A Scan Statistic based on Anscombe's Variance Stabilization Transformation

Kunihiko Takahashi, Toshiro Tango

Department of Technology Assessment and Biostatistics
National Institute of Public Health, Japan

OBJECTIVE

This paper proposes a new scan statistic which detects disease clusters more accurately than that based on the likelihood ratio.

BACKGROUND

The circular spatial scan statistic proposed by Kulldorff and Nagarwalla [1] has been widely used along with SaTScan software for cluster detection. To detect arbitrarily shaped clusters which cannot be detected by the circular scan statistic, Duczmal and Assunção [2] and Tango and Takahashi [3] have proposed different scan statistics. All of these tests are based on maximizing the likelihood ratio statistic $\lambda(Z)$ for each window Z. However, Tango and Takahashi [3] have shown examples in which Duczmal and Assunção's procedure detected quite large and peculiar shaped clusters that had the largest likelihood ratio λ among the three scan statistics applied. It cast a doubt on the validity of the model selection based on maximizing $\lambda(Z)$.

METHODS

One of reasons for detecting undesirable clusters is that $\lambda(Z)$ is derived only from the observed number of cases $n(Z)$ and the expected number $\mu(Z)$ under the null hypothesis H_0 of no clustering. $\lambda(Z)$ ignores the variability of the relative risks of regions included in Z. Then we propose an alternative scan statistic that can take such variability into account.

Assume that, under H_0, the observed number of cases X_i is a Poisson random variable with expected value μ_i in each region $i = 1, 2, \ldots, m$. Then, let us apply Anscombe[4]'s variance stabilization transformation:

$$Y_i = 2\sqrt{X_i + (3/8)} - 2\sqrt{\mu_i + (1/8)},$$

where $E(Y_i) = 0$ and $\mathrm{Var}(Y_i) = 1$ under H_0. For any Z, let $\bar{y}(Z)$ and $\bar{y}(Z^c)$ be the means of y_i within Z and outside, respectively. Then, we propose a new scan statistic T as

$$T = \max_{Z \in \mathcal{Z}} \{\bar{y}(Z) - \bar{y}(Z^c)/\sqrt{(1/l(Z)) + (1/l(Z^c))}\},$$

where $l()$ denotes the number of regions included therein. The window Z^* which attains the maximum T is defined as the most likely cluster (MLC). In the same manner as Kulldorff's scan statistic, Monte Carlo testing is required for the distribution of T under H_0.

RESULTS

Several scenarios of simulation were used to illustrate the proposed test statistic T with scanning methods of the circular [1] and the flexible [3]. The bivariate power distribution proposed by [3] shows that T has shorter tails and, consequently, better ability of pinpointing the assumed hot-spot cluster compared with the likelihood ratio.

CONCLUSIONS

The proposed scan statistic can detect disease clusters more accurately than that based on the likelihood ratio.

REFERENCES

[1] Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995; **14**:799–810.

[2] Duczmal L, Assunção R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* 2004; **45**:269–286.

[3] Tango T, Takahashi T. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 2005; **4**:11.

[4] Anscombe F. J. The transformation of Poisson, binomial and negative-binomial data. *Biometrika* 1948; **35**:246–254.

Table. Bivariate power distributions $P(l, s) \times 1000$ of the likelihood ratio statistic λ and the proposed statistic T using the flexible scanning method ($K = 15$) for the hot-spot cluster **A** with $s^* = 3$ regions, where l is the length of significant MLC, s is the number of regions identified out of the assumed true cluster (see details [3]). Total number of cases is set to be 500 in the entire $m = 113$ regions (total number of population is 19,803,618), and relative risk in the hot-spot is set to be 3.0. Nominal α-lebel is set as 0.05 and 1000 trials are carried out. Lines of $s = 0, 1, 2$ whose all the cells have zero power are not shown. The mark "$*$" is the powers of accurate detection.

likelihood ratio statistic λ (traditional power = 1000/1000)

Length l	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
Include $s = 3$			$*468$	179	123	86	78	40	19	5	2	0	0	0	0	1000

proposed statistic T (traditional power = 1000/1000)

Length l	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
Include $s = 3$			$*891$	73	18	9	5	4	0	0	0	0	0	0	0	1000