異質性が内在する 2 標本検定問題におけるロバストな統計量

総合研究大学院大学　　*大前　勝弘
統計数理研究所　　小森　理
統計数理研究所　　江口　真透

1. 研究の背景と目的

２標本検定問題は，統計学の最も古典的な問題のうちの一つである．典型的には，比較したい２つの群から得られた標本を用いて，２群の乖離度が仮定や目的に応じた統計量を用いて測られる．代表的な統計量としては，t-統計量や Wilcoxon の順位和検定統計量が挙げられる．今回は，多重な２標本検定を利用した変数フィルタリングにおいて，一部の変数に異質性が内在する場合を考える．目的は，両群間の乖離度が大きい変数のうち，群内の異質性に依らない変数を同定することである．

2. 概要

群内の異質性をとらえるには，サブサンプリングが有用である．すなわち，比較したい２群から繰り返し小標本を取り直すことで，群の異質な情報をより反映する統計量が得られる．X_{ij}を標本 i 変数 j の観測値とし，各標本は群 0 あるいは群 1 のいずれかに属するとする．それぞれの群の標本数をn_0，n_1 とし，両群からサイズ a，b のサブサンプルを取ることを考える．このとき，このような小標本の取り方はそれぞれ $\mathrm{l} = 1, 2, \ldots, k_1$，$m = 1, 2, \ldots, k_0$ の組だけあり，サブサンプルで評価される t-統計量は，

$$U_j^T = \frac{1}{k_0 k_1} \sum_{l=1}^{k_1} \sum_{m=1}^{k_0} \frac{\sqrt{a+b}(\bar{X}_{j1l} - \bar{X}_{j0m})}{s_j}, \qquad (1)$$

ここで，$k_1 = {}_{n_1}\mathrm{C}_a$，$k_0 = {}_{n_0}\mathrm{C}_b$ で，$\bar{X}_{jyt}$は変数 j 群 y におけるt番目のサブセットの標本平均である．s_j に関してはいくつかのオプションが考えられるが．ここでは Welch タイプの non-pooled な標本分散とする．我々は，これらのサブサンプルで評価された t-統計量の符号の不安定性に着目し，次の符号和統計量を提案する．

$$U_j^S = \frac{1}{k_0 k_1} \sum_{l=1}^{k_1} \sum_{m=1}^{k_0} \mathrm{H}(\bar{X}_{j1l} - \bar{X}_{j0m}), \qquad (2)$$

ここで，H はヘヴィサイド関数である．これら(1)，(2)式は，その定義から U-統計量であることがわかる[1]．特に(2)式は，$a = 1$，$b = 1$ のとき，Wilcoxon の順位和検定統計量に一致する．本発表では，U-統計量の性質からこれら２つの統計量を比較し，符号和統計量が，群内の異質性に対してロバストであることを示す．この評価は U-統計量の漸近的性質およびシミュレーションによって与えられる．適用例として，遺伝子発現データによる癌関連遺伝子フィルタリングの例を用い，従来のいくつかのフィルタリング手法と比較して，符号和統計量による変数フィルタリングが安定した結果をもたらすことを示す．

参考文献

[1] Lehmann, E, L. (1999). *Elements of large-sample theory*, Springer.