data conditional on the observed data. This is similar to the idea underlying the EM algorithm of Dempster, Laird and Rubin (1977). This feature of (2.2), that the contribution of the censored observations is a conditional expectation, holds quite generally for score tests with censored data; it is not a special property of the family (1.1), but rather is a general property of the parametric likelihood approach adopted here.

Applying standard asymptotic results (see, e.g., Cox and Hinkley, 1974, pages 323–324) to the scores (2.2) gives that under the null hypothesis, subject to mild regularity conditions,

$$n^{-1/2}\partial\ell(\hat\gamma_0, 0)/\partial\theta \to_{\mathcal{L}} N(0, I_{\theta\theta|\gamma}), \tag{2.3}$$

where $\to_{\mathcal{L}}$ denotes convergence in a distribution and $I_{\theta\theta|\gamma}$ is the adjusted average

expected Fisher information. That is, $I_{\theta\theta|\gamma} = I_{\theta\theta} - I_{\theta\gamma}I_{\gamma\gamma}^{-1}I_{\gamma\theta}$, where for example

$$I_{\theta_\ell\theta_k} = \lim_{n\to\infty} I_{\theta_\ell\theta_k}^{(n)} \quad \text{and} \quad I_{\theta_\ell\theta_k}^{(n)} = n^{-1}\mathrm{Cov}_{H_0}(\partial\ell(\gamma, 0)/\partial\theta_\ell, \partial\ell(\gamma, 0)/\partial\theta_k).$$

Although sufficient conditions are complex to state, (2.3) is just the standard asymptotic result for scores with independent but not identically distributed data, and so will hold quite generally for common distributions provided the ψ_ℓ functions are reasonably smooth. See also the conditions given by Borgan (1983), who uses a martingale approach to derive asymptotic likelihood results for censored data. The expected information terms under the null hypothesis can be expressed

$$nI_{\theta_\ell\theta_k}^{(n)} = \sum_{i=1}^n \int s_{i\ell}(y; \gamma)s_{ik}(y; \gamma)\bar{C}_i(y)\, dF_i(y; \gamma), \tag{2.4}$$

$$nI_{\theta_\ell\gamma_j}^{(n)} = \sum_{i=1}^n \int s_{i\ell}(y; \gamma)q_{ij}(y; \gamma)\bar{C}_i(y)\, dF_i(y; \gamma), \tag{2.5}$$

and

$$nI_{\gamma_j\gamma_k}^{(n)} = \sum_{i=1}^n \int q_{ij}(y; \gamma)q_{ik}(y; \gamma)\bar{C}_i(y)\, dF_i(y; \gamma), \tag{2.6}$$

where

$$s_{i\ell}(y; \gamma) = \psi_\ell(F_i(y; \gamma)) - E(\psi_\ell(U) \mid U > F_i(y; \gamma)),$$

$$q_{ij}(y; \gamma) = \partial \log(f_i(y; \gamma)/\bar{F}_i(y; \gamma))/\partial\gamma_j,$$

and as before U has a uniform distribution on (0, 1). This form for the expected information components is particularly convenient for survival models, since they are often formulated in terms of the hazard function $f_i/\bar{F}_i$, which appears in q_{ij}.

The tests we consider are quadratic score statistics of the form

$$Q_m = n^{-1}(\partial\ell(\hat{\gamma}_0, 0)/\partial\theta)'\,\hat{V}^{-1}(\partial\ell(\hat{\gamma}_0, 0)/\partial\theta)$$

where $\hat{V}$ is a suitable estimate of $I_{\theta\theta|\gamma}$. From (2.3), $Q_m \to_{\mathcal{L}} \chi_m^2$ when $\hat{V}$ is consistent, under the null hypothesis. The primary remaining difficulty is the choice of $\hat{V}$, which is discussed in Section 3. Generally we have found that the only reliable method is to use the expected information in the sample, (2.4)–(2.6). Although these formulas appear quite complex, at worst they require numerical evaluation of univariate integrals, and in many situations (see the Appendix) closed form expressions can be given.

We will now show that there is a connection between tests based on the scores (2.2) and more general approaches using empirical distribution functions. In the uncensored case, these latter approaches can be thought of as follows. Let $\hat{\gamma}$ be an estimator of γ, define "generalized residuals" $\hat{u}_i = F_i(t_i; \hat{\gamma})$, and denote by $\hat{H}_n$ the empirical distribution function of $\hat{u}_1, \cdots, \hat{u}_n$. Then goodness-of-fit statistics are functionals of the stochastic process

$$\hat{y}_n(u) = \sqrt{n}(\hat{H}_n(u) - u), \quad 0 \le u \le 1 \tag{2.7}$$

the functional chosen to measure the "distance" of $\hat{H}_n$ from the uniform distri-

bution function. Basic theory for this approach in the regression case was given the Pierce and Kopecky (1979) and Loynes (1980).

In the censored case, the quantities $\hat{u}_i = F_i(y_i; \hat{\gamma})$ are possibly censored values of $F_i(t_i; \hat{\gamma})$, and the standard approach has been to replace $\hat{H}_n$ by the Kaplan-Meier (1958) estimate based on these, for use in (2.7). The scores (2.2) however, are functionals of a different estimate of the distribution of the u_i, which arises naturally from the parametric likelihood approach used here.

For complete samples, $\hat{H}_n$ assigns a mass of $1/n$ to each failure time. For a censored observation, we have no information about the true failure time except that it is greater than the observed censoring time, so the mass of $1/n$ for a censored observation is only known to lie somewhere in the interval $(\hat{u}_i, 1)$. The Kaplan-Meier estimator can be thought of here as distributing this mass in a certain nonparametric way over the uncensored values greater than $\hat{u}_i$ (see Efron, 1967). The score test approach leads to distributing this mass uniformly over $(\hat{u}_i, 1)$. This is equivalent in the time scale to distributing the mass for a censored observation y_i proportionally in t to the conditional survival function $\bar{F}_i(t; \hat{\gamma})/\bar{F}_i(y_i; \hat{\gamma})$. This idea is also discussed briefly by Aitkin and Clayton (1980), although our motivation for this came from the score test development given here, independently of their work. The result is to use in place of $\hat{H}_n$ in (2.7),

$$\hat{H}_n^*(u) = n^{-1} \sum_{i=1}^n I(u \geq \hat{u}_i)\{z_i + (1 - z_i)(u - \hat{u}_i)/(1 - \hat{u}_i)\}. \tag{2.8}$$

Defining $\hat{y}_n^*$ to be the process (2.7) with $\hat{H}_n^*$ replacing $\hat{H}_n$, the scores can then be expressed as functionals of the form $\sqrt{n} \int_0^1 \psi(u)\, d\hat{y}_n^*(u)$. This follows since

$$\begin{aligned} &\sqrt{n} \int_0^1 \psi_\ell(u)\, d\hat{y}_n^*(u) \\ &\quad = n\left\{\int_0^1 \psi_\ell(u)\, d\hat{H}_n^*(u) - \int_0^1 \psi_\ell(u)\, du\right\} \\ &\quad = \sum_{i=1}^n \{z_i \psi_\ell(\hat{u}_i) + (1 - z_i)E(\psi_\ell(U) \mid U > \hat{u}_i) - E(\psi_\ell(U))\}, \end{aligned} \tag{2.9}$$

which is the same as (2.2). From (2.9) we see that the scores compare $\hat{H}_n^*$ to the uniform distribution by taking the difference of the expectations of the functions ψ_ℓ under the two distributions. It should be noted that this representation for the scores implies that the score for any sort of distributional alternative, for example the Weibull alternative to the exponential, will be a functional of $\hat{H}_n^*$, not of the Kaplan-Meier estimator. Further properties of $\hat{H}_n^*$ are discussed in Gray and Pierce (1984).

3. Neyman's smooth test. In this section we extend the Neyman smooth test to allow for censored data. As mentioned in Section 1, this test comes from taking $\psi_\ell(u) = u^\ell, \ell = 1, \cdots, m$, in the scores. The results of Thomas and Pierce (1979) and Kopecky and Pierce (1979) in the identically distributed, uncensored case indicated that $m = 2$ is a good choice and that the resulting test is competitive, in terms of power, with other approaches. The difficulty in extending this test to the censored data case is that the information matrix depends on the

censoring distributions. Below we try several methods for handling this dependence, examining the methods by numerically evaluating the size of the Neyman smooth test for exponentiality in the identically distributed case. We then apply the test to an example of a censored regression data set, and use simulations to examine the size and the power of the test in the context of this example.

It is simpler here to take $U_i = \bar{F}_i(T_i; \gamma) = 1 - F_i(T_i; \gamma)$ in defining the alternatives (1.1). This change will not affect the ultimate test statistics, although there will be slight changes in the development. With the functions $\psi_\ell(x) = x^\ell$, the scores for the Neyman smooth test then are

$$\begin{aligned} &\partial \ell(\hat{\gamma}_0, 0)/\partial \theta_\ell \\ &\quad = \textstyle\sum_{i=1}^n \{z_i \bar{F}_i^\ell(y_i; \hat{\gamma}_0) + (1 - z_i)\bar{F}_i^\ell(y_i; \hat{\gamma}_0)/(\ell + 1) - (\ell + 1)^{-1}\} \end{aligned} \tag{3.1}$$

since $E(U^\ell \mid U < k) = k^\ell/(\ell + 1)$ and $E(U^\ell) = (\ell + 1)^{-1}$. Also formulas (2.4) and (2.5) for the expected information will now have $s_{i\ell}(y; \gamma) = \ell \bar{F}_i^\ell(y; \gamma)/(\ell + 1)$.

We will use the expected information in the sample, (2.4)–(2.6), to estimate the variance of the scores. We considered using several other approaches, and although we did not exhaust all possibilities, the methods we did try performed poorly here. One possibility we considered was to use the observed information in place of the expected information. This approach fails without some modification for the surprising reason that, even when sampling under the null hypothesis, the observed information matrix often fails to be positive definite at $(\hat{\gamma}_0, 0)$. This presumably is because the likelihood as a function of (γ, θ) is not concave everywhere, and $(\hat{\gamma}_0, 0)$ is not the global maximum of the likelihood, even if the model is true. We also tried using jackknife and bootstrap methods to get a nonparametric estimate of the variance of the scores for use in place of the expected information. The size of the test based on the normal approximation tended to be much larger than the nominal level in moderate sized samples when these estimates were used. For more details on these points see Gray and Pierce (1984).

We consider three general approaches for handling the dependence of the expected information on the censoring: (a) assuming a parametric model for the censoring, with parameters possibly estimated from the data, (b) assuming the censoring is homogeneous, at least within specified groups, and using a nonparametric estimate of its distribution, and (c) in the special case where the potential censoring times v_i are known for all observations, evaluate the expected information conditional on the v_i by taking $C_i(v) = I(v \geq v_i)$ in (2.4) to (2.6). This last approach avoids distributional assumptions about the censoring, and the conditional reference set it provides may be appropriate even if the censoring is known to be homogeneous.

Table 1 contains the results of simulations to examine the performance, in terms of size of the test, of these three methods for evaluating the expected information in the Neyman smooth test for exponentiality. Throughout this section, the quadratic score statistics for the Neyman smooth test of exponentiality with $m = 1$ and $m = 2$ will be denoted by W_1 and W_2, respectively. The simulations were limited to the identically distributed, homogeneous censorship

TABLE 1
Empirical sizes (in %) of the W_1 and W_2 tests with the expected information computed using three different approaches

	Nominal level	Conditional		Homogeneous exponential		Kaplan-Meier	
		W_1	W_2	W_1	W_2	W_1	W_2
$\sigma = 1$	10% (.95)	9.4	9.0	9.9	8.8	9.8	8.4
$p_c = .2$	5% (.69)	4.2	4.1	4.1	4.3	4.3	4.4
$\sigma = 1$	10%	10.5	9.4	9.8	8.9	9.1	8.2
$p_c = .5$	5%	5.1	5.0	4.7	4.7	4.1	4.8
$\sigma = 2$	10%	11.3	10.7	9.3	7.7*	11.3	9.6
$p_c = .2$	5%	6.5*	5.1	4.8	3.5*	6.1	4.7
$\sigma = 2$	10%	10.8	11.0	4.5*	4.3*	10.6	9.0
$p_c = .5$	5%	6.0	6.1	1.9*	2.2*	5.3	4.6
$\sigma = .5$	10%	7.0*	7.9*	8.2	9.7	7.1*	7.6*
$p_c = .2$	5%	3.9	3.0*	4.6	4.1	3.7	3.3*
$\sigma = .5$	10%	9.9	7.7*	15.3*	18.1*	9.6	8.0*
$p_c = .5$	5%	4.6	3.9	9.0*	11.4*	4.4	4.0

* The empirical size is more than 2 standard errors from the nominal level.

case, with failures generated from a unit exponential distribution and censoring times from Weibull distributions with $\bar{C}(v) = \exp\{-(v/\alpha)^\sigma\}$. The values of the Weibull shape parameter used were $\sigma = 1$, 0.5, and 2. For each value of the shape parameter, values of α were selected to make the probability of an observation being censored, p_c, equal to 0.2 and 0.5. In each of the 6 cases (3 values of $\sigma \times 2$ values of p_c), we generated 1000 samples of size 50. In each case, the 6 statistics (W_1 and W_2 with the expected information computed using the 3 different approaches) were computed from the same samples. The specific approaches used in computing the expected information were the "homogeneous exponential," where the censoring is assumed to follow an exponential distribution with the scale parameter estimated from the data; the "Kaplan-Meier," where the censoring is assumed to be homogeneous with the distribution estimated with the Kaplan-Meier estimator; and the "conditional," where the expected information is computed conditional on the actual potential censoring times. Details of the expected information calculations for the Kaplan-Meier and conditional approaches are given in the Appendix.

In Table 1 binomial standard errors are given in parentheses after the nominal levels in the first case. The same standard errors can be used in the other cases. The results indicate that in all cases the Kaplan-Meier and conditional approaches performed well. When the censoring is not exponential ($\alpha = 0.5$ or 2) and $p_c = 0.5$, use of the homogeneous exponential censoring model can be quite bad. However, when $p_c = 0.2$ there seems to be little effect.

Next we turn to an example data set taken from Glasser (1967). The data are

days survived after surgery for 131 lung cancer patients, with 66 of the observations censored. Two of the times are 0. These were replaced by the value 0.1 to facilitate the analysis. The patients are in two groups: group one patients have "low" vital capacity/predicted vital capacity ratios, and group two patients have "high" ratios. The age of the patients is also given as a continuous covariate. Throughout the discussion of this example, we will assume the censoring is homogeneous and estimate its distribution with the Kaplan-Meier estimator when calculating the expected information. Formulas for the expected information are given in the Appendix. We first fit an exponential model of the form $\bar{F}_i(t; \beta) = \exp\{-t \cdot \exp(x_i' \beta)\}$. In this model both age and group were significant. The values of the Neyman smooth test for this example were $W_1 = 4.95$ and $W_2 = 5.06$. The p-value for W_1 is approximately 0.03, so the Neyman smooth test gives significant evidence of lack of fit.

In addition to the tests, we also use graphical methods to examine lack of fit, as discussed by Aitkin and Clayton (1980). If we compute the estimate $\hat{H}_n^*$, given by (2.8), then a plot of $\log(1 - \hat{H}_n^*(1 - e^{-t}))$ against t or a plot of $\log\{-\log(1 - \hat{H}_n^*(1 - e^{-t}))\}$ against $\log(t)$ should lie roughly on the straight line through the origin with slope 1, when the model is true. Because of the censoring, there will, in general, be little information about lack of fit in the right tail, so we will use the second of these plots. Note that if we fit an exponential model, and the true model is a corresponding Weibull model, then the plot should lie roughly on a straight line through the origin with slope equal to the Weibull shape parameter.

The plot from the fit of the exponential model is given in Figure 1. Examination of this indicates a Weibull model may fit better. This is confirmed in Figure 2, which contains the plot from the fit of the Weibull model, $\bar{F}_i(t; \gamma) = \exp\{-(t \cdot \exp(x_i' \beta))^{\sigma}\}$, with the same covariables as used in the exponential model. The values of the Neyman smooth test for the Weibull model were 0.78 for the 1 degree of freedom test and 3.19 for the 2 degree of freedom test.

The disturbing fact about this example is that the likelihood ratio test for the

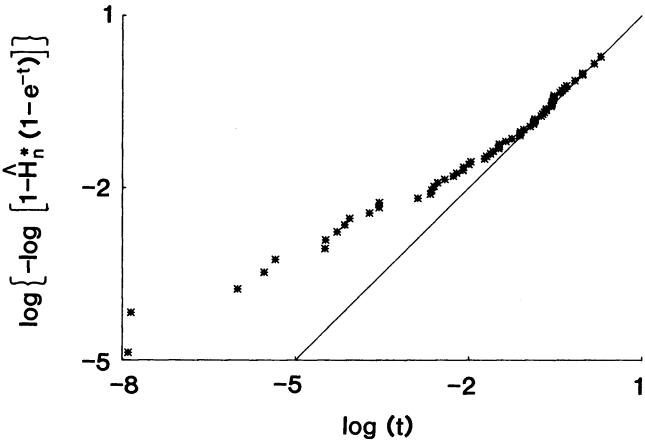

FIG. 1. *Plot of a modified empirical distribution function from the exponential model.*

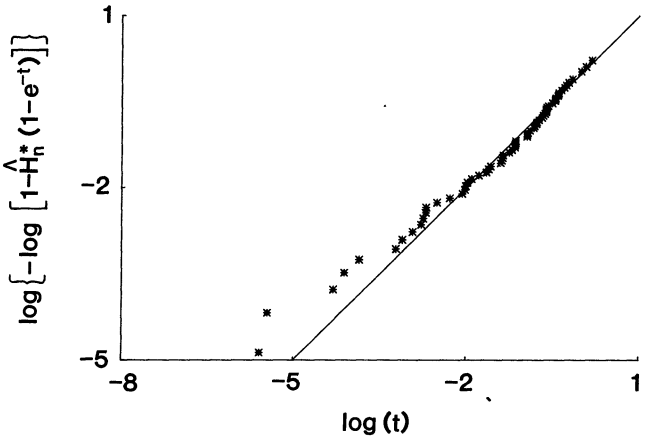

FIG. 2. *Plot of a modified empirical distribution function from the Weibull model.*

exponential model versus the Weibull model is 12.96 on 1 degree of freedom. In the uncensored, identically distributed case, Kopecky and Pierce (1979) give the local asymptotic relative efficiency of W_1 relative to the efficient score test for the Weibull departure from the exponential as 0.876, so it is surprising that here there should be such a large discrepancy between the Neyman smooth test and this particular likelihood ratio test.

The Neyman smooth test is known to perform well with uncensored data. In view of the above discrepancy, further numerical investigation into the performance of the test was called for. To keep as close as possible to the situation in the above example, we chose to simulate data from the fitted exponential and Weibull models, although neither is probably the true distribution underlying the observed data there. Thus, to examine the size of the Neyman smooth test for exponentiality, we generated censored samples of size 131 from the exponential regression model, using the covariates of the original data, taking the "true" parameter values to be the maximum likelihood estimates from the fit of the exponential model to the original data, and using the Kaplan-Meier estimate of the censoring distribution from the original data as the censoring distribution. For each of the 500 generated samples, we fit an exponential model and computed the values of W_1 and W_2. For comparison, we also fit a Weibull model and computed the likelihood ratio test for the exponential model versus the Weibull model. To examine the power, we repeated the procedure, only generating the sample from the estimated Weibull model from the original data (shape parameter = 0.693). The results are given in Table 2. Standard errors are given in parentheses after the entries. The results seem to indicate that the size is adequate and that the power of the Neyman smooth test for this Weibull alternative is fairly good.

Although we have not performed a comprehensive study of the Neyman smooth test with censored data, we believe overall the results of this section are quite encouraging. We have demonstrated that there are methods for handling the

TABLE 2
Empirical size and power against a Weibull (Shape = .693) alternative of the Neyman smooth test for exponentiality and the likelihood ratio test in the setting of the example data set

	$W1$		$W2$		Likelihood ratio test	
Nominal Level	10%	5%	10%	5%	10%	5%
Size	9.0 (1.3)*	4.2 (0.9)	10.2 (1.4)	6.4 (1.1)	8.8 (1.3)	5.0 (1.0)
Power	86.8 (1.5)	81.2 (1.7)	87.4 (1.5)	79.2 (1.8)	96.6 (0.8)	91.6 (1.2)

* The value in parentheses is the binomial standard error of the entry.

dependence on the censoring distributions which are computationally feasible and maintain the size of the test. Although we have only given closed form expressions for the exponential and Weibull distributions, this is possible for some other distributions as well, and at worst one would be faced with numerical evaluation of univariate integrals. In the case where we have examined power, the performance of the Neyman smooth test relative to the likelihood ratio test seems consistent with the results of Kopecky and Pierce (1979) for the uncensored case. We expect that a similar relationship will hold with other alternatives, so that the Neyman smooth test with censored data will be a reasonably powerful omnibus test, as it is in the uncensored case.

APPENDIX

In this Appendix, expressions are given for the components of the expected information in the Neyman smooth tests for the exponential and Weibull distributions (i) assuming a homogeneous censoring distribution and estimating it with the Kaplan-Meier estimator and (ii) conditioning on potential censoring times for all observations.

Let $y_1^c, \cdots, y_{n_c}^c$ denote the values of the censored observations, where

$$n_c = \sum_{i=1}^{n} (1 - z_i)$$

is the number of censored observations. Let $\Delta\hat{C}(y_h^c)$ denote the jump in the Kaplan-Meier estimate of the censoring distribution at y_h^c and $\hat{C}(\infty)$ the value of the Kaplan-Meier estimate of the survivor function beyond the largest censoring time. Then from (2.4)–(2.6) and (3.1), the components of the expected information using the Kaplan-Meier estimate of the homogeneous censoring distribution are

$$nI^{(n)}_{\theta_\ell\theta_k} = k\ell[(k+1)(\ell+1)(k+\ell+1)]^{-1} \sum_{i=1}^{n} \{1 - \sum_{h=1}^{n_c} \Delta\hat{C}(y_h^c)\bar{F}_i^{k+\ell+1}(y_h^c; \gamma)\},$$

$$nI^{(n)}_{\theta_\ell\gamma_j} = \ell(\ell+1)^{-1} \sum_{i=1}^{n} \{\hat{\bar{C}}(\infty)P_{ij\ell}(\infty) + \sum_{h=1}^{n_c} \Delta\hat{C}(y_h^c)P_{ij\ell}(y_h^c)\},$$

and

$$nI^{(n)}_{\gamma_j\gamma_k} = \sum_{i=1}^{n} \{\hat{\bar{C}}(\infty)Q_{ijk}(\infty) + \sum_{h=1}^{n_c} \Delta\hat{C}(y_h^c)Q_{ijk}(y_h^c)\},$$

where

$$P_{ij\ell}(w) = \int_0^w q_{ij}(y;\gamma)\bar{F}_i^{\ell}(y;\gamma)\, dF_i(y;\gamma)$$

and

$$Q_{ijk}(w) = \int_0^w q_{ij}(y;\gamma)q_{ik}(y;\gamma)\, dF_i(y;\gamma).$$

If $v_1, \cdots, v_n$ denote the potential censoring times, then the components of the expected information under the conditional approach are

$$nI^{(n)}_{\theta_\ell\theta_k} = k\ell[(k+1)(\ell+1)(k+\ell+1)]^{-1}\sum_{i=1}^n \{1 - \bar{F}_i^{k+\ell+1}(v_i;\gamma)\},$$

$$nI^{(n)}_{\theta_\ell\gamma_j} = \ell(\ell+1)^{-1}\sum_{i=1}^n P_{ij\ell}(v_i),$$

and

$$nI^{(n)}_{\gamma_j\gamma_k} = \sum_{i=1}^n Q_{ijk}(v_i).$$

For the exponential model with $\bar{F}_i(t;\gamma) = \exp\{-t\cdot\exp(x_i'\beta)\}$, we have

$$P_{ij\ell}(y) = x_{ij}[1 - \exp\{-y(\ell+1)\exp(x_i'\beta)\}]/(\ell+1)$$

$$Q_{ijk}(y) = x_{ij}x_{ik}[1 - \exp\{-y\cdot\exp(x_i'\beta)\}].$$

For the Weibull model with $\bar{F}_i(t;\gamma) = \exp\{-a_i(t)\}$, where

$$a_i(t) = [t\cdot\exp(x_i'\beta)]^{\sigma}, \quad \gamma_i = \beta_i, \quad i = 1, \cdots, p, \quad \text{and} \quad \gamma_{p+1} = \sigma,$$

we have

$$P_{ij\ell}(w) = \sigma x_{ij}[1 - \exp\{-(\ell+1)a_i(w)\}]/(\ell+1), \quad j = 1, \cdots, p,$$

$$P_{i,p+1,\ell}(w) = [\sigma(\ell+1)]^{-1}\{[1 - \log(\ell+1)][1 - \exp\{-(\ell+1)a_i(w)\}] + \Gamma_1[(\ell+1)a_i(w)]\},$$

$$Q_{ijk}(w) = \sigma^2 x_{ij}x_{ik}[1 - \exp\{-a_i(w)\}], \quad j, k = 1, \cdots, p,$$

$$Q_{ij,p+1}(w) = Q_{i,p+1,j}(w) = x_{ij}\{[1 - \exp\{-a_i(w)\}] + \Gamma_1(a_i(w))\}, \quad j = 1, \cdots, p,$$

and

$$Q_{i,p+1,p+1}(w) = \sigma^{-2}\{1 - \exp\{-a_i(w)\} + 2\Gamma_1[a_i(w)] + \Gamma_2[a_i(w)]\},$$

where $\Gamma_h(x) = \int_0^x [\log(u)]^h e^{-u}\, du$. The functions Γ_h are derivatives of the incomplete gamma function $\Gamma(w, \alpha) = \int_0^w t^{\alpha-1}e^{-t}\, dt$. In particular, $\Gamma_1(w) = \partial\Gamma(w, 1)/\partial\alpha$ and $\Gamma_2(w) = \partial^2\Gamma(w, 1)/\partial\alpha^2$. A general algorithm for evaluating derivatives of the incomplete gamma function has been given by Lindstrom (1981).

REFERENCES

AITKIN, M. and CLAYTON, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist.* **29** 156–163.

BARGAL, A. I. (1981). Efficiency comparisons of goodness-of-fit tests for Weibull and gamma distributions with singly censored data. Ph.D. thesis, Dept. of Statist., Oregon State University.

BORGAN, O. (1983). Maximum likelihood estimation in a parametric counting process model, with applications to censored failure time data and multiplicative models. Technical Report No. 18, Dept. of Math. and Statist., Agricultural University of Norway.

CHEN, C. (1981). Correlation-type goodness-of-fit tests for randomly censored data. Technical Report No. 73, Division of Biostatist., Stanford University.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B.* **39** 1–38.

EFRON, B. (1967). The two-sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 831–853.

FAREWELL, V. T. and PRENTICE, R. L. (1977). A study of distributional shape in life testing. *Technometrics* **19** 69–76.

GLASSER, M. (1967). Exponential survival with covariance. *J. Amer. Statist. Assoc.* **62** 561–568.

GRAY, R. J. and PIERCE, D. A. (1984). Goodness-of-fit tests for censored survival data. Technical Report No. 95, Dept. of Statist., Oregon State University.

HABIB, M. G. (1981). A chi-square goodness-of-fit test for censored data. Ph.D. thesis, Dept. of Statist., Oregon State University.

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

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

KOPECKY, K. J. and PIERCE, D. A. (1979). Efficiency of smooth goodness-of-fit tests. *J. Amer. Statist. Assoc.* **74** 393–397.

LINDSTROM, F. T. (1981). An algorithm for the evaluation of the incomplete gamma function and the first two partial derivatives with respect to the parameter. *Comm. Statist.—Simulation Comput.* **B10** 465–478.

LOYNES, R. M. (1980). The empirical distribution function of residuals from generalized regression. *Ann. Statist.* **8** 285–298.

MIHALKO, D. P. and MOORE, D. S. (1980). Chi-square tests of fit for type II censored data. *Ann. Statist.* **8** 625–644.

NEYMAN, J. (1937). "Smooth test" for goodness of fit. *Skan. Aktuaretidskr.* **20** 150–199.

PETTIT, A. N. (1976). Cramér-von Mises statistics for testing normality with censored samples. *Biometrika* **63** 475–481.

PETTIT, A. N. (1977). Tests for the exponential distribution with censored data using Cramér-von Mises statistics. *Biometrika* **64** 629–632.

PIERCE, D. A. and KOPECKY, K. J. (1979). Testing goodness of fit for the distribution of errors in regression models. *Biometrika* **66** 1–5.

RAO, K. C. and ROBSON, D. S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. *Comm. Statist.* **3** 1139–1153.

SMITH, R. M. and BAIN, L. J. (1976). Correlation type goodness-of-fit statistics with censored sampling. *Commun. Statist.—Theory Methods* **A5** 115–132.

THOMAS, D. R. and PIERCE, D. A. (1979). Neyman's smooth goodness-of-fit test when the hypothesis is composite. *J. Amer. Statist. Assoc.* **74** 441–445.

TURNBULL, B. W. and WEISS, L. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics* **34** 367–375.

DANA-FARBER CANCER INSTITUTE
44 BINNEY STREET
BOSTON, MASSACHUSETTS 02115

OREGON STATE UNIVERSITY
DEPARTMENT OF STATISTICS
CORVALLIS, OREGON 97331