NAG 8-555

AIAA 38.

GODDARD GRANT

IN-89-CR

146706

SURVIVAL ANALYSIS, OR WHAT TO DO WITH UPPER LIMITS IN ASTRONOMICAL SURVEYS

TAKASHI ISOBE and ERIC D. FEIGELSON

Department of Astronomy
Pennsylvania State University
University Park, PA 16802

Abstract: A field of applied statistics called 'survival analysis' has been developed over several decades to deal with 'censored data', which occur in astronomical surveys when objects are too faint to be detected. We review here briefly, and elsewhere in more detail, how these methods can assist in the statistical interpretation of astronomical data.

(NASA-CR-182936) SURVIVAL ANALYSIS, OR WHAT TO DO WITH UPPER LIMITS IN ASTRONOMICAL SURVEYS (Pennsylvania State Univ.) 3 p CSCL 03A

N88-24551

Unclas G3/89 0146706

Astronomers often need to analyze samples of objects for which incomplete information is available. For example, a study of the X-ray properties of optically selected quasars might find more than half of the X-ray observations are upper limits. In statistics, these are called left censored data points. Some researchers have omitted these censored data from consideration. Others divided them into detected and undetected subsamples and compared them by two sample tests. These methods can introduce bias to the results and do not make efficient use of the data. The presence of upper limits make it difficult to measure even simple statistical quantities, such as the mean X-ray luminosity of the sample.

Similar cases occur in many fields of research (e.g. biomedical research, actuarial science, industrial reliability testing, econometrics) and many statisticians have studied these problems. The methods they have developed comprise a field generally known as 'Survival Analysis'. Although it may seem that observations in astronomy are very different from data in these fields, we find that the mathematical problems are the same. The field of survival analysis is now quite extensive, and is described in several recent books (e.g. Miller 1981, Lawless 1982). We have summarized some of the methods that might be particularly useful to astronomers in two recent papers (Feigelson and Nelson 1985, Isobe et al. 1986; Papers I and II).

The method used for several centuries by actuaries to determine the distribution function of a censored data set is in fact identical to the "fractional luminosity function" frequently used by radio astronomers (e.g. Auriemma et al. 1977). Its limitations are that accuracy is lost due to binning, error analysis is not simple ($\sqrt{N}$ errors are not correct, the actuarial "Greenwood's formula" must be used), and few useful statistical tests are available. The situation changed dramatically in the 1950-70's. With the advent of Kaplan and Meier's unbinned maximum-likelihood estimator for the distribution function, the discovery of several non-parametric 2-sample tests, and the development of several tests for correlation and linear regression. Some of this progress has been paralleled in astronomy with the independent discovery of the Kaplan-Meier estimator (Avni et al. 1980; Hummel 1981; Pfleiderer and Krommidas 1982) and of the EM-algorithm linear regression (Avni and Tananbaum 1986). Avni's procedures are now widely used in X-ray astronomy. We believe that the power and reliability of astronomers' interpretation of censored data will be greatly improved if we have the full range of survival analysis techniques available to us.

We can briefly review the principal elements of survival analysis discussed in Papers I and II. The Kaplan-Meier estimator finds the maximum-likelihood distribution of a censored data set. For example, we can find the X-ray luminosity function of optically selected quasars, its mean value and its error. Assuming there are no ties among the data z_i, the Kaplan-Meier estimator is given by

$$S(z_i) = \prod_{z_j < z_i} (1-1/n_j)^{\delta_j} \qquad \text{where } z_i > z_1,$$

$$S(z_i) = 1 \qquad \text{where } z_i < z_1,$$

where $\delta_j=1$ if z_j is detected, $\delta_j=0$ if z_j is undetected, and n_i = number of data points in $R(z_i)$. $R(z_i)$ is called the risk set, which consists of all data points which have not been detected before z_i. The Kaplan-Meier function has simple analytic formulae for error analysis, and gives the mean and standard deviation of the distribution.

If we need to test the hypothesis that two or more populations have the same distributions (e.g., the X-ray luminosity functions of optically selected and radio selected quasars), several procedures are available (Paper I). Gehan's extension of the Wilcoxon test and the logrank test are most commonly used in biomedical applications. The former is more effective than the latter when the underlying distribution function is normal. The logrank test is more effective at finding differences in the two samples at the censored end of the distribution, while the Gehan test is more sensitive at the uncensored end.

Now if we want to see whether a correlation exists between two variables (e.g., the optical and X-ray luminosities), Cox regression and a generalization of Kendall's τ correlation coefficient are very useful (Paper II). The former one is very popular in many fields of study if only the dependent variable contains censored data. The latter method is not yet widely used, but it permits any type of censoring. If a correlation is present, linear regression can be performed. The EM (estimate and maximize) algorithm with a normal distribution (Wolynetz 1979) or with a Kaplan-Meier distribution (Buckley and James 1980) can treat this problem. Schmitt (1985) has developed a linear regression method for data censored in both variables.

The main advantages of survival analysis over methods traditionally used by astronomers are its wide variety of statistical tools and their mathematical robustness well-established by professional statisticians. The principal disadvantages are its complexity and inconvenience compared to previous methods. We have used a combination of programs in commercial statistical software packages (BMDP for the Kaplan-Meier estimator, 2-sample tests and Cox regression), published FORTRAN codes (the Kaplan-Meier and 2-sample tests in Lee 1976, and the EM algorithm linear regression in Wolynetz 1979), and codes written by ourselves (generalized Kendall's τ and Schmitt's linear regression for dual censored data). We are glad to share our experience and codes with others and, if sufficient interest is present, may produce a more coherent package for astronomical use. Please feel free to contact us.

While astronomers can substantially benefit from the survival analysis methods already developed, some additional statistical work needs to be done. An obvious deficiency is that existing methods assume the censored value is known exactly, while most astronomical upper limits are estimated from the absence of a signal in (assumed) Gaussian noise. A weighting scheme based on the noise value, rather than an artificial limit such as "3-σ", should be developed. Study is also needed on the statistics of truncated rather than censored data. Truncated data sets in astronomy are those where objects fainter than a given flux level are missing entirely from the sample. Statisticians have only recently realize (Woodruffe 1985) that the "C-method" derived by Lynden-Bell (1971) to overcome selection effects in quasar surveys is the analog of the Kaplan-Meier estimator for truncated data. We are beginning to study statistics for truncated data in some detail.

This work was supported in part by NASA grant NAG 8-555.

References

Auriemma, C. et al., Astr. Ap. 57, 41 (1977).
Avni, Y., et al., Ap.J. 238, 800 (1980).
Avni, Y. and Tananbaum, H., Ap.J. 305, 83 (1986).
Buckley, J., and James, I., Biometrika 66, 429 (1979).
Feigelson, E.D. and Nelson, P.I., Ap.J. 293, 192 (1985). [Paper I]
Hummel, E., Astr. Ap. 93, 93 (1981).
Isobe, T., Feigelson, E.D. and Nelson, P.I., Ap.J. 306, 490 (1986). [Paper II]
Lawless, J.F. Statistical Models and Methods for Lifetime Data, New York:Wiley (1982).
Lee, E.T. Statistical Methods for Survival Data Analysis, Belmont CA:Lifetime Learning Pub. (1980).
Lynden-Bell, D., Mon. Not. R. Astr. Soc. 155, 95 (1971).
Miller, R.G., Jr., Survival Analysis, New York:Wiley (1981).
Pfleiderer, J. and Krommidas, P., Mon. Not. R. Astr. Soc. 198, 281 (1982).
Schmitt, J.H., Ap.J. 293, 178 (1985).
Wolynetz, M.S. Appl. Statistics 28, 185 (1979).
Woodroofe, M., Ann. of Statistics 13, 163 (1985).

Authors' address

Eric D. Feigelson and Takashi Isobe
Department of Astronomy
525 Davey Laboratory
The Pennsylvania State University
University Park, PA 16802
U.S.A.

Submitted to Bull. Inform. CDS.