○パーミュテーションを用いたマルチアレルHWE検定

岡田　随象

○背景: HWE検定おける下記問題点は、アレル数の増大により顕著になると考えられる。

・χ^2検定…χ^2値に対する不適切な漸近近似による、type I error の上昇。*1
・正確確率検定…計算負荷の増大 *2
⇒適切な計算負荷で、正確なP値を得る検定手法が必要である。

*1:Janis E. Wigginton et al. A Note on Exact Test of HWE. AJHG 76:887-:2005.
*2: Edward J. Louis et al. An Exact Test for HWE and Multiple Alleles. BIOMETRICS 43:805-:1987.

○目的: マルチアレル多型HWE検定の計算負荷を評価する。

○方法

①:HWE検定手法の定義

・1 locus における kアレル多型で構成された、Nサンプルに対するディプロタイプデータを対象とする。
・アレル i の観測度数を m_i, 頻度を、f_i とする。(i=1…k)
・アレル i, j で構成されるディプロタイプの観測度数を n_{ij}, 期待度数を e_{ij} とする。

①-1: χ^2検定

・下記定義式に従い、HWE χ^2統計量(=X_{HWE}) 及び対応するP値(=$P_{HWE, Chi}$) を求める。

$$E_{ij} = 2Nf_i f_j \quad (i \neq j),$$
$$Nf_i f_j \quad (i = j).$$
$$X_{HWE} = \sum_{i=1}^{k}\sum_{j=i}^{k} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}.$$
$$X_{HWE} \sim \chi^2_{(df)}, \quad df = \left(\frac{k(k+1)}{2} - 1\right) - (k-1) = \frac{k(k-1)}{2}.$$

①-2:Permutaion χ2検定

・ハプロタイプのパーミュテーションを行い、各ステップでのX_{HWE}を得る。
・得られた分布を、帰無仮説下でのX_{HWE}の分布と仮定し、パーミュテーションP値(=$P_{HWE, Perm}$)を求める。

①-3:正確確率検定

・与えられたアレル度数分布 (=**m**) のもとでの、ディプロタイプ度数分布 (=**n**) の生起確率 (=Pr) を①式の通り定義する。*2
・存在しうる全てのディプロタイプ度数分布について Pr を計算する。
・②式に従いP値 (=$P_{HWE, Exact}$) を求める。

$$\Pr(n \mid m) = \frac{N!\prod_i m_i!}{(2N)!\prod_{i\le j} n_{ij}!} 2^{\sum_{i<j} n_{ij}} \quad \ldots.① \qquad P_{HWE,Exact} = \sum_{\Pr \le \Pr_{Observed}} \Pr(n \mid m) \ldots.②$$

・Prの大小比較においては、可変部のみの比較とした。
・階乗値は対数変換値を配列変数に格納し、繰り返し使用した。

・全てのディプロタイプ度数分布の数え上げ手法

①-3-1:全ヘテロディプロタイプの列挙 (all)
・各ヘテロディプロタイプ毎に、0 ～最大値 までの組み合わせを列挙する。
・ホモディプロタイプ全てが非負整数を満たす組み合わせを採用する。
⇒採用される割合が低く、非効率的

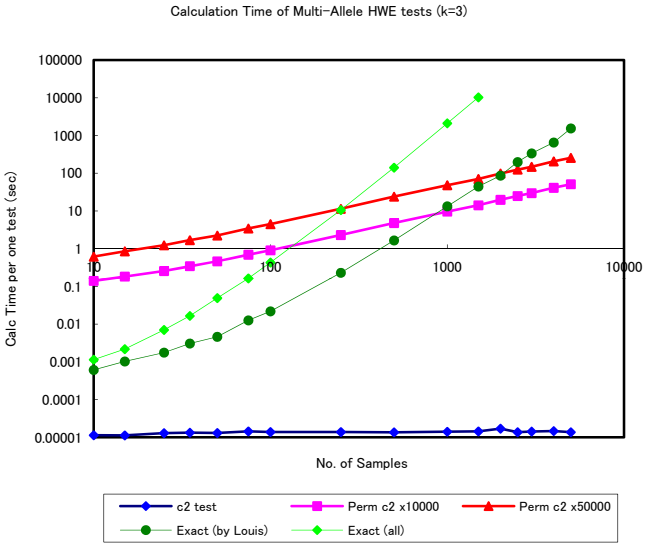
①-3-2:Louis et al による再帰的手法 (by Louis) *2
(1): kアレルのうち、アレル観測度数が最も少ないアレル(=m_A)を選択する。
(2): n_{AA}の度数を 0～[m_A/2] の範囲で+2ずつ変化させる。
(3): 各n_{AA}の度数において、$\mathbf{n_{A*}}$の組み合わせを網羅する。
(4): 各組み合わせににおいて、k-1アレルを対象とし、手順(1)へ戻る。
⇒有効なディプロタイプ組み合わせのみを網羅するため、効率的

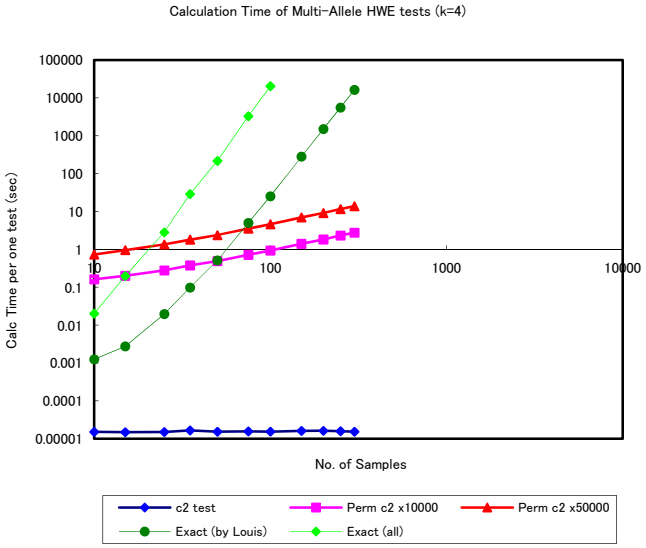
②:対象ジェノタイプデータの作成
・k=3, f=(0.333, 0.333, 0.333) 及び k=4, f=(0.25, 0.25, 0.25, 0.25)、n=10～5000 を対象とした。
・HWE平衡を仮定してシミュレーションデータを作成した。

③:計算時間の測定
・各シミュレーションデータに対し下記検定手法を実行した。
χ^2検定、Permutation χ^2検定(x10000, x50000)、正確確率検定(all, by Louis)
・各検定実行前後で時刻を測定し、差分を計算時間とした。

・各条件下で10個のデータを作成し、計算時間の平均値を取得した。
・計算環境：Dual Core AMD Opteron 2.85GHz @ solaris 10.

○：結果





①-1：χ^2検定
・サンプル数N・アレル数kに応じた計算時間の増加は認められなかった。
・約10^-5秒と、手法間で最も早い計算時間であった。

①-2：Permutaion χ2検定
・N・permutation回数に比例した計算時間の増加が認められた。
・kに応じた増加は認められなかった。

①-3：Exact検定
・N・kに応じた計算速度の増加を認めた。
・Louis et al によるアルゴリズム(by Louis)は、全ヘテロディプロタイプの列挙(all)より小さい計算時間を達成した。
・Nに対する増加は一次より大きく、Nが大きいときは①-1,2と比較して多大な計算時間が必要となった。

⇒（もしPermutation P値が適切なP値を出力するのであるならば）、
k=4, N=100 が正確率検定に替えてPermutation法が推奨される一基準と考えられた。

○：正確確率検定(all) における計算負荷

計算負荷 = 対数階乗値の計算 + 共通部分の計算 + 可変部分の計算 * 組み合わせ数 + α
≒ $O(N^2/2) + O(N+N+2N) + O(N+df)*O((2N/k+1)^{df})$
≦ $O(N^2) + O(N)*O((N)^{df})$ （ df=k(k-1)/2, N>>k ）
＝ $O(N^{(df+1)})$

⇒k=3・4 のとき、≦$O(N^4)$・$O(N^7)$ となり、観測値から得られた$O(N^{3.73})$・$O(N^{6.37})$ と矛盾しない結果となった。