2011年7月14日　統計数理研究所　オープンハウス

# 遺伝子発現差解析におけるP値推定

藤澤　洋徳　　数理・推論研究系　准教授

## 1. はじめに

遺伝子発現データに対する遺伝子発現差解析を考える．データの特徴は標本数は小さいのに遺伝子数が大きいことである．そのため，ある遺伝子に着目して遺伝子発現差解析を考えて，そのP値を並べ替え技法で推定するとき，標本数が小さいために，精度が低いという問題点がある．そこで，遺伝子数が大きいことに着目して，ターゲットとする遺伝子のデータだけでなく，他の遺伝子のデータを援用してP値を推定するという方法がしばしば採用されている．しかしながら，それぞれの遺伝子は，遺伝子発現差があったりなかったりで，等質ではない．そのため，他の遺伝子のデータを何の工夫もなしに援用することは，大きな問題を引き起こす．本研究では，他の遺伝子のデータを援用するためには，どのようにすれば良いかを議論する．

## 2. 遺伝子発現差解析

遺伝子発現値　（ある遺伝子上の二つのグループ$X$と$Y$に対して）

$$\boldsymbol{Z} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

帰無仮説　$H : \mu_X = \mu_Y$　（棄却されると遺伝子発現差あり）

通常の検定統計量　$N = n + m$

$$T(\boldsymbol{Z}) = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\{(n-1)S_X^2 + (m-1)S_Y^2\}/(N-2)}}$$

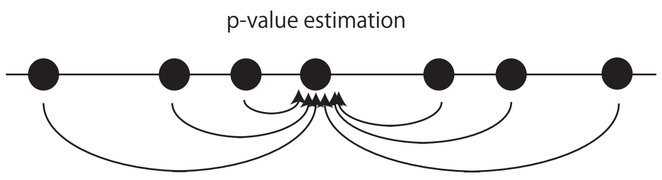
$p$値　$p = \Pr(T(\boldsymbol{Z}) > t \,|\, H)$　$t = T(\boldsymbol{z})$

$p$値推定　$\hat{p} = (1/B)\sum_{b=1}^{B} I(T(\boldsymbol{Z}_b^{\#}) > t)$

$\boldsymbol{Z}_b^{\#}$: $\boldsymbol{Z}$の第$b$番目の並べ替え標本

問題点　しばしば$n$と$m$は小さい．$n = m = 4$のとき$B = 8!/(4!4!) = 56$．
$B$が小さすぎる．0でない$p$値の推定値は高々0.02程度．

## 3. 他の遺伝子のデータを援用する




$p$値推定　（遺伝子$g$に対して）

$$\hat{p}\left(T(\boldsymbol{Z}^{\#}); t_g\right) = \frac{1}{BG}\sum_{b=1}^{B}\sum_{g'=1}^{G} I(T(\boldsymbol{Z}_{bg'}^{\#}) > t_g)$$

問題点　見掛け上の標本数$BG$は増えているが，
帰無仮説が正しい遺伝子と正しくない遺伝子が混在している．

## 4. 妥当な方法

検定統計量 (Pan, 2003)

$$T_{\mathrm{Pan}}(\boldsymbol{Z}) = \left|\frac{\bar{X}_{(1)} + \bar{X}_{(2)}}{2} - \frac{\bar{Y}_{(1)} + \bar{Y}_{(2)}}{2}\right| \Bigg/ \sqrt{\frac{1}{4}\left(\frac{S_{X(1)}^2}{n_1} + \frac{S_{X(2)}^2}{n_2} + \frac{S_{Y(1)}^2}{m_1} + \frac{S_{Y(2)}^2}{m_2}\right)}$$

帰無統計量

$$\begin{aligned} T_{\mathrm{Pan}}^{\mathrm{null}}(\boldsymbol{Z}) &= T_{\mathrm{Pan}}\left(\boldsymbol{X}_{(1)}, -\boldsymbol{X}_{(2)}, -\boldsymbol{Y}_{(1)}, \boldsymbol{Y}_{(2)}\right) \\ &= \left|\frac{\bar{X}_{(1)} - \bar{X}_{(2)}}{2} + \frac{\bar{Y}_{(1)} - \bar{Y}_{(2)}}{2}\right| \Bigg/ \sqrt{\frac{1}{4}\left(\frac{S_{X(1)}^2}{n_1} + \frac{S_{X(2)}^2}{n_2} + \frac{S_{Y(1)}^2}{m_1} + \frac{S_{Y(2)}^2}{m_2}\right)} \end{aligned}$$

問題点の克服　基礎となる確率分布が原点対称であるとき，
帰無統計量の分布は，帰無仮説$H$が正しいかどうかに依存しない！

検定統計量と帰無統計量の関係

$$p_{\mathrm{Pan}} = \Pr(T_{\mathrm{Pan}}(\boldsymbol{Z}) > t \,|\, H) = \Pr(T_{\mathrm{Pan}}^{\mathrm{null}}(\boldsymbol{Z}) > t)$$

帰無統計量を用いて経験的な$p$値推定が可能になる．

## 5. 最適な方法

より一般的な観点　他の遺伝子のデータを援用した$p$値経験推定が妥当になる検定統計量$T$の条件

次を満たす適当な帰無統計量$h(\boldsymbol{Z})$が存在する：
$(*)$　$p = \Pr(T(\boldsymbol{Z}) > t \,|\, H) = \Pr(h(\boldsymbol{Z}) > t)$.

ポイント！　この条件を満たす検定統計量$T$の適当なクラスを考え，
その中で最適な検定統計量を考える．（詳細は論文を見てください．）

基礎となる分布が原点対称の場合

$$T_{\mathrm{S}}(\boldsymbol{Z}) = |Q_{\mathrm{S}}|/\sqrt{S^2}$$

$$\begin{aligned} Q_{\mathrm{S}} &= \left(\frac{\bar{X}_{(1)} + \bar{X}_{(2)}}{2} - \frac{\bar{Y}_{(1)} + \bar{Y}_{(2)}}{2}\right) \Bigg/ \sqrt{\frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{m_1} + \frac{1}{m_2}\right)} \\ (N-4)S^2 &= (n_1-1)S_{X(1)}^2 + (n_2-1)S_{X(2)}^2 + (m_1-1)S_{Y(1)}^2 + (m_2-1)S_{Y(2)}^2 \end{aligned}$$

この検定統計量はPanの検定統計量とほぼ同じ．

基礎となる分布が同じ位置分布族に属している場合

$$T_{\mathrm{p}}(\boldsymbol{Z}) = |Q_{\mathrm{p}}|/\sqrt{\left\{(N-4)S^2 + Q_0^2\right\}/(N-3)}$$

$$\begin{aligned} Q_{\mathrm{p}} &= \left\{\frac{n_1 m_1}{n}(\bar{X}_{(1)} - \bar{Y}_{(1)}) + \frac{n_2 m_2}{m}(\bar{X}_{(2)} - \bar{Y}_{(2)})\right\} \Bigg/ \sqrt{\frac{n_1 m_1}{n} + \frac{n_2 m_2}{m}} \\ Q_0 &= \sqrt{\frac{n_1 m_1 n_2 m_2/nm}{n_1 m_1/n + n_2 m_2/m}}\left\{(\bar{X}_{(1)} - \bar{Y}_{(1)}) - (\bar{X}_{(2)} - \bar{Y}_{(2)})\right\} \end{aligned}$$

検定統計量$T_{\mathrm{p}}$は$T_{\mathrm{s}}$と比べると自由度が1小さい．普通は自由度1の差は小さいが，遺伝子発現差解析では標本数が小さいことがあるので，その場合は，パフォーマンスの非常に大きな改良となる．

## 6. 数値例

Golden Spike Data の解析 ($n = m = 3$)

| $\alpha$ | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | Bon* |
|---|---|---|---|---|---|
| | 検出された遺伝子数 | | | | |
| $T_{\mathrm{s}}$ | 340 | 9 | 0 | 0 | 0 |
| $T_{\mathrm{p}}$ | 1013 | 525 | 161 | 16 | 16 |
| | 正しく検出された遺伝子数 | | | | |
| $T_{\mathrm{s}}$ | 317 | 8 | 0 | 0 | 0 |
| $T_{\mathrm{p}}$ | 890 | 508 | 158 | 16 | 16 |

Bon*: 全体の有意水準を0.01とした場合のボンフェロニ補正．
$\alpha = 0.01/3866 \approx 2.59 \times 10^{-6}$.