Empirical/Asymptotic P-Values for Monte Carlo-Based Hypothesis Testing: an Application to Cluster Detection Using the Scan Statistic

Allyson M. Abrams, Martin Kulldorff, Ken Kleinman

Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care

Objective

Our goal was to increase the precision of the p-value produced from SaTScan [1] while reducing the amount of CPU time needed by decreasing the number of Monte Carlo replicates.

Background

SaTScan is a freely available software that uses the scan statistic to detect clusters in space, time or space-time [2]. SaTScan uses Monte Carlo hypothesis testing in order to produce a p-value for the null hypothesis that no clusters are present. Monte Carlo hypothesis testing can be a powerful tool when asymptotic theoretical distributions are inconvenient or impossible to discover; the main drawback to this approach is that precision for small p-values can only be obtained through greatly increasing the number of Monte Carlo replications, which is both computer-intensive and time consuming. Depending on the type of analysis being done, the number of geographical areas included, the amount of historical data, and the number of Monte Carlo replications, SaTScan can take anywhere from seconds to hours to run. In doing daily surveillance of many syndromes, we need to limit the amount of time it takes to generate each p-value while still retaining enough precision in the p-value to determine how unusual a cluster is. Since the type of analysis done and the geographic regions being used cannot be changed in most cases, we focus here on trying to reduce the number of Monte Carlo replicates needed.

Methods

We ran SaTScan on a sample map using 100,000,000 Monte Carlo replicates in order to generate the 'true' log-likelihood ratio needed to obtain certain p-values. We also ran SaTScan 1000 times on the same map, each time generating 999 Monte Carlo replicates. In each of these 999 replicates the maximum log-likelihood ratio, among all distinct circles, is the statistic reported. The ordinary Monte Carlo p-value is the rank of the observed maximum log-likelihood ratio among the 999 maximum log-likelihood ratios from the Monte Carlo replicates, divided by 1000.

We found the maximum likelihood estimates of the parameters of various distributions, assuming the 999 replicates came from that distribution, for each of the 1000 SaTScan runs. The empirical/asymptotic p-value under a given distribution is the area to the right of the observed log-likelihood assuming the estimated parameters for that distribution. For each distribution, we generated: (1) empirical/asymptotic p-values based on the 'true' log-likelihood value and (2) the log-likelihoods that would have been required to generate a specified set of p-values.

We also repeated this entire process generating 99 and 9999 Monte Carlo replicates in each of 1000 SaTScan runs. In these cases, the Monte Carlo p-value is the rank of the observed maximum log-likelihood ratio among the 99 (or 9999) Monte Carlo log-likelihood ratios, divided by 100 (or 10,000).

Results

Intuitively, an extreme value distribution should be the best fit since the Monte Carlo replicates generate maximum log-likelihood ratios, and in fact the empirical/asymptotic p-values from the Gumbel distribution [3] appear unbiased. In contrast, other tested distributions, including the Gamma, Normal, and Lognormal, all resulted in biased p-values. Interestingly, the ordinary Monte Carlo p-values reported from SaTScan had greater variance than the Gumbel-based p-values. These results were similar regardless of the number of Monte Carlo replicates used, but as the number of replicates increased, the precision in the empirical/asymptotic p-values increased and the variance decreased.

Conclusions

Empirical/asymptotic p-values based on the Gumbel distribution can be preferable to true Monte Carlo p-values even when both are generated from an equal number of Monte Carlo replicates. Empirical/asymptotic p-values can also accurately generate p-values smaller than is possible with Monte Carlo p-values with a given number of replicates. We suggest empirical/asymptotic p-values as a hybrid method to accurately obtain small p-values with a relatively small number of Monte Carlo replicates.

References

[1] Kulldorff, M, *SaTScan: Software for the spatial and space-time scan statistics*. 2004, www.satscan.org.

[2] Kulldorff, M, *A spatial scan statistic.* Communications in Statistics: Theory and Methods, 1997. **26**: p. 1481-1496.

[3] Gumbel, EJ, *Statistics of Extremes*. Dover. 2004.