TESTING TO DETERMINE THE UNDERLYING DISTRIBUTION
USING RANDOMLY CENSORED DATA

by

Myles Hollander and Frank Proschan

FSU Statistics Report M458
AFOSRN Technical Report No. 6
AFOSR Technical Report No. 79

April, 1978
The Florida State University
Department of Statistics
Tallahassee, Florida 32306

Research sponsored by the Air Force Office of Scientific Research, AFSC, USAF, under Grants AFOSR 76-3109 and AFOSR 74-2581D. The United States Government is authorized to reproduce and distribute reprints for governmental purposes.
Testing to Determine the Underlying Distribution Using Randomly Censored Data

by

Myles Hollander and Frank ProsCHAN

SUMMARY

For right-censored data, we develop a goodness-of-fit procedure for whether the underlying distribution is a specified function $G$. Our test $C$ is the one-sample limit of Efron's (1967) two-sample statistic $W$. The test based on $C$ is compared with recently proposed competitors due to Koziol-Green (1976) and Hyde (1977). The comparisons are on the basis of (i) applicability of competitors, (ii) the extent to which the censoring distribution affects the inferential procedure, and (iii) power. It is shown that in certain situations the $C$ test compares favorably with the tests of Koziol-Green and Hyde.

Some Key Words: Goodness-of-fit test; Kaplan-Meier estimator; Right-censored
1. INTRODUCTION

In the classical non-censored one-sample goodness-of-fit problem, or a random sample \( X_1, \ldots, X_n \) from a population with distribution function \( F(x) = P(X \leq x) \); the corresponding survival function is \( F(x) = P(X > x) \). The null hypothesis asserts that \( F(x) = G(x) \), where \( G \) is completely specified. The need to generalize this problem to encompass censored data arises because some situations, such as clinical trials, or other types of events, such as the occurrence of an end-point event before all patients, or items on test, have experienced the event. In the trials context the end-point event could, for example, be relapse, pregnancy death. In the life-testing framework, the end-point event could be failure or death. In these cases the observations can be viewed as pairs \( (Z_i, \delta_i) \), \( i = 1, \ldots, n \), where

\[
Z_i = \min(X_i, T_i),
\]

\[
\delta_i = \begin{cases} 
1 & \text{if } Z_i = X_i \text{ (i\text{th observation is uncensored)},} \\
0 & \text{if } Z_i = T_i \text{ (i\text{th observation is censored)},}
\end{cases}
\]

where \( T_i \) is the time to censorship of the \( i\text{th observation} \). Here we assume that \( X_1, \ldots, X_n \) are independent and identically distributed according to a continuous distribution \( F \), \( T_1, \ldots, T_n \) are independent and identically distributed according to a continuous censoring distribution \( H(x) = P(T \leq x) \), and furthermore the \( T \) and the \( X \)'s are assumed mutually independent. The censoring distribution \( H \) is typically, though not necessarily, unknown and is treated as a nuisance parameter.
2. THE ONE-SAMPLE LIMIT OF EFRON'S TWO-SAMPLE TEST

Efron (1967) has considered the two-sample problem with right-censored data. In addition to the $Z$'s and $\delta$'s defined by (1.1) and (1-2), one observes the analogous quantities

$$Z_i^* = \min(Y_i, S_i),$$

(2.1)

$$\epsilon_i = \begin{cases} 1 & \text{if } Z_i^* = Y_i, \\ 0 & \text{if } Z_i^* = S_i, \end{cases}$$

(2.2)

where $i = 1, \ldots, m$, and $m$ denotes the size of sample 2. Here $Y_1, \ldots, Y_m$ are independent and identically distributed according to an unknown distribution $G$, $S_1, \ldots, S_m$ are independent and identically distributed according to $I(x) = P(S \leq x)$, and the $X$'s, $T$'s, $Y$'s, and $S$'s are assumed mutually independent.

In the two-sample problem, the null hypothesis is

$$H_0^* : F(x) = G(x), \text{ all } x,$$

(2.3)

where the hypothesized common distribution is unspecified. Efron's test of $H_0^*$ is based on the statistic

$$\hat{W} = -\int \hat{G}(x) d\hat{F}(x),$$

(2.4)

where $\hat{F}, \hat{G}$ are the Kaplan-Meier (1958) estimators of $F, G$ respectively. Letting $Z(1) < Z(2) \ldots < Z(n)$ denote the ordered $Z$'s,

$$\hat{F}(x) = \prod_{j=1}^{k-1} \frac{\delta_j}{(n-j)/(n-j+1)}, \quad x \in (Z_{k-1}, Z_k],$$

and $\hat{F}(x) = 0$ for $x > Z(n)$. Of course $\hat{G}$ is defined analogously.
approaches 0. If \( \sigma_0^2 \) is finite and if \( \hat{\sigma}^2 \) is a consistent estimator of \( \sigma_0^2 \), then under \( H_0 \),

\[
C^* \overset{\text{def.}}{=} n^2 (C - \frac{1}{2}) / \hat{\sigma} \rightarrow N(0, 1).
\]

Assuming \( \sigma_0^2 \) is finite, one consistent estimator of \( \sigma_0^2 \) is

\[
\hat{\sigma}^2 = 4^{-1} \int \frac{z^3 (\overline{K}_n (G^{-1}(z))^*)^{-1} dz}{\overline{G}(Z(n))}, \tag{2.7}
\]

where \( \overline{K}_n \), the empirical survival function of the \( Z \)'s, is

\[
\overline{K}_n(x) = \begin{cases} 
(n-i+1)/n, & Z(i-1) \leq x < Z(i), \\
0, & Z(n) \leq x,
\end{cases}
\]

where \( Z(0) = -\infty \). Expression (2.7) can be simplified to

\[
\hat{\sigma}^2 = 16^{-1} \sum_{i=1}^{n} \left\{ n/(n-i+1) \right\} \left\{ \overline{G}(Z(i-1)) \right\}^4 - \left\{ \overline{G}(Z(i)) \right\}^4 \}
\]

Let \( X \) be distributed according to \( G \) and let \( X^* \) be independent of \( X \) and have distribution \( G^* \). To test \( H_0 \) versus one-sided alternatives \( F = G^* \) where \( P(X \geq X^*) < \frac{1}{2} \), we reject \( H_0 \) if \( C^* < -z_\alpha \) and accept \( H_0 \) otherwise. To test \( H_0 \) versus one-sided alternatives \( F = G^* \) where \( P(X \geq X^*) > \frac{1}{2} \), we reject \( H_0 \) if \( C^* > z_\alpha \) and accept \( H_0 \) otherwise. Here \( z_\alpha \) is the upper \( \alpha \) percentile point of a standard normal distribution.

When there is no censoring, our goodness-of-fit test based on \( C \) reduces to Moses' (1964) goodness-of-fit test based on the one-sample limit of Wilcoxon's two-sample statistic. That is, with no censoring, \( C = \sum_{i=1}^{n} \overline{G}(X_i) / n \) which under \( H_0 \) is distribution-free with distribution that of the average of \( n \) independent uniform random variables. The test then refers \( (12n)^{\frac{1}{2}} (C - \frac{1}{2}) \) to the standard normal distribution.
3. Comparisons of the C Test, the Koziol-Green Test, and Hyde's Test

Competing tests of $H_0$ have recently been proposed by Koziol and Green (1976) and Hyde (1977).

Koziol-Green (1976) Test: Apply the probability integral transformation to the Z's to form new pairs $(V_i, \delta_i)$, where $V_i = \min(U_i, L_i)$, $U_i = G(X_i)$, $L_i = G(T_i)$, and $\delta_i$ is as before. This reduces the problem to testing whether the distribution of the U's is uniform on $(0, 1)$. The Koziol-Green statistic, a generalization of the Cramér-von Mises statistic to the right-censored situation, is

$$\psi^2 = n \int_0^1 (\hat{F}_U(t) - t)^2 dt,$$

where $\hat{F}_U$ is the Kaplan-Meier estimator of the distribution of $U$.

Koziol and Green derive the asymptotic distribution of $\psi^2$ under the restriction that the censoring distribution $H$ be related to the survival distribution $F$ via

$$H = F^\beta,$$

for some $\beta$, $0 < \beta < 2$. For this model,

$$P(\delta_i = 0) = \int (1 - F^\beta) dF = \beta/(\beta+1),$$

so that Koziol and Green interpret $\beta$ as the censoring parameter. Koziol and Green's asymptotic theory for $\psi^2$ restricts $\beta$ to be less than 2, and they give asymptotic critical points of $\psi^2$ for the models $\beta = 0$ (no censoring) and $\beta = .5, 1, and 1.5$. Thus to implement the $\psi^2$ test, the user must know $\beta$ or
Hyde shows that the statistic

\[
D = A \left\{ \frac{n}{\sum_{i=1}^{n} a_i} \right\}^{\frac{1}{2}} \sim N(0, 1),
\]

where \( \left\{ \frac{n}{\sum_{i=1}^{n} a_i} \right\}^{\frac{1}{2}} \) is, under \( H_0 \), a consistent estimator of the standard deviation of \( A \). Hyde makes the additional assumption that \( E(a_i) \) be finite, but that this condition is automatically satisfied follows from Hyde's result that \( E(\delta_i - a_i) = 0 \). Hyde's statistic, when specialized to our model by setting \( v_i = 0 \) for all \( i \), becomes

\[
D = \frac{n}{\sum_{i=1}^{n} \left\{ \delta_i + \log \bar{G}(z_i) \right\}^{\frac{1}{2}} - \sum_{i=1}^{n} \log \bar{G}(z_i)} .
\]

(3.4)

If the failure rate \( r(x) = g(x)/\bar{G}(x) \) exists, where \( g(x) = (d/dx)G(x) \), then \( D \) can be written as

\[
D = \frac{n}{\sum_{i=1}^{n} \left\{ \delta_i - \int r(u) du \right\}^{\frac{1}{2}} - \sum_{i=1}^{n} \int r(u) du} .
\]

When \( D \) is significantly large (small), \( H_0 \) is to be rejected in favour of the alternative that the true average failure rate is larger (smaller) than the average failure rate of the hypothesized distribution \( G \).
To be specific, for

\[ H_1(x) = \begin{cases} 
1.98x, & 0 \leq x \leq \frac{1}{2}, \\
1 - \exp(-bx), & \frac{1}{2} \leq x,
\end{cases} \]

with \( b = -2\log_{10} \), we find \( \Delta(F_1, H_1) = -16 \). Thus if \( H_1 \) is the true censoring distribution, and the sample size \( n \) is sufficiently large, Hydts will lead to the decision that the failure rate \( 2x \) of \( F_1 \) is "larger" than the constant failure rate 1 of the hypothesized \( G \). However, for

\[ H_2(x) = \begin{cases} 
.0025x, & 0 \leq x \leq 2, \\
.005 + .99(x-2), & 2 \leq x \leq 3, \\
1 - \exp(-cx), & 3 \leq x,
\end{cases} \]

with \( c = -(1/3)\log_{10} .005 \), we find \( \Delta(F_1, H_2) = .11 \). Thus when \( H_2 \) is the true censoring distribution, and \( n \) is sufficiently large, Hydts test should lead to the decision that the failure rate of \( F_1 \) is "smaller" than that of \( G \).

An advantage of the C test proposed in Section 2 is that estimates \( \int G(x) d\delta(x) \), independently of the censoring distribution \( H \), The analogous procedure for \( \hat{W} \) in the two sample situation was a motivating factor in Efron's development of the test based on \( \hat{W} \).

In a limited study we have obtained Monte Carlo power comparisons of the test based on C, D, and \( \psi^2 \). Since the tests based on \( \psi^2 \) are only asymptotically exact, our study also provides information about the closeness of the true levels to their nominal asymptotically correct values. Though the tests based on C and D can be one-sided or two-sided, the \( \psi^2 \) test is inherently two-sided and the only two-sided counterparts based on C and D were used. Furthermore, although assumption (3.2) is restrictive and not required by the C or D tests, in fairness to the \( \psi^2 \) test we have sampled from situations where \( \beta \) is one of the values for which asymptotic percentage points of \( \psi^2 \) are tabulated by Koziol and Green.
Table 1. Estimated powers of C, D, and $\psi^2$ tests.

(a) Hypothesized normal, normal location alternatives.

\[ G(x) = \phi(x), \ F(x) = \phi(x-\theta), \ \overline{H} = (\overline{F})^\beta \]

<table>
<thead>
<tr>
<th>Test</th>
<th>C</th>
<th>D</th>
<th>$\psi^2$</th>
<th>C</th>
<th>D</th>
<th>$\psi^2$</th>
<th>C</th>
<th>D</th>
<th>$\psi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha/\theta$</td>
<td>0</td>
<td>.25</td>
<td>.5</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.01</td>
<td>.013</td>
<td>.018</td>
<td>.011</td>
<td>.078</td>
<td>.017</td>
<td>.040</td>
<td>.381</td>
<td>.154</td>
<td>.230</td>
</tr>
<tr>
<td>.05</td>
<td>.038</td>
<td>.061</td>
<td>.051</td>
<td>.199</td>
<td>.081</td>
<td>.139</td>
<td>.590</td>
<td>.384</td>
<td>.466</td>
</tr>
<tr>
<td>.10</td>
<td>.096</td>
<td>.114</td>
<td>.098</td>
<td>.279</td>
<td>.167</td>
<td>.226</td>
<td>.703</td>
<td>.525</td>
<td>.600</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n = 50, $\beta = \frac{1}{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha/\theta$</td>
</tr>
<tr>
<td>.01</td>
</tr>
<tr>
<td>.05</td>
</tr>
<tr>
<td>.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n = 20, $\beta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha/\theta$</td>
</tr>
<tr>
<td>.01</td>
</tr>
<tr>
<td>.05</td>
</tr>
<tr>
<td>.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n = 50, $\beta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha/\theta$</td>
</tr>
<tr>
<td>.01</td>
</tr>
<tr>
<td>.05</td>
</tr>
</tbody>
</table>
16 still alive. Those observations corresponding to deaths due to other causes those corresponding to the 16 survivors are treated as censored observations (drawals). As reported by Koziol and Green (1976), there is a basis for suspect that had the patients not been treated with estrogen, their survival distribution for deaths from cancer of the prostate would be exponential with mean 100 months. We thus applied the C, D, and $\psi^2$ statistics to test that the survival distribution is $\bar{G}(x) = \exp(-x/100)$.

For the data of Table 2, $C^* = .69$ with a corresponding two-sided $P$ value of .49. Hyde's statistic is $D = -.17$ with a corresponding two-sided $P$ value of .86. The value of the Koziol-Green statistic for the data of Table 2 is $\psi^2 = 1.02$. The proportion $p_c$ of censored observations is $121/211 = .573$, we find from (a) that $\hat{\beta} = 1.34$. Entering Table 1 of Koziol and Green (1976) at $\beta = 3/2$ with $\psi^2 = 1.02$ gives $P \approx .14$.

Although all $P$ values are consistent with the hypothesized exponential with mean 100, the value of $\psi^2$ is more suggestive of a possible deviation from $H_0$. The values of $C^*$ and $D$. Some insight into this is obtained from Figure 1 which contains plots of the Kaplan-Meier estimator $\hat{F}$ and the hypothesized survival function $\bar{G}$.

The visual indication from Figure 1 is consistent with an underlying life distribution $F$ having the property that $\int FdG$ is close to $1/2$. Indeed, the value of $C$ for the data of Table 2 is .51. Recall that $C$ estimates $\int FdG$. Similarly, Figure suggests that the average failure rates of $F$ and $G$ are close and thus it is not surprising that Hyde's statistic assumes a value that is close to its null expected value of zero. However, in a Cramér-von Mises type statistic such as the Koziol-Green $\psi^2$, the $\hat{F}(x) - G(x)$ differences are squared. Thus the negative deviations, found here mostly for the middle month values, do not "cancel" the positive deviations, found in the early and late months portions of the axis. This is a possible explanation for the relatively lower $P$ value achieved by $\psi^2$. 
The value of $\psi^2 = 1.02$ was computed for an updated version of the
even allowing for updating, our
of 1.02 is not close to their reported value of
value $\psi^2 = 4.84$. Although we to obtain the earlier data, we believe the Koziol - Green value is incorrect
Koziol and Green may have arised their use, in Appendix 2 of their paper, of the same symbol $n$ to denote a
fixed sample size and the random number of unsorted observations. The possibility of an error of this nature has been confirmed by Drs. Koziol and Green in conversations with the authors of this paper.

Table 2: Survival times and withdrawal times in months for 211 patients
(with number of ties given in parentheses)


This research was sponsored by the U.S. Air Force Office of Scientific Research.
The authors express their gratitude to John Kitchin for programming the Monte Carlo calculations of Section 3, the computations of $C$, $D$, and $\psi^2$ for the data of Table 2, and a computer plot of Figure 1. We also thank James Koziol and Sylvan Green for furnishing the data of Table 2, and John Hyde, George Marsaglia, John McCool, and Jayaram Sethuraman for helpful discussions.
REFERENCES


For right-censored data, we develop a goodness-of-fit procedure for testing whether the underlying distribution is a specified function $G$. Our test statistic $C$ is the one-sample limit of Efron's (1967) two-sample statistic $\hat{W}$. The test based on $C$ is compared with recently proposed competitors due to Koziol and Green (1976) and Hyde (1977). The comparisons are on the basis of (i) applicability, (ii) the extent to which the censoring distribution can affect the inference, and (iii) power. It is shown that in certain situations the $C$ test compares favourably with the tests of Koziol-Green and Hyde.