

時空間スキャン統計量の p 値計算のための逐次計算法

栗木 哲 数理・推論研究系 教授

本研究は、高橋邦彦客員准教授（名古屋大学准教授）、原尚幸客員准教授（新潟大学准教授）との共同研究です。

1 空間疫学のスキャン統計量とその p 値

n カ所の各地域 $i \in V = \{1, \dots, n\}$ について、着目するイベント（例：患者）の発生数 X_i が、その期待度数 λ_i （例：患者の期待度数）とともに得られているとする。次の確率モデルを想定する。

$$X_i \sim \text{Poisson}(\theta_i \lambda_i) \quad (\text{独立に}).$$

θ_i はSMR (Standardized Mortality Ratio) とよばれる。空間疫学では、ある地域クラスターではそれ以外の地域よりもSMRが大ききという状況を想定し、その地域クラスターをホットスポットとよぶ。

ホットスポットを検出するためには、事前にホットスポットの候補（スキャンウィンドウ） $Z = \{Z_1, \dots, Z_M\}$, $Z_i \subset V$ を設定し、帰無仮説 $H_0 : \theta_i \equiv \theta_0$ (一定), 対立仮説 $H_1 : \exists Z \in \mathcal{Z} \text{ s.t. } \theta_i = \theta_Z (i \in Z), = \theta_Z (i \notin Z), \theta_Z > \theta_0$ とする多重検定を行う。尤度比検定統計量は $\max_{Z \in \mathcal{Z}} \varphi_N(X_Z)$, ただし

$$\varphi_N(X_Z) = \begin{cases} N \left\{ p \left(\frac{\hat{p}}{p} \log \frac{\hat{p}}{p} - \frac{\hat{p}}{p} + 1 \right) + (1-p) \left(\frac{1-\hat{p}}{1-p} \log \frac{1-\hat{p}}{1-p} - \frac{1-\hat{p}}{1-p} + 1 \right) \right\}, & \text{if } \frac{\hat{p}}{p} \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

$$X_Z = (X_i)_{i \in Z}, N = \sum_{i \in V} X_i, \hat{p} = \sum_{i \in Z} X_i / \sum_{i \in V} X_i, p = \sum_{i \in Z} \lambda_i / \sum_{i \in V} \lambda_i \text{ で与えられる.}$$

検出されたホットスポットの有意性を評価するための多重性調整 p 値は、最大値統計量の帰無仮説 H_0 の下での分布（ N 所与の条件付分布）から定義される。

$$P\left(\max_{Z \in \mathcal{Z}} \varphi_N(X_Z) \leq c\right) = E\left[\prod_{Z \in \mathcal{Z}} \chi_Z(X_Z) \mid N\right],$$

ただし $\chi_Z(X_Z) = 1\{\varphi_N(X_Z) \leq c\}$. 総和 N を与えたときの $(X_i)_{i \in V}$ の分布は多項分布 $\text{Mult}(N; (p_i)_{i \in V})$, $p_i = \lambda_i / \sum_{i \in V} \lambda_i$ である。

2 多項分布の分解と逐次計算

多項分布ベクトル

$$X_V = (X_1, \dots, X_l, X_{l+1}, \dots, X_m, X_{m+1}, \dots, X_n) \sim \text{Mult}(N; (p_1, \dots, p_n))$$

とその2つの周辺ベクトル

$$X_{B_1} = \overbrace{(X_1, \dots, X_l, X_{l+1}, \dots, X_m)}^{X_{R_1}} \\ \underbrace{(X_{l+1}, \dots, X_m, X_{m+1}, \dots, X_n)}_{X_{C_1} = X_{B_1 \cap B_2}} = X_{B_2}$$

を考える。 X_V の乱数は3ステップに分けて生成できる：

$$(M_2, N_1) \mid N \sim \text{Mult}\left(N; \left(\sum_{i \in B_2} p_i, \sum_{i \in R_1} p_i\right)\right), \\ X_{B_2} \mid M_2 \sim \text{Mult}\left(M_2; (p_i)_{i \in B_2} / \sum_{i \in B_2} p_i\right), \\ X_{R_1} \mid N_1 \sim \text{Mult}\left(N_1; (p_i)_{i \in R_1} / \sum_{i \in R_1} p_i\right).$$

この分解に対応して

$$E[\chi_1(X_{B_1}) \chi_2(X_{B_2})] = E^{(M_2, N_1) \mid N} [E^{X_{B_2} \mid M_2} [E^{X_{R_1} \mid N_1} [\chi_2(X_{B_2}) \chi_1(X_{B_1})]]] \\ = E^{(M_2, N_1) \mid N} [E^{X_{B_2} \mid M_2} [\chi_2(X_{B_2}) \xi(N_1, X_{C_1})]]$$

ただし

$$\xi(N_1, X_{C_1}) = E^{X_{R_1} \mid N_1} [\chi_1(\underbrace{X_{C_1}}_{X_{B_1}}, X_{R_1})].$$

この関係式より、次の逐次計算が提案できる。

- $X_{B_1} = (X_{C_1}, X_{R_1})$ の要素のうち、 X_{C_1} を固定し、 X_{R_1} について期待値をとる。 $\xi(N_1, X_{C_1})$ は数表の形で保持する。
- 残りの X_{B_2} について期待値をとる。

計算量（期待値計算における「和」の数）は、逐次計算による場合 $O(N^{\max(|B_1|, |B_2|)})$, 逐次計算を用いない場合 $O(N^{|B_1 \cup B_2| - 1})$ 。

3 マルコフ構造の抽出

$B_1, \dots, B_m \subset V$ は、各 $1 \leq i \leq m-1$ について、 $k(i) > i$ が存在し $B_i \cap (\bigcup_{j>i} B_j) = B_i \cap B_{k(i)}$ となるとき、RIP (Runnig Intersection Property) を持つという。 $m=2$ のときは自動的にRIPとなる。 B_1, \dots, B_m がRIPのとき、 $E[\prod_{i=1}^m \chi_i(X_{B_i}) \mid N]$ の逐次計算可能性を証明できる。

V のRIP部分集合列 B_1, \dots, B_m が存在し、各 j について Z_j が B_1, \dots, B_m のいずれかに含まれるならば（それを $B_{\tau(j)}$ とおく）、期待値

$$E\left[\prod_j \chi_j(X_{Z_j})\right] = E\left[\prod_{i=1}^m \chi_i(X_{B_i})\right], \quad \chi_i(X_{B_i}) := \prod_{j \in \tau^{-1}(i)} \chi_j(X_{Z_j})$$

は逐次計算が可能である。RIP部分集合列 B_1, \dots, B_m は次の「コーダル拡張」で得られる。

- 頂点集合を $V = \{1, \dots, n\}$, 辺集合を $E = \{(i, j) \in V \times V \mid i, j \in Z, \exists Z \in \mathcal{Z}\}$ とする無向グラフ $G = (V, E)$ を定義する。
- グラフに辺を加え（加える辺の集合を E_1 とおく）拡張グラフ $\tilde{G} = (V, E \cup E_1)$ をコーダルグラフとする（コーダル拡張）。
- 拡張グラフ \tilde{G} の「極大クリークの完全列」 B_1, \dots, B_m が求めるRIP列となる。

4 山形県胆嚢ガンデータの解析

1996–2000年、山形県市町村別の胆嚢ガン件数データ（丹後・高橋・横山, 2007; Tango, 2010）について2種類のスキャンウィンドウを試みた：

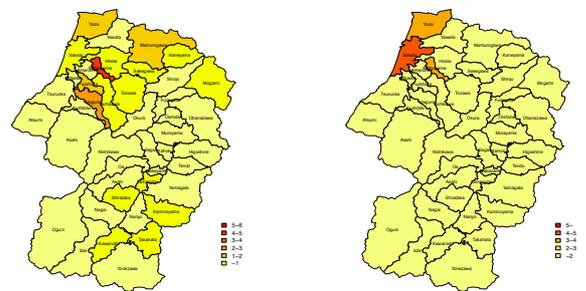


図. SMR (左) と Kulldorff 統計量 (右)

[solo] 各市町村を単独でスキャン、ウィンドウ数は市町村数44. [solo+pair] 各市町村単独と隣接する2市町村のペアをスキャン、ウィンドウ数は154 (ランダムに約半分76+78に分け、それぞれの p 値を計算し、それを足し合わせた)。

表. 検出された地域とその p 値

statistic	window	solo+pair	solo
7.651	{酒田, 遊佐}	0.00953	–
4.578	{酒田}	0.1847	0.0433
4.356	{酒田, 平田}	0.2247	–
4.247	{酒田, 三川}	*	–
3.924	{酒田, 余目}	*	–
3.570	{酒田, 八幡}	*	–
3.364	{松山}	*	0.1739
3.205	{藤島, 羽黒}	0.6458	–
3.071	{遊佐}	*	0.2065