疑似相関を用いた多重性調整法の開発

二宮 嘉行　　数理・推論研究系 教授

目的と設定

多次元データの期待値を比較する最も基本的な多重検定問題において，検定間に未知の相関があるとき，検出力を上げるため，仮説によっては一致性をもたない疑似相関を用い，多重性調整済 p 値を与える

コントロール群とケース群の独立な二次元データの期待値を比較する多重検定問題に対し，前者を $\{(y_{i1}^{(0)}, y_{i2}^{(0)}) \mid 1 \le i \le 10\}$，後者を $\{(y_{i1}^{(1)}, y_{i2}^{(1)}) \mid 1 \le i \le 10\}$，期待値と分散を次とする（$u \in \{0,1\}$，未知）

$$\mathrm{E}\begin{pmatrix} y_{i1}^{(u)} \\ y_{i2}^{(u)} \end{pmatrix} = \begin{pmatrix} \mu_1^{(u)} \\ \mu_2^{(u)} \end{pmatrix}, \qquad \mathrm{V}\begin{pmatrix} y_{i1}^{(u)} \\ y_{i2}^{(u)} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

検定問題

$$H_1^{(1)} : \mu_1^{(1)} = \mu_1^{(0)} \quad \text{vs} \quad K_1^{(1)} : \mu_1^{(1)} > \mu_1^{(0)}$$
$$H_2^{(1)} : \mu_2^{(1)} = \mu_2^{(0)} \quad \text{vs} \quad K_2^{(1)} : \mu_2^{(1)} > \mu_2^{(0)}$$

に対して通常の t 統計量 $T_1^{(1)}$ と $T_2^{(1)}$ を考え，共通の閾値 c により棄却域を設定する，つまり，有意水準が α のとき $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ が既知なら

$$\mathrm{FWER} \equiv \mathrm{P}_{H_1^{(1)} \cap H_2^{(1)}}\{\max(T_1^{(1)}, T_2^{(1)}) > c\} \le \alpha$$

なる c を棄却限界として求めればよい（P_H は仮説 H のもとでの確率）

本設定では σ_{jk} は未知であるため，それらには推定量を代入し，FWER を漸近的にコントロールすることを考える

MaxT 法

$(T_1^{(1)}, T_2^{(1)})^{\mathrm{T}}$ の漸近帰無分布である期待値 0，分散 1，相関 $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ の二次元正規分布において，σ_{jk} の代わりにその不偏推定量

$$\tilde{\sigma}_{jk} = \frac{1}{18}\sum_{u=0}^{1}\sum_{i=1}^{10}(y_{ij}^{(u)} - \bar{y}_j^{(u)})(y_{ik}^{(u)} - \bar{y}_k^{(u)})$$

を用いるのが最も自然であり（$j,k \in \{1,2\}$），$\tilde{\sigma}_{jk}$ は真の仮説に依らず一致性をもつため，この方法は FWER を漸近的にコントロールする

提案手法

検出力を上げるためには，$H_1^{(1)} \cap H_2^{(1)}$ を真としたモデルのもとでの σ_{jk} の次の不偏推定量を用いたい

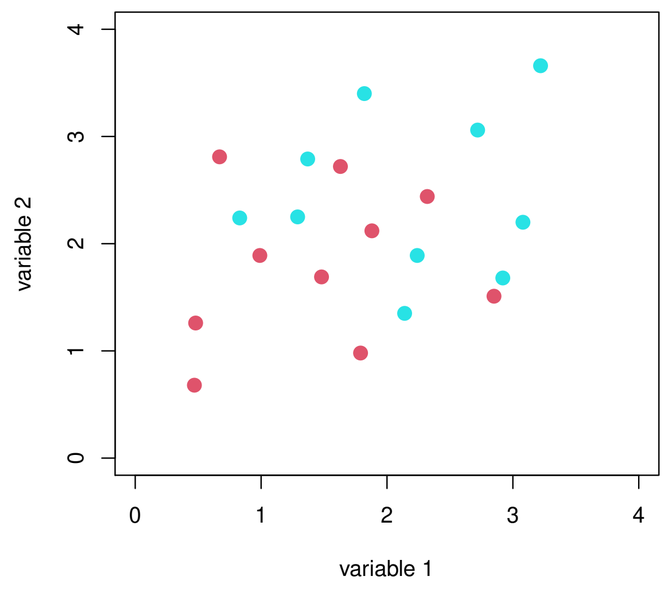
$$\hat{\sigma}_{jk} = \frac{1}{19}\sum_{u=0}^{1}\sum_{i=1}^{10}\left(y_{ij}^{(u)} - \frac{\bar{y}_j^{(0)} + \bar{y}_j^{(1)}}{2}\right)\left(y_{ik}^{(u)} - \frac{\bar{y}_k^{(0)} + \bar{y}_k^{(1)}}{2}\right)$$

対立仮説が真のとき，いわゆる疑似相関である　$\hat{\sigma}_{12}/\sqrt{\hat{\sigma}_{11}\hat{\sigma}_{22}}$　は $\tilde{\sigma}_{12}/\sqrt{\tilde{\sigma}_{11}\tilde{\sigma}_{22}}$ より大きくなる傾向があるため

疑似相関は仮説によっては意味のない値となるため，その利用は一般に FWER のコントロールを保証しないが，実は今のケースでは OK

人工データへの適用

対立仮説 $K_1^{(1)} \cap K_2^{(1)}$ の適当なモデルのもとで人工データを発生させ，多重検定問題を有意水準 5% でおこなうと，MaxT 法では二つの検定とも棄却されないのに対し，提案手法では二つの検定がともに棄却される



図（2 群 2 特徴量の人工データ）：群 1（青）の相関は 0.18，群 2（赤）の相関は 0.12，疑似相関は 0.30 である

反例

ケース群がもう一種あり，その独立な二次元データ $\{(y_{i1}^{(2)}, y_{i2}^{(2)}) \mid 1 \le i \le 10\}$ に対し，$u = 2$ として設定で述べたモデルが成立しているとする

$$H_1^{(2)} : \mu_1^{(2)} = \mu_1^{(0)} \quad \text{vs} \quad K_1^{(2)} : \mu_1^{(2)} > \mu_1^{(0)}$$
$$H_2^{(2)} : \mu_2^{(2)} = \mu_2^{(0)} \quad \text{vs} \quad K_2^{(2)} : \mu_2^{(2)} > \mu_2^{(0)}$$

を前述の二つの検定に加え，t 検定統計量をそれぞれ $T_1^{(2)}$ と $T_2^{(2)}$ と表し，$(T_1^{(1)}, T_2^{(1)}, T_1^{(2)}, T_2^{(2)})$ の漸近帰無分布となる四次元正規分布を考える

$$\hat{\sigma}_{jk} = \frac{1}{29}\sum_{u=0}^{2}\sum_{i=1}^{10}\left(y_{ij}^{(u)} - \frac{\bar{y}_j^{(0)} + \bar{y}_j^{(1)} + \bar{y}_j^{(2)}}{3}\right)\left(y_{ik}^{(u)} - \frac{\bar{y}_k^{(0)} + \bar{y}_k^{(1)} + \bar{y}_k^{(2)}}{3}\right)$$

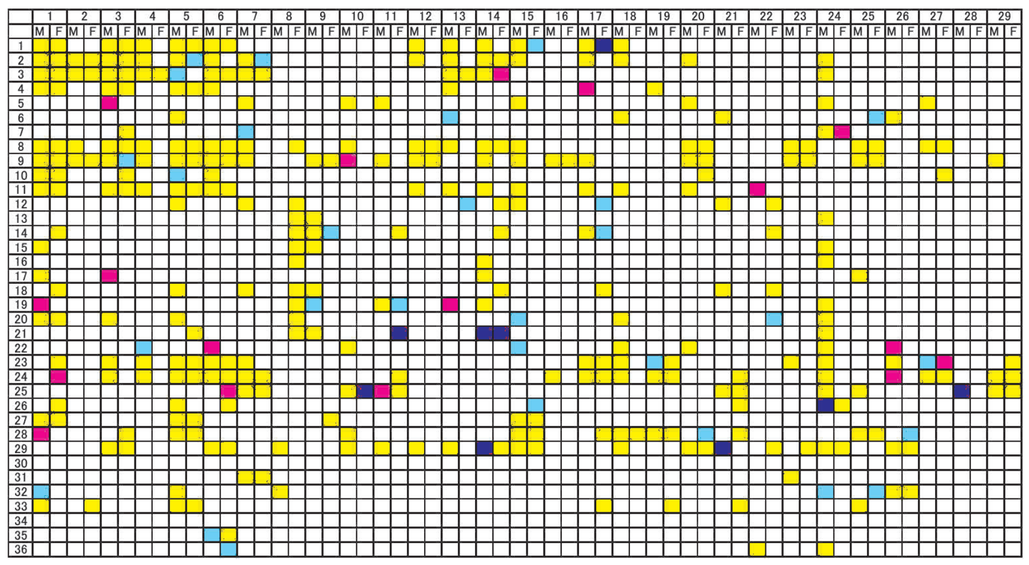
という $H_1^{(1)} \cap H_1^{(2)} \cap H_2^{(1)} \cap H_2^{(2)}$ を真としたモデルのもとでの σ_{jk} の不偏推定量を用いると，FWER がコントロールされない反例を作れる

数値実験

表（提案手法と既存手法の検出力比較）：数値は検出力，カッコ内の数値は平均 p 値，ρ は相関，p は検定数，μ は平均の差

ρ	n	p	μ	Bon		MaxT		SDMaxT		Proposal	
0.0	12	50	1.2	30.4	[34.4]	30.9	[33.8]	34.7	[28.8]	38.8	[23.7]
0.6	12	50	1.2	30.5	[34.1]	36.3	[25.1]	41.3	[23.3]	54.6	[16.5]
0.3	6	50	1.2	6.9	[65.9]	7.7	[59.6]	8.2	[58.6]	12.4	[50.3]
0.3	18	50	1.2	59.8	[14.1]	61.6	[12.5]	70.2	[8.7]	74.3	[7.0]
0.3	12	20	1.2	46.8	[20.4]	48.6	[17.1]	57.1	[13.4]	64.0	[10.3]
0.3	12	80	1.2	26.3	[39.9]	27.3	[35.6]	31.8	[32.2]	41.6	[24.7]
0.3	12	50	0.9	14.3	[55.0]	14.9	[50.6]	16.9	[48.5]	21.8	[42.3]
0.3	12	50	1.5	55.8	[14.9]	57.8	[12.9]	67.8	[8.8]	77.1	[5.6]

実データ解析



図（マウスの量的形質を比較する 36 個の検定からなる多重検定 58 種）：行は量的形質の種類，列はマウスの種類，黄色・水色・桃色・紺色は Bonferroni・MaxT・SDMaxT・Proposal で新たに棄却された仮説を表す

引用文献

Ninomiya, Y., Kuriki, S., Shiroishi, T. and Takada, T. (2021). A modification of MaxT procedure using spurious correlations. Journal of Statistical Planning and Inference, 214, 128–138.