ゲノム網羅的解析における薬物動態関連遺伝子検出のための スチューデント化修正最大対比法の提案

長島 健悟
城西大学 薬学部

Kengo Nagashima
Faculty of Pharmaceutical Sciences, Josai University

統計数理研究所リスク解析戦略研究センター研究会
2010 年 3 月 26 日

薬物動態関連遺伝子

薬物動態 (pharmacokinetics)

- 薬物の体内動態 (体内での動き) をいくつかのパラメータに要約
  - 吸収量: 薬物血中濃度–時間曲線下面積 AUC
  - 吸収速度: 最高血中濃度 $C_{\max}$
  - 代謝や排泄: 消失速度定数 K_{el}

薬物動態関連遺伝子

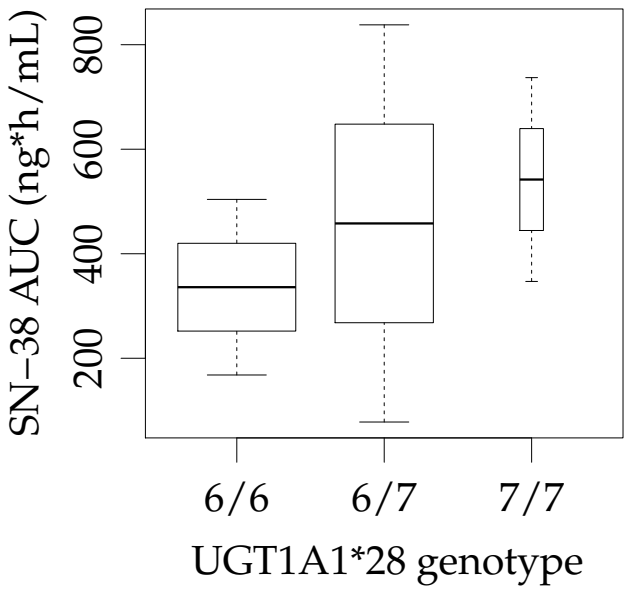
- 薬物に対する曝露量を計る指標, 薬効や副作用の情報を与える
- 薬物動態関連遺伝子は, 薬物の体内動態に関連する遺伝子
- 個別化医療などに繋がる

薬物動態関連遺伝子の例

UGT1A1 遺伝子 (Innocenti, et al. 2004. [7])

- 塩酸イリノテカン (CPT-11) の副作用発現に関与
- UGT1A1*28, UGT1A1*6 などの多型が特定の遺伝子型を持つ場合
  UDP グルクロン酸転移酵素 (UGT) の活性が低下
  薬剤 (中間代謝物 SN-38) の代謝, 排泄速度が低下
  好中球減少や重篤な下痢などの副作用の発現が高まる

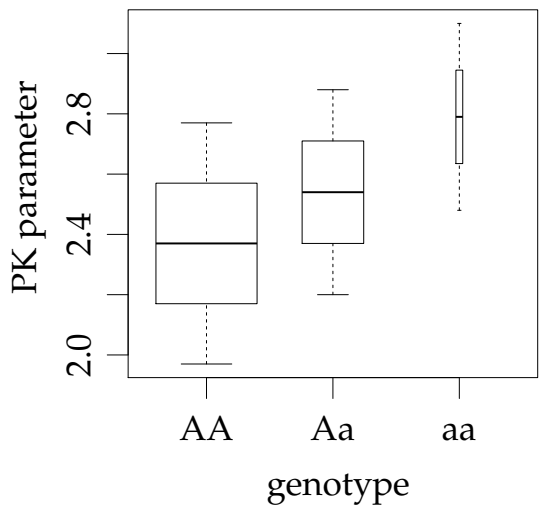
予測に利用できる新規バイオマーカーの同定が期待され, 薬物動態関連遺伝子の探索研究が数多く行われている



薬物動態関連遺伝子のスクリーニングにおけるデータ構造

測定項目

- 薬物動態パラメータ
  左に裾を引いた分布, 経験的に対数正規分布 (Gabrielsson, Weiner. 2000.[3])
- 遺伝子情報
  本研究では SNPs を想定 (10 万 ~100 万), "AA", "Aa", "aa"の三群



データ構造

データ構造の続きと記号の定義

- ANOVA 型のデータ構造
- 第 i 群, j 番目の測定値を表わす確率変数 Y_{ij}
  $i = 1$: "AA", 2: "Aa", 3: "aa"とする
- 薬物動態パラメータの測定値を対数変換し $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$ を仮定する

- 第 i 群の標本平均 $\bar{Y}_i$, 標本平均ベクトル $\mathbf{Y} = (\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)^{\mathrm{t}}$
- 併合不偏分散 $V = \frac{1}{\gamma} \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$
  とその自由度 $\gamma = \sum_{i=1}^{a} (n_i - 1)$
- 対角行列 $\mathbf{D} = \mathrm{diag}\left(\frac{1}{n_1}, \frac{1}{n_2}, \frac{1}{n_3}\right)$

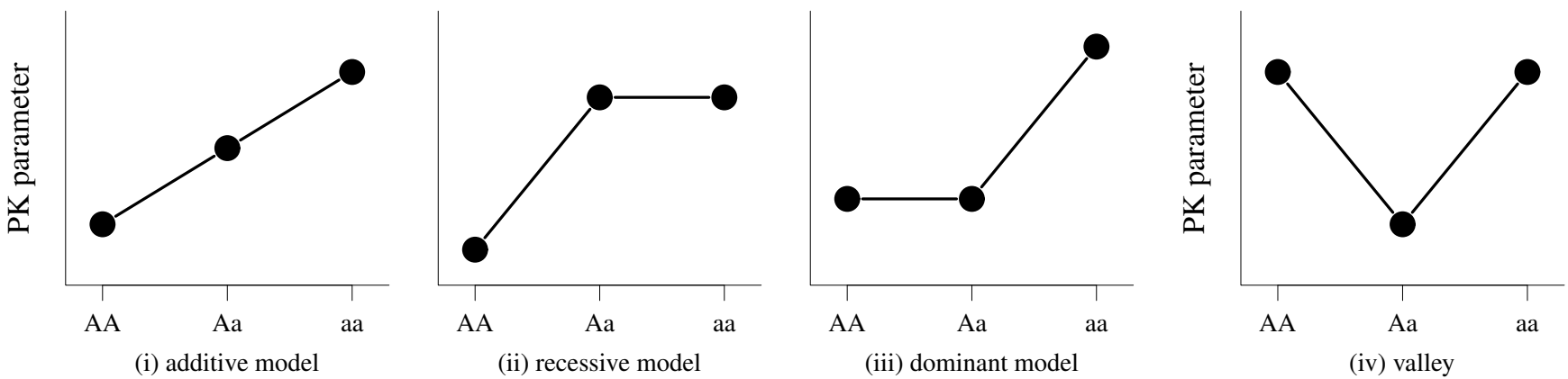
- 母平均ベクトル $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^{\mathrm{t}}$

既存法

- 対数変換した後に ANOVA (対数正規分布するから)
- Kruskal–Wallis 検定

この後に . . .

- 検出された遺伝子の反応関係を 目視 でチェック
  OK (i)~(iii) 生物学的に妥当な反応関係
  NG (iv) 生物学的にはまず起こりえない反応関係



既存法の問題点

ANOVA や Kruskal–Wallis 検定は
大量の (iv) 谷型パターンを検出

理由

- 包括的帰無仮説 $H_0 : \mu_1 = \mu_2 = \mu_3$, 対立仮説 $H_1 :$ not H_0 の仮説検定
- 対立仮説には谷型を含むあらゆるパターン

谷型パターンの検出を減らすことが出来ないか？

最大対比法

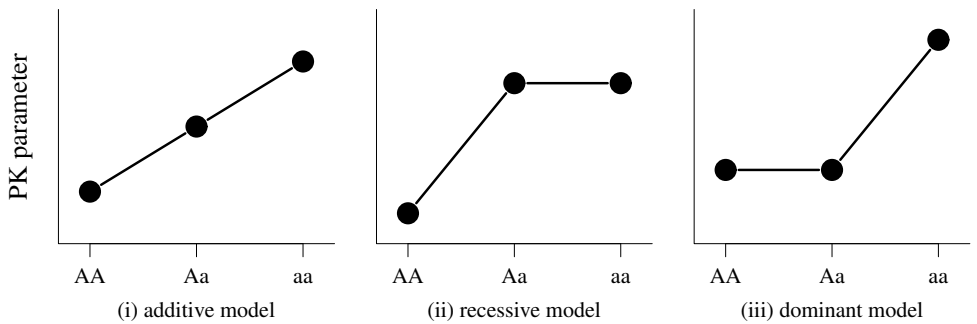
- 包括的帰無仮説 $H_0 : \mu_1 = \mu_2 = \mu_3$, 対立仮説 $H_1 : \mathbf{C}\boldsymbol{\mu} > \mathbf{0}$ の仮説検定
- $\mathbf{C}$ は対比係数行列とよばれ, m 個の対比係数ベクトル $\mathbf{c}_k$ で構成

$$\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m)^{\mathrm{t}}, \quad \mathbf{c}_k = (c_{k1}, c_{k2}, c_{k3})^{\mathrm{t}}$$

$$\mathbf{c}_k^{\mathrm{t}}\boldsymbol{\varepsilon} = \mathbf{0}, \boldsymbol{\varepsilon} = (1, 1, 1)^{\mathrm{t}}$$

例

- $\mathbf{c}_1 = (-1, 0, 1)^{\mathrm{t}}, \mathbf{c}_2 = (-1, 1, 1)^{\mathrm{t}}, \mathbf{c}_3 = (-1, -1, 1)^{\mathrm{t}}$ のとき



- 対立仮説 $H_1 : \mu_1 < \mu_2 < \mu_3,\ \mu_1 < \mu_2 = \mu_3,\ \mu_1 = \mu_2 < \mu_3$

対立仮説に制限が加わる

最大対比法

$$T_{\max} = \max\left(T_1, T_2, \ldots, T_k, \ldots, T_m\right), \quad T_k = \frac{\mathbf{c}_k^{\mathrm{t}} \bar{\mathbf{Y}}}{\sqrt{V \mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k}} \tag{1}$$

- P 値は同時分布の積分 ($t_{\max}$ は観測値)

$$\begin{aligned} P\text{-value} &= \Pr(T_{\max} > t_{\max}) \\ &= 1 - \Pr(T_1 \leq t_{\max}, T_2 \leq t_{\max}, \ldots, T_k \leq t_{\max}, \ldots, T_m \leq t_{\max}) \end{aligned}$$

- H_0 のもとでの $\mathbf{T} = (T_1, T_2, \ldots, T_k, \ldots, T_m)^{\mathrm{t}}$ の同時分布

$$\mathbf{T} \sim t_m(\gamma, \mathbf{R}_T)$$

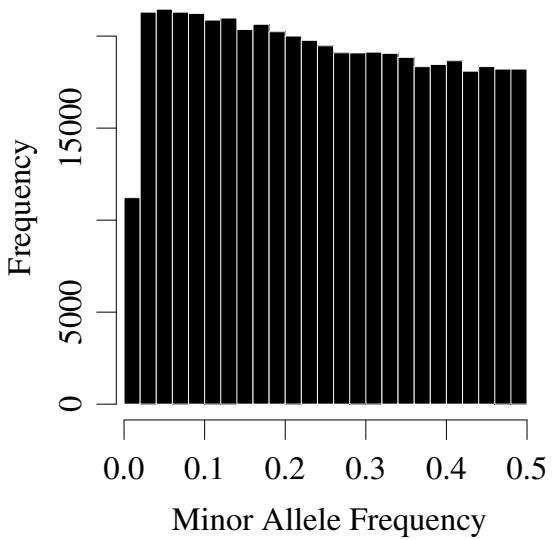
$$\mathbf{R}_T \text{の第 } (k,l) \text{ 要素} = \frac{\mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_l}{\sqrt{\mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k} \sqrt{\mathbf{c}_l^{\mathrm{t}} \mathbf{D} \mathbf{c}_l}}$$

最大対比法の問題点とマイナー対立遺伝子頻度

- 不等標本の場合, (1) 式の分母 (分散に相当) が不安定

$$\sqrt{V\mathbf{c}_k^{\mathrm{t}}\mathbf{D}\mathbf{c}_k} = \sqrt{V\left(\frac{c_{k1}^2}{n_1} + \frac{c_{k2}^2}{n_2} + \frac{c_{k3}^2}{n_3}\right)}$$

- マイナー対立遺伝子頻度 (Minor Allele Frequency; MAF)
  集団における, ある SNP から求まる"a"の頻度 $\frac{n_2+2n_3}{2(n_1+n_2+n_3)}$



JSNP データベースの MAF の分布

$n = 300$ の場合の MAF と (n_1, n_2, n_3)

MAF	$n = 300$		
	n_1	n_2	n_3
0.12	234	61	5
0.25	168	113	19
0.33	133	133	34
0.5	75	150	75

※ HWE を仮定した場合

修正最大対比法

$$M_{\max} = \max\left(M_1, M_2, \ldots, M_k, \ldots, M_m\right), \quad M_k = \frac{\mathbf{c}_k^{\mathrm{t}} \bar{\mathbf{Y}}}{\sqrt{\mathbf{c}_k^{\mathrm{t}} \mathbf{c}_k}}$$

- H_0 のもとでの $\mathbf{M} = (M_1, M_2, \ldots, M_k, \ldots, M_m)^{\mathrm{t}}$ の同時分布

$$\mathbf{M} \sim \mathrm{N}_m(\mathbf{0}, \mathbf{R}_M)$$

$$\mathbf{R}_M\text{の第 } (k,l) \text{ 要素} = \sigma^2 \frac{\mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_l}{\sqrt{\mathbf{c}_k^{\mathrm{t}} \mathbf{c}_k} \sqrt{\mathbf{c}_l^{\mathrm{t}} \mathbf{c}_l}}$$

- 未知パラメータ σ^2 を含むので, P 値をリサンプリング法で近似

(Westfall, Young 1993[12])

修正最大対比法とその問題点

- 最大対比法の弱点を補うようなパフォーマンス
  組み合わせた適用を提案 (Sato et al. 2009[1])

Sato et al. 2009[1] による性能評価

	(i) additive 	(ii) recessive 	(iii) dominant 	(iv) valley 
修正最大対比法	○	○	△	○
最大対比法	△	△	○	△
Kruskal-Wallis 検定	△	△	×	×

リサンプリング法は計算に非常に時間がかかる

- $n = 250$ 程度, 10 万 SNPs 程度のデータに適用すると数週間
  一般的な PC 上で計算 (Intel Core 2 Duo 3.0 GHz)

本研究の目的

修正最大対比法の改良 (計算時間減・精度向上)

修正最大対比法と同等の性能を持ち
帰無分布が求まる統計量を提案

スチューデント化修正最大対比統計量

$$S_{\max} = \max\left(S_1, S_2, \ldots, S_k, \ldots, S_m\right), \quad S_k = \frac{\mathbf{c}_k^{\mathrm{t}} \bar{\mathbf{Y}}}{\sqrt{V \mathbf{c}_k^{\mathrm{t}} \mathbf{c}_k}}$$

- P 値は同時分布の積分 ($s_{\max}$ は観測値)

$$P\text{-value} = 1 - \Pr(S_1 \leq s_{\max}, S_2 \leq s_{\max}, \ldots, S_k \leq s_{\max}, \ldots, S_m \leq s_{\max})$$

- H_0 のもとでの $\mathbf{S} = (S_1, S_2, \ldots, S_k, \ldots, S_m)^{\mathrm{t}}$ の同時分布

$$\mathbf{S} \sim t_m(\gamma, \mathbf{R}_S)$$

$$\mathbf{R}_S \text{の第 } (k,l) \text{ 要素} = \frac{\mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_l}{\sqrt{\mathbf{c}_k^{\mathrm{t}} \mathbf{c}_k} \sqrt{\mathbf{c}_l^{\mathrm{t}} \mathbf{c}_l}}$$

多変量 t 分布の積分について

- R の mvtnorm package や, SAS/IML の関数が利用可能

(Genz and Bertz 1999, 2002[4, 5])

- ランダム化準モンテカルロ法
  - モンテカルロ法と比較して精度が高い場合が多い

(Niederreiter 1987[9], Lemieux 2009[8])

- Separation-of-Variables Methods
  - 非積分関数を分散が小さくなるように変換

(Genz and Bertz 1999, 2002[4, 5])

性能評価

目的

- 修正最大対比法に対する, スチューデント化修正最大対比法の精度・計算速度の改善を評価

条件

- 同一精度: $3.5\times$ モンテカルロ誤差 $< 10^{-2}$
- 仮想データは $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$ を仮定して生成
  - $\mu_i = \Delta \times c_{ki}$, $\sigma^2 = 1$
  - 包括的帰無仮説 $\Delta = 0$, 反応パターンを 4 通り
    $\mathbf{c}_1 = (-1/2, 0, 1/2)^{\mathrm{t}}$, $\mathbf{c}_2 = (-1/3, 1/3, 2/3)^{\mathrm{t}}$,
    $\mathbf{c}_3 = (-2/3, -1/3, 1/3)^{\mathrm{t}}$, $\mathbf{c}_4 = (1/3, -2/3, 1/3)^{\mathrm{t}}$
- $n = 300$, MAF$= 0.12, 0.25, 0.33, 0.5$ とし, n_1, n_2, n_3 を決定
- 用いる対比係数行列は両方とも $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3)^{\mathrm{t}}$

評価指標

- 100 回の実行時間

結果および考察

計算時間の比較

条件	Δ	MAF	計算時間 (秒)	
			MMCM	sMMCM
overall null hypothesis	0	0.33	298.25	0.92
(i) additive	0.25	0.12	94.77	0.90
(ii) dominant	1	0.5	16.78	0.92
(iii) recessive	0.5	0.25	71.85	0.90
(iv) valley	0.25	0.33	254.31	0.92

MMCM: 修正最大対比法; sMMCM, スチューデント化修正最大対比法

20~300 倍速い

- 帰無仮説に近いほど時間がかかる (モンテカルロ誤差が大きいため)
  - 実際は多くの SNPs が帰無仮説に近いと思われるため, 適用する上で好都合

まとめと今後の課題

まとめ

- 精度・計算速度を改善したスチューデント化修正最大対比法を提案
- 適用可能性が高まった (R package 公開済)
  - $3.5\times$ モンテカルロ誤差 $< 10^{-4}$ で 1 日程度 (スライド 12 と同程度の規模)

今後の課題

- $\sqrt{V\mathbf{c}_k^{\mathrm{t}} \bigcirc \mathbf{c}_k} \to \sqrt{V\mathbf{c}_k^{\mathrm{t}}\mathbf{G}\mathbf{c}_k}$, $\mathbf{G} = \mathrm{diag}\,(g_1, g_2, \ldots, g_a)$ のような一般化
- 不等分散の場合に拡張

参考文献 I

[1] Sato Y, Laird NM, Nagashima K, Kato R, Hamano H, Yafune A, Kaniwa N, Saito Y, Sugiyama E, Kim S-R, Furuse J, Ishii H, Ueno H, Okusaka T, Saijo N, Sawada J, Yoshida T. A new statistical screening approach for finding pharmacokinetics-related genes in genome-wide studies. *The Pharmacogenomics Journal* 2009; **9**: 137–146.

[2] Evans WE, McLeod HL. Pharmacogenomics — drug disposition, drug targets, and side effects. *The New England Journal of Medicine* 2003; **348**(6): 538–549.

[3] Gabrielsson J, Weiner D. *Pharmacokinetic and pharmacodynamic data analysis: concepts and applications*. Taylor & Francis: Sweden, 2000.

[4] Genz A, Bretz F. Numerical computation of multivariate t probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation* 1999; **63**: 361–378.

[5] Genz A, Bretz F. Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics* 2002; **11**: 950–971.

[6] Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Research* 2002; **30**: 158–162.

参考文献 II

[7] Innocenti F, Undevia SD, Iyer L, Chen PX, Das S, Kocherginsky M, Karrison T, Janisch L, Ramirez J, Rudin CM, Vokes EE, Ratain MJ. Genetic variants in the UDP-glucuronosyltransferase 1A1 gene predict the risk of severe neutropenia of Irinotecan. *Journal of Clinical Oncology* 2004; **22**(8): 1382–1388.

[8] Lemieux C. *Monte Carlo and quasi-Monte Carlo sampling*. Springer, 2009.

[9] Niederreiter H. *Random number generation and quasi-Monte Carlo methods*. Society for Industrial Mathematics, 1987.

[10] Sugiyama E, Kaniwa N, Kim S R, Kikura-Hanajiri R, Hasegawa R, Maekawa K, Saito Y, Ozawa S, Sawada J, Kamatani N, Furuse J, Ishii H, Yoshida T, Ueno H, Okusaka T, Saijo N. Pharmacokinetics of gemcitabine in Japanese cancer patients: the impact of a cytidine deaminase polymorphism. *Journal of Clinical Oncology* 2007; **25**(1): 32–42.

[11] The International HapMap Consortium. The international HapMap project. *Nature* 2003; **426**: 789–796.

[12] Westfall PH, Young SS. *Resampling-based multiple testing: Examples and methods for p-Value adjustment (Wiley series in probability and statistics)*. New York: John Wiley & Sons, Inc. 1993.

[13] Yoshimura I, Wakana A, Hamada C. A performance comparison of maximum contrast methods to detect dose dependency. *Drug Information Journal* 1997; **31**: 423–432.

モンテカルロ誤差

- 確率変数 X の期待値 $\theta = \mathrm{E}(X)$ について

$$\hat{\theta}_n = \frac{1}{N}\sum_{i=1}^{n} X_i$$

により, モンテカルロ法による近似を行った場合の標準誤差は

$$\hat{\sigma^2_N} = \hat{\sigma^2}/N = \left(\frac{1}{N-1}\sum_{i=1}^{n}(X_i - \hat{\theta_N})^2\right)/N$$

であり, これを用いて誤差評価を行う

- 通常は信頼区間 $\left[\hat{\theta_N} - k\sqrt{\hat{\sigma^2_N}},\ \hat{\theta_N} + k\sqrt{\hat{\sigma^2_N}}\right]$ の幅がある一定の値以下になるように N を決定する
  中心極限定理から $k = 2$ の場合が約 95%信頼区間、$k = 3$ の時は約 99%信頼区間

R の mvtnorm package (Genz and Bertz 2002[5])

$$\hat{\mathbf{T}}_{N,P} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2P}\sum_{j=1}^{P}\left[f(|2\{\mathbf{p}_j + \mathbf{w}_i\} - 1|) + f(1 - |2\{\mathbf{p}_j + \mathbf{w}_i\} - 1|)\right]$$

- $\mathbf{p}_j + \mathbf{w}_i$ が quasi-random points
- 関数を 2 つに分けているのは, モンテカルロ法における分散減少法の一つ

$$Q_P(\mathbf{w}) = \frac{1}{2P}\sum_{j=1}^{P}\left[f(|2\{\mathbf{p}_j + \mathbf{w}\} - 1|) + f(1 - |2\{\mathbf{p}_j + \mathbf{w}\} - 1|)\right]$$

$$\sigma_N^2 = \frac{1}{N(N-1)}\sum_{i=1}^{N}\left(Q_P(\mathbf{w}_i) - \hat{\mathbf{T}}_{N,P}\right)^2$$

- $3.5\sigma_N^2 < \epsilon$ を用いたと記載

The permutation method for the modified statistic $M_{\max}$

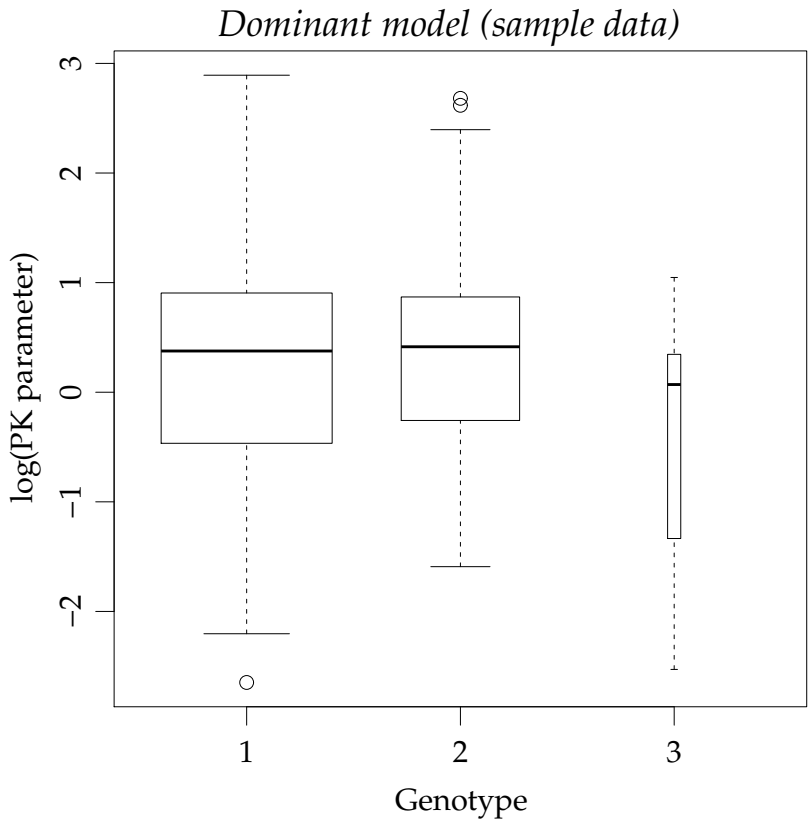
1. Initialize counting variable: $COUNT = 0$.
   Input parameters: $NRESAMPMIN$ (minimum resampling count, set to 1000), $NRESAMPMAX$ (maximum resampling count), and ϵ (absolute error tolerance).
2. Calculate $m_{\max}$, the observed value of the test statistic.
3. Let $y_{ij}^{(r)}$ donate data, which are sampled without replacement and independently form observed value y_{ij}. Here r is the resampling index ($r = 1, 2, \ldots, NRESAMP$).
4. Calculate $m_{\max}^{(r)}$ from $y_{ij}^{(r)}$. If $m_{\max}^{(r)} > m_{\max}$, then increment the counting variable: $COUNT = COUNT + 1$. Calculate approximate P-value $\hat{p}^{(r)} = COUNT/r$, and the simulation standard error $\hat{\sigma}^{(r)} = \mathrm{SE}(\hat{p}^{(r)}) = \sqrt{\hat{p}^{(r)}(1 - \hat{p}^{(r)})/r}$.
5. Repeat steps 3–4, while $r > NRESAMPMIN$ and $3.5\hat{\sigma}^{(r)} < \epsilon$ (corresponding to corresponding to an approximate confidence level of 99.95%; this is the accuracy of the randomized quasi-Monte-Carlo method of Genz and Bretz 2002), or $NRESAMPMAX$ times. Output the approximate P-value $\hat{p}^{(r)}$ and the standard error $\hat{\sigma}^{(r)}$.

mmcm package (R パッケージ)

```
library(mmcm)

set.seed(136885)
x <- c(
  rnorm(130, mean =  1 / 6, sd = 1),
  rnorm( 90, mean =  1 / 6, sd = 1),
  rnorm( 10, mean = -2 / 6, sd = 1)
)
g <- rep(1:3, c(130, 90, 10))
boxplot(
  x ~ g,
  width=c(length(g[g==1]),length(g[g==2]),
    length(g[g==3])),
  main="Dominant model (sample data)",
  xlab="Genotype", ylab="log(PK parameter)"
)

# coefficient matrix
# c_1: additive, c_2: recessive, c_3: dominant
contrast <- rbind(
  c(-1, 0, 1), c(-2, 1, 1), c(-1, -1, 2)
)
y <- mmcm.mvt(x, g, contrast)
y
```




```
  Studentized modified maximum contrast method

observed statistics = 0.535
contrast = 3    (-1 -1 2)
P-value = 0.044844
```