

## SHORTER COMMUNICATIONS

EDITOR:  
LOUISE M. RYAN

### Efficiency Robust Tests for Survival or Ordered Categorical Data

Boris Freidlin,<sup>1,\*</sup> Marvin J. Podgor,<sup>2</sup> and Joseph L. Gastwirth<sup>3</sup>

<sup>1</sup>Biometric Research Branch, National Cancer Institute,  
Bethesda, Maryland 20892, U.S.A.

<sup>2</sup>National Eye Institute, Bethesda, Maryland 20892, U.S.A.

<sup>3</sup>Department of Statistics, George Washington University,  
Washington, D.C. 20052, U.S.A.

\**email*: freidlinb@ctep.nci.nih.gov

**SUMMARY.** The selection of a single method of analysis is problematic when the data could have been generated by one of several possible models. We examine the properties of two tests designed to have high power over a range of models. The first one, the maximum efficiency robust test (MERT), uses the linear combination of the optimal statistics for each model that maximizes the minimum efficiency. The second procedure, called the MX, uses the maximum of the optimal statistics. Both approaches yield efficiency robust procedures for survival analysis and ordinal categorical data. Guidelines for choosing between them are provided.

**KEY WORDS:** Asymptotic relative efficiency; Censored data; Contingency tables; Dose-response data; Efficiency robustness; Stratified data; Survival analysis; Weighted log-rank statistic.

#### 1. Introduction

In many applications, the precise form of the model underlying the data is not known; however, several scientifically plausible ones are available. Often optimal tests for each of them exist. Unfortunately, use of any one optimal test may lead to a loss of power under another model. Two approaches have been developed to obtain a single test with good power properties over the range of the models. In survival analysis, Tarone (1981) and Fleming and Harrington (1991) use the maximum of the standardized optimum statistics (MX). Alternatively, the maximum efficiency robust test (MERT) developed by Gastwirth (1966, 1985) and Birnbaum and Laska (1967) uses a linear combination of the optimal tests.

This paper shows how the efficiency robustness properties of the two methods depend on the null correlation matrix of the optimal tests and compares their power properties in survival and dose response settings. Our aim is to provide guidelines for choosing a robust test. The necessary background is provided in Section 2. Section 3 presents the comparison of the two methods in the survival setting. Section 4 is devoted

to the analysis of  $2 \times K$  tables. Section 5 presents recommendations for use of the methods.

#### 2. General Background

Often the precise distribution underlying the data is not known, although a family,  $\Psi : \{f_i; i = 1, \dots, I\}$ , of plausible alternatives can be specified. Consider a situation where the following three conditions hold: (1) asymptotically most powerful tests,  $\{S_i\}$ ,  $i = 1, \dots, I$ , for the respective members of the alternative family  $\Psi$  exist; (2) under the null hypothesis, standardized versions  $\{T_i\}$  of the  $\{S_i\}$  are asymptotically jointly multivariate normal with correlation matrix  $\{\rho_{ij}\}$ , where all  $\rho_{ij} > 0$ ; and (3) the Pitman asymptotic relative efficiency (ARE) of the test  $T_i$  relative to the test  $T_j$  when  $T_j$  is optimum is  $\rho_{ij}^2 = \langle T_i, T_j \rangle^2$ . These conditions are satisfied in a wide variety of applications (Van Eeden, 1964; Gross, 1981).

For any asymptotically normal test statistic  $T$ , denote its relative efficiency to the optimal test  $T_i$  for model  $f_i$  by  $e(T, i)$ . The lowest ARE  $T$  has when a model in  $\Psi$  is true is denoted

$e(T, \Psi) = \inf_{1 \leq i \leq I} \{e(T, i)\}$ . MERT satisfies

$$e(\text{MERT}, \Psi) = \sup_{T \in \Gamma} \left[ \inf_{1 \leq i \leq I} \{e(T, i)\} \right],$$

where  $\Gamma$  is the set of all consistent asymptotically normal tests for the problem. Gastwirth (1966) showed that, when the minimum correlation of the optimal tests  $T_i$ ,  $\rho^* = \min(\rho_{ij})$ , is  $> 0$ , the MERT exists, is unique, and is a linear combination of the  $\{T_i\}$ . Another robust test statistic is  $MX = \max_{1 \leq i \leq I} (T_i)$ . Asymptotically, under the null hypothesis, MX is distributed as  $\max[MN(0, \{\rho_{ij}\})]$ .

The correlation matrix,  $\{\rho_{ij}\}$ , of the optimal statistics summarizes the structure of the family of alternative models, as each correlation reflects how close the two models are (Hall and Joiner, 1982). When the models are very far apart, an adaptive procedure (Bickel, 1982) is needed.

### 3. Applications to Survival Analysis

We assume that  $n_i$  ( $i = 1, 2$ ) patients are allocated between two treatment groups and the random censorship model applies. Associated with patient  $j$  in group  $i$  is a survival time  $U_{ij}$  and a censoring time  $V_{ij}$ , which are independent random variables with survival functions  $S_i(t)$  and  $C_i(t)$ , respectively. We observe  $T_{ij} = \min(U_{ij}, V_{ij})$ . Let  $T_k$  denote the ordered failure times in the pooled sample. Weighted log-rank tests are used to test the equality of two survival distributions  $H_0: S_1(t) = S_2(t) = S(t)$ , i.e.,

$$LRW = \sum_k W(T_k)(O_k - E_k), \tag{1}$$

where  $O_k$  and  $E_k$  are the observed and conditionally expected numbers of failures in group 1 at time  $T_k$  under  $H_0$  and  $W(\cdot)$  is a weight function. Let  $\hat{S}(t-)$  denote the Kaplan-Meier estimator in the pooled sample. Fleming and Harrington (1991, p. 257) introduced the family  $G^{a,b}$ , with the weight function  $W(t) = \{\hat{S}(t-)\}^a \{1 - \hat{S}(t-)\}^b$  (for  $0 \leq a, 0 \leq b$ ), and showed that conditions 1, 2, and 3 of Section 2 hold.

We consider four members of this family:  $G^{0,0}$ ,  $G^{2,0}$ ,  $G^{0,2}$ , and  $G^{2,2}$ . These cover a wide range of possible differences in the survival distributions. Specifically, the four statistics are designed to detect constant difference, early difference, late difference, and middle difference, respectively. The correlation matrix (uncensored case) for the family is

	$G^{0,0}$	$G^{2,0}$	$G^{0,2}$	$G^{2,2}$
$G^{0,0}$	1	.745	.745	.837
$G^{2,0}$		1	.167	.535
$G^{0,2}$			1	.535
$G^{2,2}$				1

Tests  $G^{0,0}$  and  $G^{2,0}$  were shown (Fleming and Harrington, 1991) to be asymptotically most powerful for specific location alternatives. While the analytical forms of the survival distributions for which the tests  $G^{0,2}$  and  $G^{2,2}$  are optimal are not known, they can be approximated since the optimal weight function for a test of the form (1) is proportional to the log hazard ratio (Schoenfeld, 1981). The alternatives for which the four statistics  $G^{0,0}$ ,  $G^{2,0}$ ,  $G^{0,2}$ , and  $G^{2,2}$  are optimal are labeled models A, B, C, and D, respectively.

Monte Carlo simulations were performed for the four alternative models in order to compare the powers of the MERT and MX procedures. We present a brief summary of how the

simulation studies were conducted; a more detailed description appears in Freidlin, Podgor, and Gastwirth (1998). The survival times in group 1 were exponential ( $\lambda = 1$ ) in all cases. For alternative model A, group 2 survival times were exponential ( $\lambda = 2$ ). For alternative model B, group 2 survival times were generated according to formula (4.24) from Fleming and Harrington (1991, p. 275) with  $\Delta = 1.526$ . The alternative models C and D were approximated by taking the log hazard ratio equal to the weight function. We considered both censored and uncensored data for all models. Censoring was simulated using the uniform (0, 2) distribution, which resulted in approximately 43% patients censored under  $H_0$ . Each treatment group had 50 patients, and 5000 replications were done for each model in all cases. The following families of alternative models were considered: (1) family of models A, B, C, and D,  $\rho^* = .167$ ; (2) family of models A, B, and D,  $\rho^* = .535$ ; and (3) family of models A and D,  $\rho^* = .837$ .

For each family of alternatives, Table 1 gives the powers of the one-sided .05 level tests optimal for the alternative models, the MERT, and the MX test for the family. For the family that consists of all four models A, B, C, and D ( $\rho^* = .167$ ), the MX statistic is more powerful than the MERT under three of the four models. For the family of models A, B, and D ( $\rho^* = .535$ ), the power advantage of the MX is small, especially in the censored case. For the family consisting of models A and D ( $\rho^* = .837$ ), there is no loss in power from using the MERT statistic.

### 4. Efficiency Robust Procedures for Ordered Categorical Data

Consider an experiment where response rates in  $K$  ordered groups are compared. Let  $n_i$ ,  $X_i$ , and  $\pi_i$  denote the sample size, number of responses, and the probability of response in the  $i$ th group, respectively ( $i = 1, \dots, K$ ), i.e., the  $X_i$  are binomial random variables with parameters  $n_i$  and  $\pi_i$ . Suppose it is expected that response probabilities,  $\pi_i$ , are monotone and can be modeled by a function  $\pi_i = H(\alpha + \beta\nu_i)$ , where  $H$  is a twice differentiable monotone link function, e.g., logistic  $\pi_i = \exp(\alpha + \beta\nu_i) / [1 + \exp(\alpha + \beta\nu_i)]$  and  $\nu_i$  are the monotone column scores. We are interested in testing the null hypothesis  $H_0: \beta = 0$  corresponding to  $\pi_1 = \pi_2 = \dots = \pi_K = \pi$  against  $H_1: \beta \neq 0$  (two sided) or  $H_2: \beta > 0$  (one sided). Tarone and Gart (1980) showed that, for the data generated by the model with link function  $H$  and scores  $\nu_s = \{\nu_{s,1}, \dots, \nu_{s,K}\}$ , the asymptotically most powerful test does not depend on  $H$  and is given by

$$T_s = \sum_{i=1}^K \nu_{s,i}(X_i - n_i\hat{\pi}) / \sqrt{V_T},$$

where

$$\hat{\pi} = \sum_{i=1}^K x_i / \sum_{i=1}^K n_i$$

and

$$V_T = \hat{\pi}(1 - \hat{\pi}) \left[ \sum_{i=1}^K \nu_{s,i}^2 n_i - \left( \sum_{i=1}^K \nu_{s,i} n_i \right)^2 / \sum_{i=1}^K n_i \right].$$

Suppose only a family of alternative sets of scores can be spec-

**Table 1**  
Empirical power estimates, survival setting

Family: Models A, B, C, and D						
Uncensored MERT weights = (0, .5949, .5949, .1248); $\rho^* = .167$						
	$G^{0.0}$	$G^{2.0}$	$G^{0.2}$	$G^{2.2}$	MERT4	MX4
Uncensored						
Model A	.951	.805	.836	.892	.952	.927
Model B	.815	.950	.207	.592	.817	.899
Model C	.688	.150	.921	.549	.764	.848
Model D	.879	.449	.769	.916	.853	.895
Null	.048	.048	.061	.052	.051	.052
Censored						
Model A	.873	.757	.645	.783	.866	.839
Model B	.873	.944	.266	.583	.824	.900
Model C	.310	.121	.492	.336	.382	.402
Model D	.729	.367	.726	.805	.732	.778
Null	.050	.050	.056	.053	.051	.052
Family: Models A, B, and D						
Uncensored MERT weights = (0, .5708, .5708); $\rho^* = .535$						
	$G^{0.0}$	$G^{2.0}$		$G^{2.2}$	MERT3	MX3
Uncensored						
Model A	.951	.805		.892	.920	.925
Model B	.815	.950		.592	.897	.915
Model D	.879	.449		.916	.828	.886
Null	.048	.048		.052	.051	.049
Censored						
Model A	.868	.747		.787	.852	.839
Model B	.874	.941		.577	.886	.909
Model D	.714	.350		.794	.683	.734
Null	.052	.052		.054	.053	.054
Family: Models A and D						
Uncensored MERT weights = (.522, .522); $\rho^* = .837$						
	$G^{0.0}$			$G^{2.2}$	MERT2	MX2
Uncensored						
Model A	.951			.892	.940	.942
Model D	.873			.917	.918	.912
Null	.048			.052	.053	.051
Censored						
Model A	.882			.788	.864	.861
Model D	.735			.801	.788	.789
Null	.049			.050	.050	.050

ified (Graubard and Korn, 1987). Gross (1981) and Podgor, Gastwirth, and Mehta (1996) showed that conditions 1, 2, and 3 hold in this setting.

A summary of a simulation study that examined the relationship between the minimum correlation,  $\rho^*$ , of the family and powers of the MERT and MX procedures is given below. For a  $2 \times 5$  table with 10 subjects per group and using logit links, we constructed 8 families each consisting of two alternative models with optimal test correlations .5, .6, .7, .75, .8, .85, .9, and .95, respectively. The alternatives were chosen so that the power of the optimal test would be near .80 ( $\alpha = .05$ ). We tabulated empirical power estimates for the

one-sided tests optimal for each of the alternative models as well as the MERT and MX tests (based on 100,000 replications).

The simulation results are given in Table 2. Both efficiency robust procedures protect against a potentially substantial loss of power that would occur if an incorrect model was utilized. For example, when  $\rho^* = .6$ , the loss of power can reach .40, while the powers of the MX and MERT tests were within .08 and .11, respectively, of the power of the optimum procedure. When  $\rho^* \geq .75$ , the results indicate that MX and MERT have similar power properties. When  $\rho^* \leq .5$ , the MX has higher minimum power.

**Table 2**  
Empirical power estimates, ordered categorical data

$\rho^*$	Under model 1				Under model 2			
	Test optimal for		MERT	MX	Test optimal for		MERT	MX
	Model 1	Model 2			Model 1	Model 2		
.5	.794	.272	.688	.726	.290	.803	.643	.706
.6	.794	.379	.711	.734	.410	.802	.689	.727
.7	.794	.502	.733	.746	.541	.792	.729	.736
.75	.794	.558	.745	.749	.604	.791	.748	.746
.8	.794	.613	.755	.764	.656	.788	.758	.755
.85	.794	.663	.764	.768	.707	.794	.776	.769
.9	.794	.700	.774	.775	.758	.794	.799	.790
.95	.794	.748	.782	.781	.772	.794	.792	.786

## 5. Conclusion

Because the asymptotic power properties of the MERT and MX procedures depend on the null correlation matrix of the optimal tests for the alternative models, the relationships we found apply across many areas of applications. The correlation matrix of the optimal statistics can guide one's choice of which robust procedure to use. With the increasing availability of powerful PCs, MX should be used, especially when  $\rho^* \leq .5$ . In situations where investigators are accustomed to the normal form of a statistic, MERT should be useful.

A SAS IML program for estimating MERT and MX  $p$ -values is available from the authors (BF).

## ACKNOWLEDGEMENTS

We thank the associate editor and two referees for their valuable comments. This research was partially supported by a grant from the National Science Foundation. The first and third authors would like to dedicate this paper to the memory of their coauthor, Dr. Marvin J. Podgor, who passed away in October 1998.

## RÉSUMÉ

La sélection d'une méthode unique d'analyse pose des problèmes quand les données auraient pu être générées par différents modèles. Nous examinons les propriétés de deux tests construits pour avoir une puissance élevée sur un ensemble de modèles. Le premier, le MERT, utilise la combinaison linéaire des statistiques optimales correspondant à chaque modèle qui maximise l'efficacité minimum. La seconde méthode, le MX, utilise le maximum des statistiques optimales. Les deux approches conduisent à des procédures robustes en terme d'efficacité pour l'analyse de données de survie et de données qualitatives ordonnées. Le choix entre les deux tests sera fait en fonction des recommandations proposées.

## REFERENCES

- Bickel, P. J. (1982). On adaptive estimation. *Annals of Statistics* **10**, 647-671.
- Birnbaum, A. and Laska, E. (1967). Efficiency robust two-sample rank tests. *Journal of the American Statistical Association* **62**, 1241-1257.
- Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
- Freidlin, B., Podgor, M. J., and Gastwirth, J. L. (1998). *Efficiency robust tests for survival or ordered categorical data*. Technical Report, George Washington University, Washington, D.C.
- Gastwirth, J. L. (1966). On robust procedures. *Journal of the American Statistical Association* **61**, 929-948.
- Gastwirth, J. L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of the American Statistical Association* **80**, 380-384.
- Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered  $2 \times K$  contingency tables. *Biometrics* **43**, 471-476.
- Gross, S. T. (1981). On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. *Journal of the American Statistical Association* **76**, 935-941.
- Hall, D. L. and Joiner, B. L. (1982). Representations of the space of distributions useful in robust estimation of location. *Biometrika* **69**, 55-59.
- Podgor, M. J., Gastwirth, J. L., and Mehta, C. R. (1996). Efficiency robust tests of independence in contingency tables with ordered classifications. *Statistics in Medicine* **15**, 2095-2105.
- Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68**, 316-319.
- Tarone, R. E. (1981). On the distribution of the maximum of the log-rank statistics and the modified Wilcoxon statistics. *Biometrics* **37**, 79-85.
- Tarone, R. E. and Gart, J. J. (1980). On the robustness of combined tests for trends in proportions. *Journal of the American Statistical Society* **75**, 110-116.
- Van Eeden, C. (1964). The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. *Annals of Mathematical Statistics* **34**, 1442-1451.

Received June 1998. Revised December 1998.

Accepted December 1998.