An Efficient Algorithm for Exact Distributions of Discrete Scan Statistics

Tangan Gao
California State University, Long Beach
tgao@csulb.edu

Waiting time random variables and related scan statistics have been studied extensively during the last four decades. There are four recent books in the area of runs, scan statistics, and their applications: Scan Statistics and Applications by Glaz and Balakrishnan (1999); Scan Statistics by Glaz, Naus and Wallenstein (2001); Runs and Scans with Applications by Balakrishnan and Koutras (2002); and Distribution Theory of Runs and Patterns and its Applications by Fu and Lou (2003). In these books chapters are devoted to discrete scan statistics and many of their interesting and useful applications. Due to complexity of computing exact distributions of discrete scan statistics, most attention has been given to providing bounds and approximations for the distributions. There are several methods for investigating exact distributions of discrete scan statistics. Most notably, the combinatorial method and the finite Markov chain imbedding method. To the best of our knowledge, the leading algorithm up to date for calculating exact distributions of scan statistics is the finite Markov chain imbedding method given in Fu and Lou (2003).

In this paper, exact distributions of discrete scan statistics $S_n(w)$ for the cases of homogeneous two-state Markov dependent trials as well as i.i.d. Bernoulli trials are discussed by utilizing conditional probability generating functions of waiting time random variables $WT(w, s)$. We first provide an efficient method to generate all conditional probability generating functions related to the random variable $WT(w, s)$. Then we show that the exact distribution of $WT(w, s)$ can be obtained by simple multiplications of sparse matrices and vectors, which leads to an efficient algorithm for calculating the exact distribution of the scan statistic $S_n(w)$. Other scan statistics closely related to $S_n(w)$ and $WT(w, s)$ are also briefly discussed. A computer algorithm in C++ has been developed to calculate the distributions based on our method. Numerical results show that our algorithm leads existing codes for exact distributions of discrete scan statistics in speed by a substantial margin. (Joint work with Morteza Ebneshahrashoob and Mengnien Wu).
