多重検定法

㈱ヒューマノーム研究所・代表取締役社長
(兼務)産業技術総合研究所・AIRC・招聘研究員
瀬々　潤
sesejun@humanome.jp

目次

科学は再現性の時代へ

- Finantial Times. 2014年3月

科学は再現性の時代へ

- Nature. 2014年1月
  - NIH plans to enhance reproducibility by Collins and Tabak
- The Lancet. 2014年1月
  - increasing value, reducing wasteと題して5本の論文を掲載.
    - 再現性を上げるための方策を議論
  - その中で，以下の再現性の問題が触れられている(Ioannidis の稿)
    - Bayerの研究者のレポートによると，67件中43件が再現できない
    - Amgenの研究者のレポートによると，53件中47件が再現できない
- 情報分野でも，再現性の担保の仕組みが模索されている
  - SIGMOD Reproducibility
  - プログラムの公開，データの共有

問題と対策

- なぜ，このような問題が起きるのか？
  - 計算をするのが大変な時代から，計算してあたりまえな時代へ
  - 再現データを取得するのが困難．再現性が無い．
  - そもそも，扱うデータが増えてくると，偶発的に起きたものか，本当に「新規の」現象なのかの判断が難しくなる
- 数理・計算機的には２つの対策
  - 統計的有意性：「まれに」しか起きない現象であることを確認する
    - 創薬を始め，広く利用されている．
    - 基準は「正答率」ではない
  - 交差検証（Cross Validation）：学習に用いたデータとは，独立したデータを用意して，確からしさを確認する．
    - ROCカーブ等

統計的有意性を
利用した特徴選択

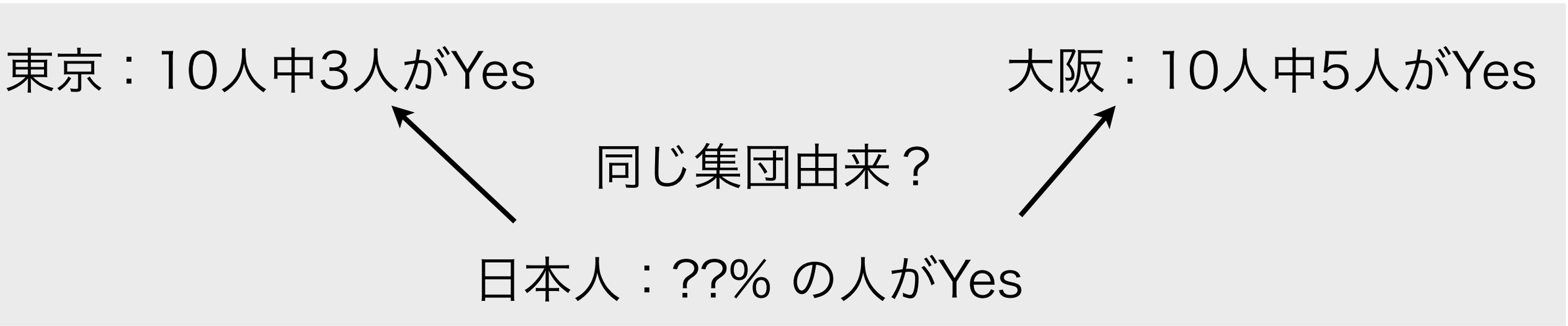
目次

「検定」とは？

- ざっくばらんに言って
  - 観察した2つ（あるいは，それ以上）の集団に，**差があるか無いかを調べる**ために，使うもの

東京：10人中3人がYes ⟷ 差がある？ ⟷ 大阪：10人中5人がYes

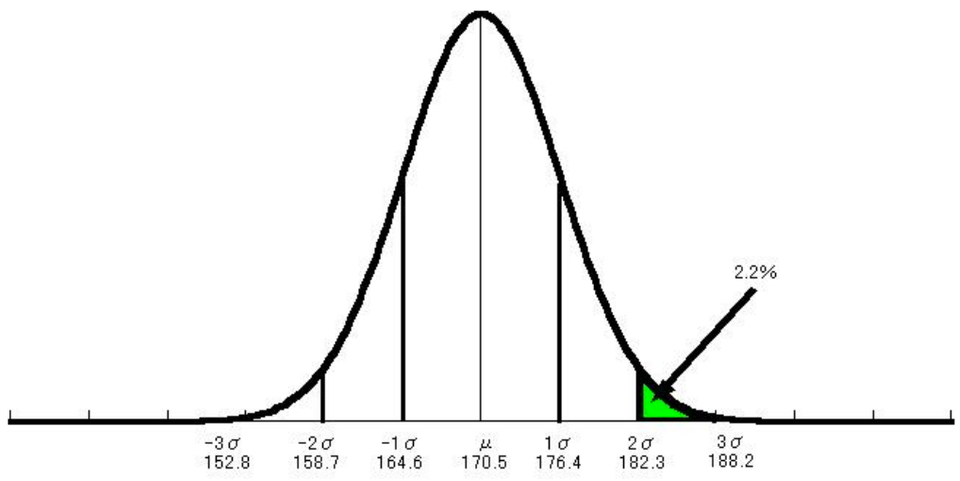
- 少しだけ厳密に言って
  - 観察した２つの集団は，**同一の集団から得られた**ものか？



- なぜ、そんなことを考えるのか？
  - 2回実験したら、2回同じ結果とは限らない←**ばらつきを考える**

「有意である（有意差がある）」とは？

- めったに起こらないだろうという，**差が生まれている**
  - 同一集団から得られた結果とは思えない状態
- データが与えられた時，同一集団から**それ以上の差が得られるだろう確率＝P値**
  - 数が小さいほど，「まれ」な差である.
- 「めったに起こらない」の**基準は，大抵は5%** (p ≦ 0.05)
  - 同一集団から，ランダムに選択した時，5%未満の確率でしか起こらないような差.
- **レアケースなので「有意差がある」 と判断**する.
  - = 同一集団から得られた，という仮説に，**ダウトと言う**



少数サンプルから確率を推定する

- 全体サンプル（＝母集団）があれば、どれくらいレアか判断する事は容易
  - しかしながら，**全部は調べられない**
- 今ある結果から**推定する**（少数のサンプルで全体を推定）

- Fisher’s exact test (フィッシャーの正確確率検定)の例：

	▲	✕	Total
High	2	3	5
Low	3	2	5
Total	5	5	10

p値は？

1. 観測された以上に「偏り」がある場合を列挙する

	▲	✕	Total
High	**1**	4	5
Low	4	1	5
Total	5	5	10

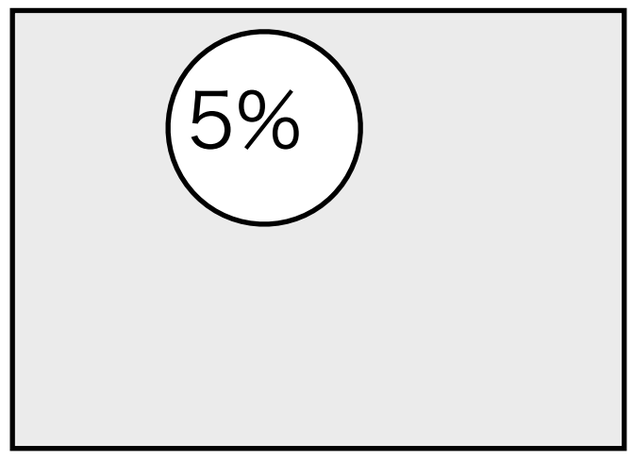
	▲	✕	Total
High	**0**	5	5
Low	5	0	5
Total	5	5	10

2. 生起確率の和を計算する

より偏りのある状態の総計＝p値

有意水準以下なら，発見とする

- 検定では，予め有意水準 (significance level)を指定する
- 有意水準α=0.05とは？
  - 100回実験したら，5回未満しか現れないくらい，レアなケースを取り出す
- 言い換えると
  - α=0.05で100回比較を行ったら，差がない場合でも5回は「差がある」と考えられてしまう
- それでも，p<0.05となるケースは稀なので，p<0.05であるときには「差がある」と言うことが多い
  - p<0.05の時，発見とみなされる.
- 検定は本来，強力なブレーキとしての機能



実験機器やセンサーの発達によって複数の検定が行われるようになった

- 薬候補と治癒の関係が知りたい
  - 薬A が 疾病X の治癒と関係しているか？→検定
  - 薬B が 疾病X の治癒と関係しているか？→検定
  - 薬C が 疾病X の治癒と関係しているか？→検定
  - ・・・沢山の検定が必要
- 大量の検定が必要となる場合
  - 遺伝子やSNPs毎の検定(遺伝子発現量，GWAS)
  - （遺伝子）機能毎の検定（Enrichment解析）
  - 創薬におけるハイスループットスクリーニング
  - モチーフ毎の検定（機能しているモチーフの発見）
  - 脳機能解析における画像解析（発火部位の特定）
  - 材料開発の素材探索（どの元素が重要か）
  - 大規模アンケートの解析

	変異				疾患
被験者	s_1	s_2	s_3	s_4	c
t_1	0	1	1	0	1
t_2	1	1	1	0	1
t_3	0	1	1	1	1
t_4	1	0	0	0	0
t_5	0	1	1	1	0
t_6	1	1	0	0	0
t_7	1	0	0	1	0
t_8	0	1	0	0	0

検査項目

ユーザ
（患者）

一般的なビッグデータ

ユーザの拡大

ライフ系
ビッグデータ

項目の拡大

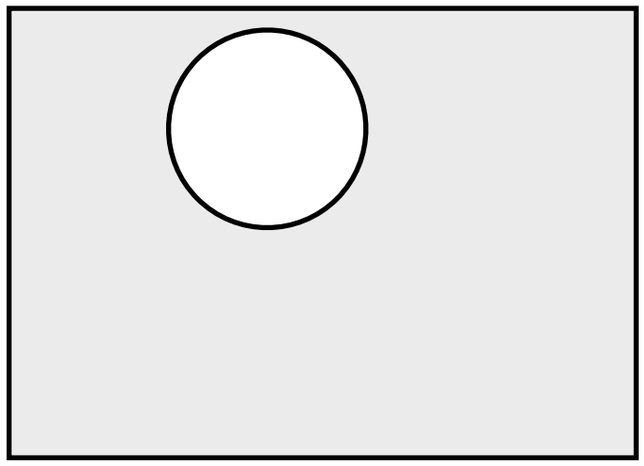
ゲノム、遺伝子発現、メチル化、プロテオーム、
メタボローム、環境情報、画像情報、etc

検定を複数回行うと，偽陽性を生む

有意水準α

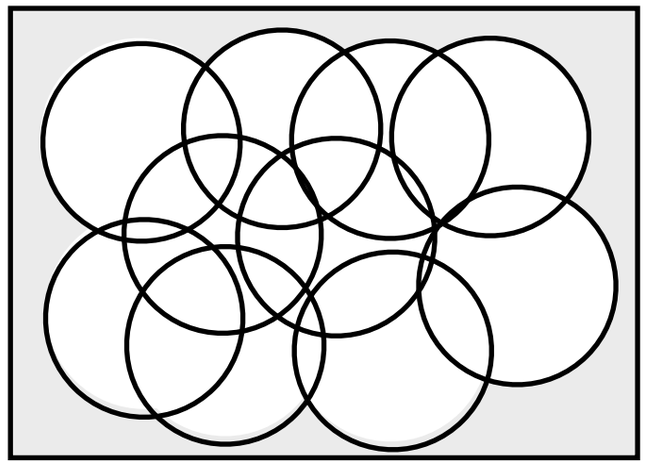
5%

1回の検定



偽陽性の起こる確率　5%

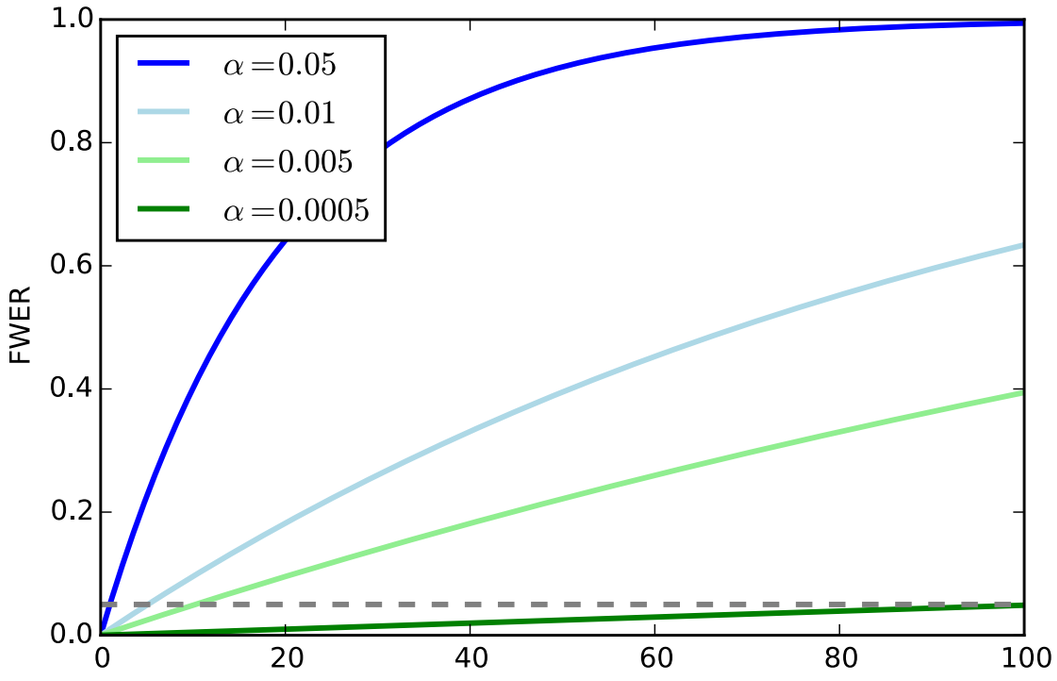
10回の検定



1-(1-0.05)^10

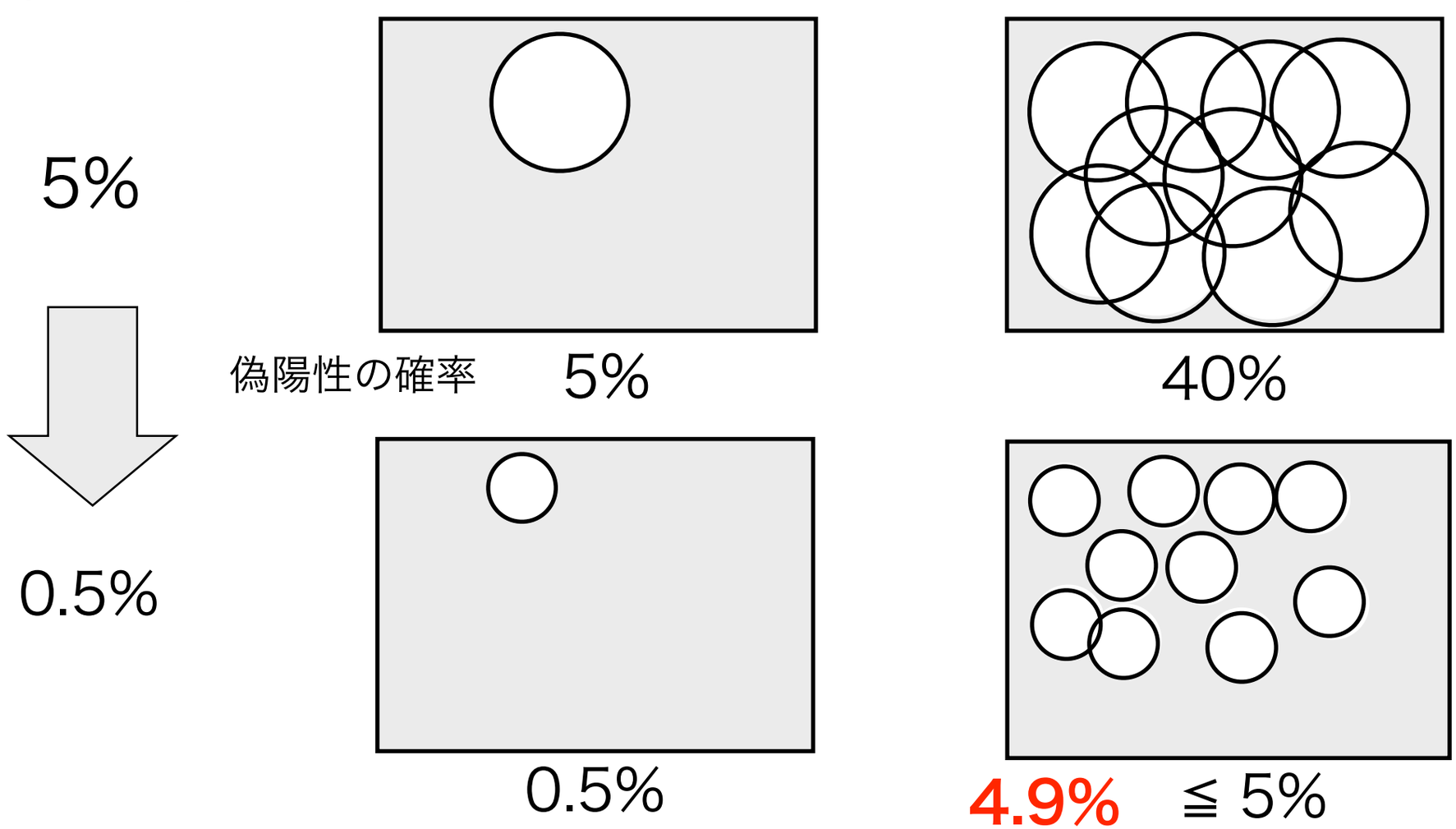
= 40%

- 1回の検定だと，「誤検出」は5%
- 10回の検定をすると，1回でも「誤検出」が現れる確率は，40%
- 有意→発見 となるので，同一の検定をランダムなデータで行った時に有意なものが生まれてしまう確率＝偽陽性を抑えたい.



多重検定補正：複数回の検定に起因する偽陽性を避ける

- 有意水準を，α=0.05ではなく，異なる数値にして，全体として偽陽性の起こる確率を抑える.
- 検定そのものは変更しない．有意水準を補正する.

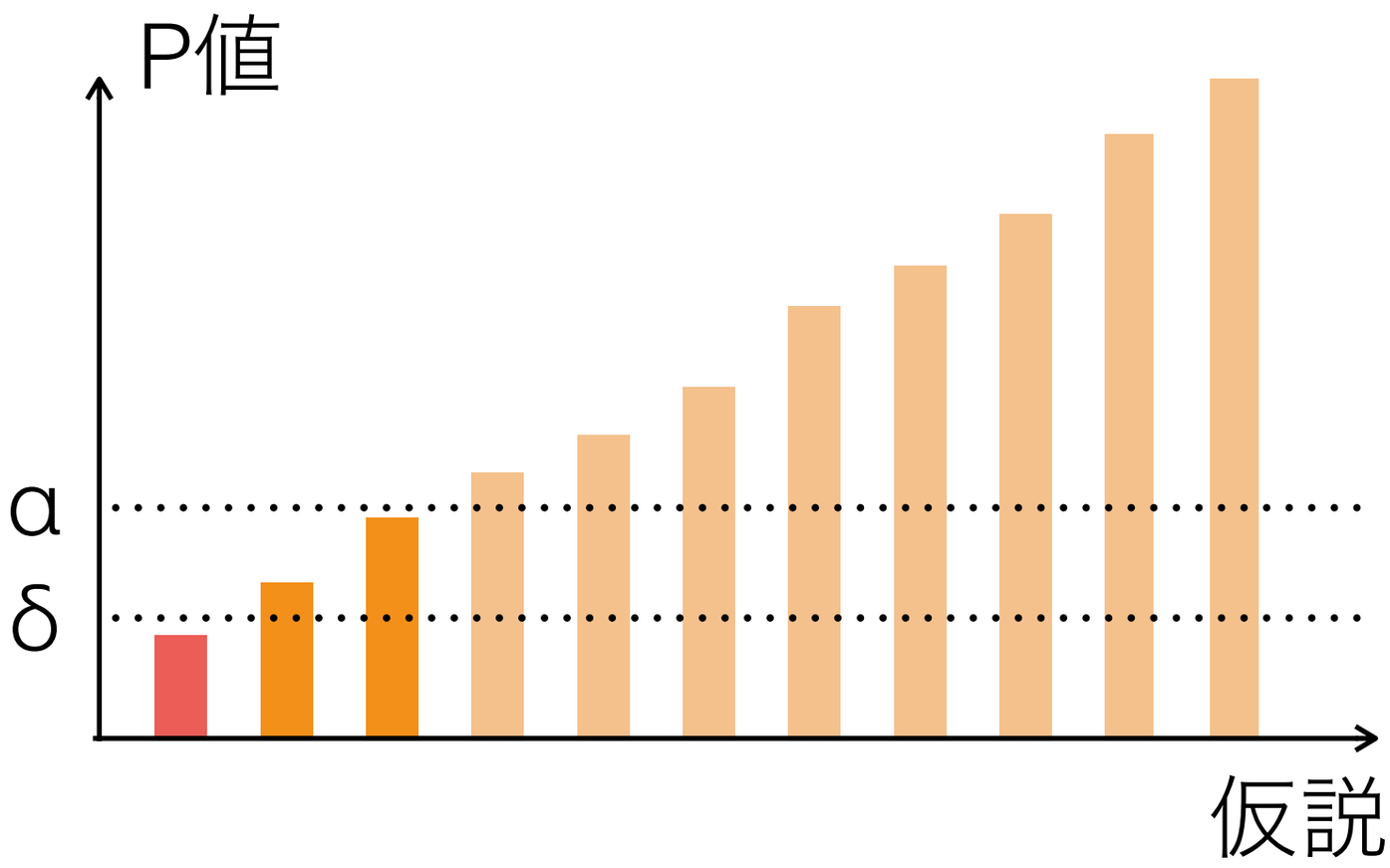


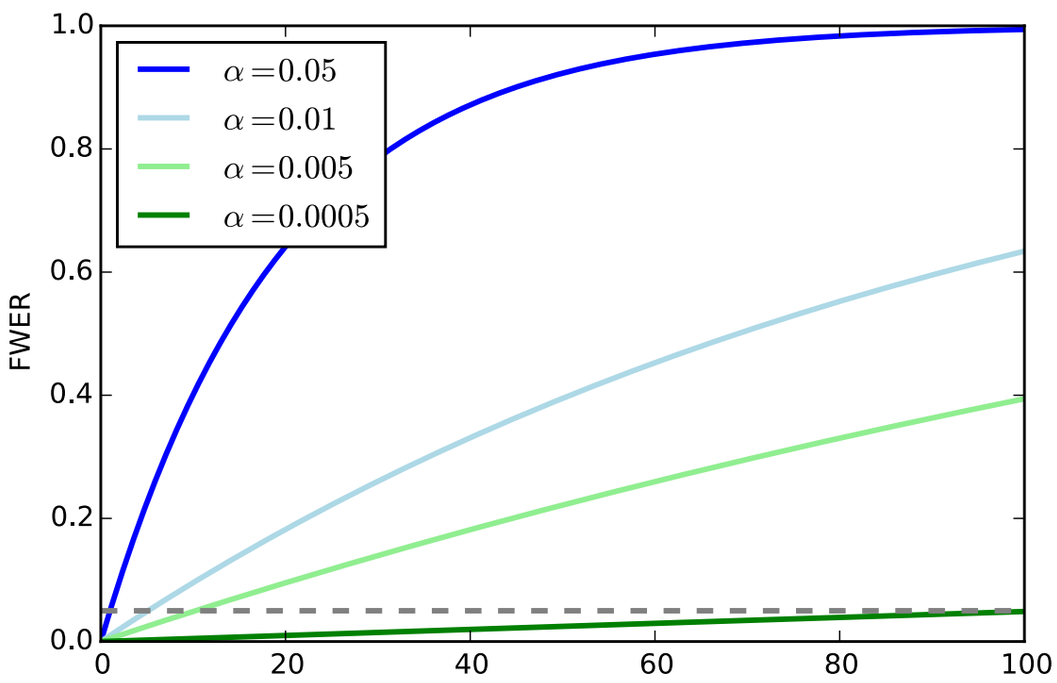
多重検定補正の手順

単一の検定：P値の計算 ⟶ α以下が有意

複数の検定：P値の計算 ⟶ 有意水準の補正 ⟶ δ以下が有意

αをδにする多重検定補正





前半に紹介する多重検定補正の例はRやSciPyで利用可能

- R
  - p.adjust() 関数で利用。

- Scipy
  - scikits.statsmodels の multicomp

- いずれもp値の一覧と、方法を入力すると、補正後の（有意水準と）p値が帰ってくる

補正の基準は，大きく分けて2通り

- 目的は**M回の検定をしても，（なんらかの）偽陽性がα以下**で抑えられること.
- **Family-Wise Error Rate (FWER)**
  - N回の検定をした場合に，1回も偽陽性が生まれない確率をαに抑える.
  - 今までの説明は，FWERを想定している
  - 有名な計算法にBonferroni法，Holm法，Westfall-Young法など
- **False Discovery Ratio (FDR)**
  - 検出された検定の内，偽陽性がα以下になるように抑える
  - 生命科学を中心に使われる
  - 有名な方法に Benjamini-Hochberg法，Efron法など.
- いずれも，有意水準の補正を行う方法（検定自身は変更しない）

様々な多重検定補正法

FWER		FDR	
Theoretical	Empirical	Theoretical	Empirical
Bonferroni (1959), **Holm (1979),** etc.	**Westfall-Young (1993)** HWY[Terada, Kim] (2015)	**Benjamini-Hochberg (1995), Storey-Tibshirani (2001),** Efron (2005), etc.	Romano *et al.* (2008)

目次

Bonferroni補正

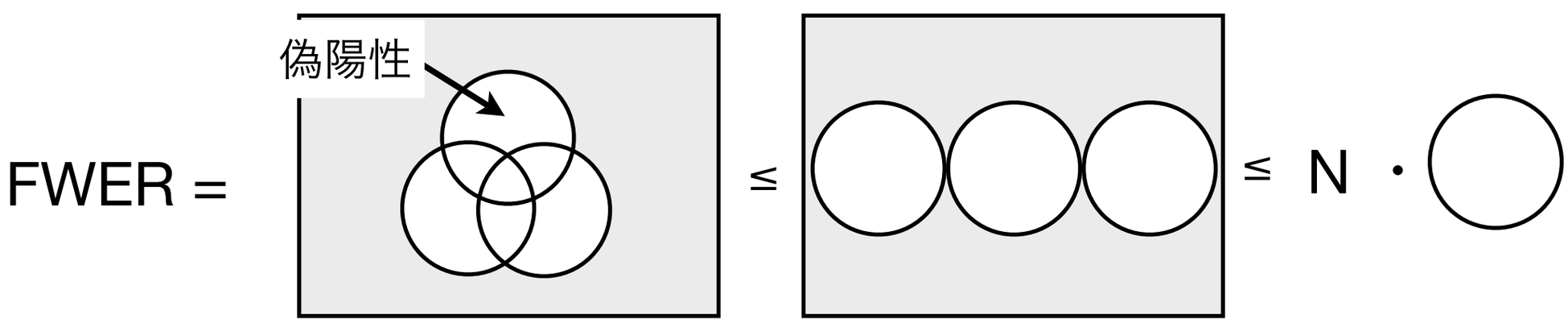
- FWERを α 以下に制御する方法
- 「補正のp値 = 検定数 * 元のp値」で知られる
- 理論的には，補正有意水準 δ を α /検定数 に設定

δ: 補正有意水準, M: 検定数, I_0: 帰無仮説に従う検定

$$
\begin{aligned}
\mathrm{FWER} &= \Pr(\text{帰無仮説に従う仮説 } H_i \text{が，1つ以上棄却される}) \\
&= \Pr\left(\bigcup_{i \in I_0} \{p_i \leq \delta\}\right) \leq \Pr\left(\bigcup_{i=1}^{M} \{p_i \leq \delta\}\right) \\
&\leq \sum_{i=1}^{M} \Pr\left(p_i \leq \delta\right)
\end{aligned}
$$

帰無仮説に従う H_i に対し，P 値の定義より $\Pr(p_i \leq \delta) = \delta$なので，

$$
= M\delta
$$



Bonferroni補正

- FWERをα以下に制御する方法
- 「補正のp値 = 検定数 * 元のp値」で知られる
- 理論的には，補正有意水準δをα/検定数 に設定

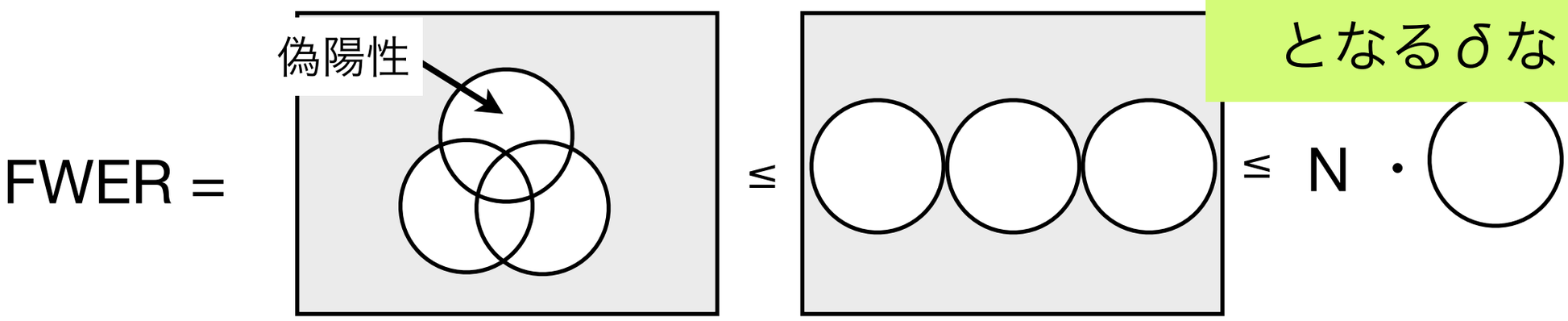
δ: 補正有意水準, M: 検定数, I_0: 帰無仮説に従う検定

$$
\begin{aligned}
\mathrm{FWER} &= \Pr(\text{帰無仮説に従う仮説 } H_i \text{が，1つ以上棄却される}) \\
&= \Pr\left(\bigcup_{i \in I_0} \{p_i \leq \delta\}\right) \leq \Pr\left(\bigcup_{i=1}^{M} \{p_i \leq \delta\}\right) \\
&\leq \sum_{i=1}^{M} \Pr\left(p_i \leq \delta\right) \\
&\quad \text{帰無仮説に従う } H_i \text{ に対し，P 値の定義より} \\
&= M\delta
\end{aligned}
$$

αとNは固定なので

$$\delta \leq \alpha / M$$

となるδなら良い.



Bonferroni補正を利用した検定手順

$p_{(1)}$	$p_{(2)}$	$p_{(3)}$	$p_{(4)}$	$p_{(5)}$	$p_{(6)}$	$p_{(7)}$	$p_{(8)}$	$p_{(9)}$	$p_{(10)}$
0.001	0.0026	0.0028	0.0029	0.006	0.007	0.02	0.03	0.04	0.05

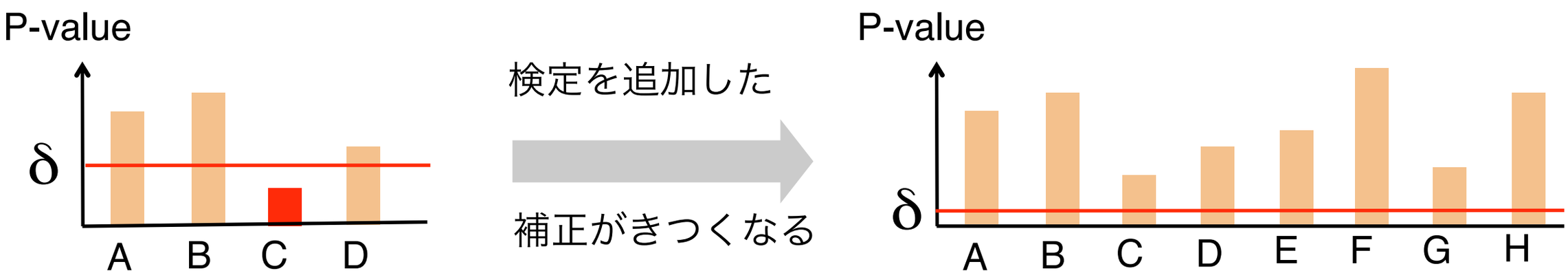
$p_{(11)}$	$p_{(12)}$	$p_{(13)}$	$p_{(14)}$	$p_{(15)}$	$p_{(16)}$	$p_{(17)}$	$p_{(18)}$	$p_{(19)}$	$p_{(20)}$
0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00

- 有意水準α=0.05での検定を考える
- 20個の検定があるので，補正後の有意水準δ=0.05/20 = 0.0025
- δ以下の$p_{(1)}$のみが，帰無仮説を棄却する（有意な差がある）
- P値を補正する計算の場合には，p(i) * 20 を行い，0.05以下の仮説を棄却する

ビッグデータのパラドックス

- 小さなデータで，有意な結果が見つかった
- データを足してみたら，有意な結果が消えてしまった.

- Bonferroni法は，「検定数」に依存して有意水準を変更する
ので，検定数が増えると有意な結果が生まれにくくなる



$$\delta = \alpha / N$$

Holm法

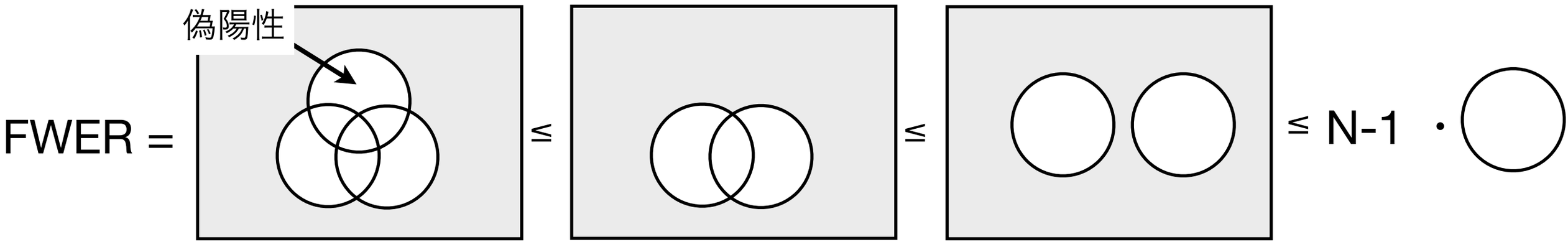
- Bonferroni補正を改良した方法.
- 棄却された仮説は，帰無仮説に従わないことを利用
- I_0を帰無仮説に従う仮説集合として，以下の様になる

$$\begin{aligned}\mathrm{FWER} &= \Pr(\text{帰無仮説に従う仮説 } H_i\text{が，1つ以上棄却される}) \\ &= \Pr\left(\bigcup_{i\in I_0}\{p_i \le \delta\}\right) \le \Pr\left(\bigcup_{i=1}^{M}\{p_i \le \delta\}\right)\end{aligned}$$

ここを考える

例えば，仮説H_1が棄却される場合，I_0にH_1は含まれないので

$$\le \Pr\left(\bigcup_{i=2}^{M}\{p_i \le \delta\}\right)$$ でも上限は抑えられる



Holm法を利用した検定手順

- Holm法では，(P値の昇順に並べた仮説に対し)i番目の仮説に対しては，α/(M-i+1)を補正後の有意水準として，検定を行う

$p_{(1)}$	$p_{(2)}$	$p_{(3)}$	$p_{(4)}$	$p_{(5)}$	$p_{(6)}$	$p_{(7)}$	$p_{(8)}$	$p_{(9)}$	$p_{(10)}$
0.001	0.0026	0.0028	0.0029	0.006	0.007	0.02	0.03	0.04	0.05

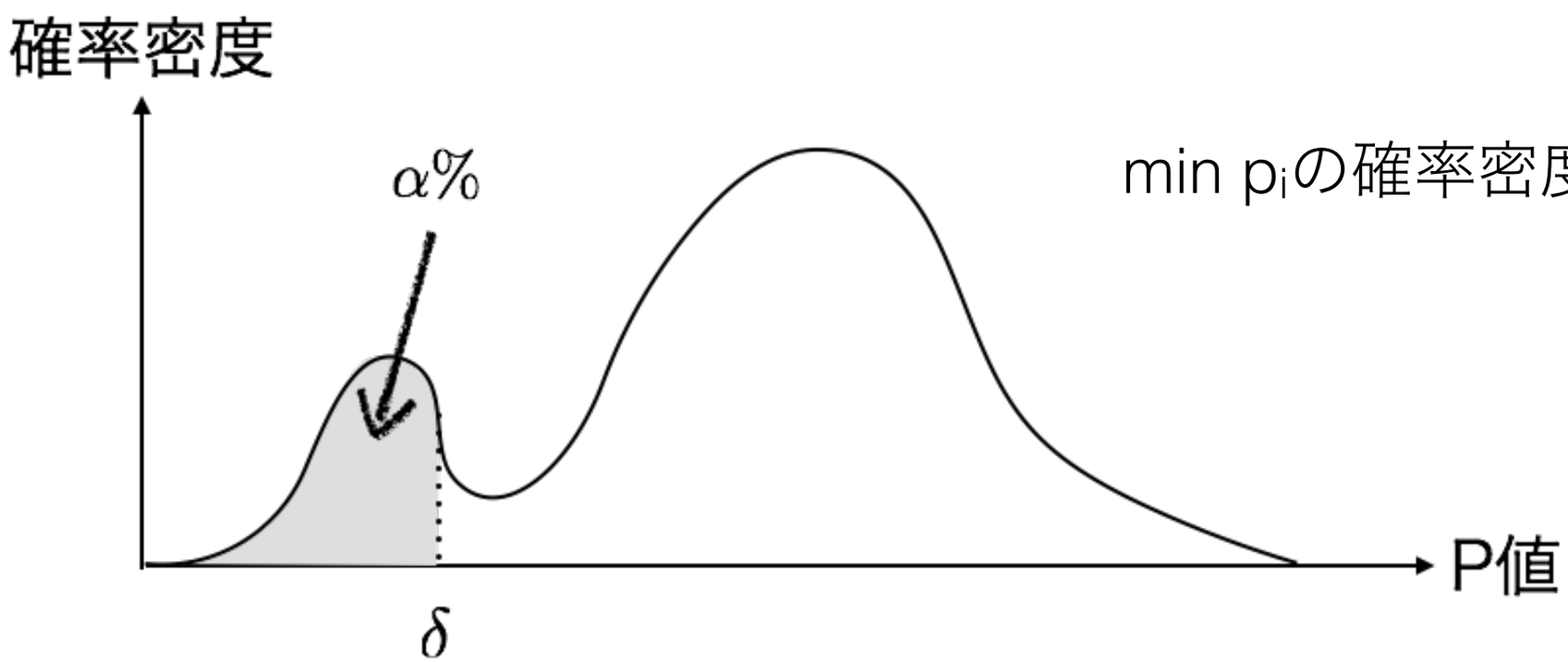
$p_{(11)}$	$p_{(12)}$	$p_{(13)}$	$p_{(14)}$	$p_{(15)}$	$p_{(16)}$	$p_{(17)}$	$p_{(18)}$	$p_{(19)}$	$p_{(20)}$
0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00

- 有意水準α=0.05での検定を考える
- p(1)に対する補正後の有意水準は，0.05/20 = 0.0025なので，H(1)は棄却される.
- p(2)に対しては，0.05/19 = 0.00263なので，やはり棄却される
- p(3)に対しては，0.05/18 = 0.00278なので，棄却されずに終了.

Westfall-Young法

- BonferroniやHolm法が理論的に上限を求めていたのに対し，モンテカルロ検定を用いる方法．仮説間の従属性を扱える．FWERを（BonferroniやHolmに比べ）αに近くできる．

$$\begin{aligned}\mathrm{FWER} &= \Pr(H_i(i \in I_0) \text{が，} \ 1\text{つ以上棄却される}) \\ &= \Pr(1\text{つ以上の} \ i \in I_0 \text{に対して} \ p_i \leq \delta) \\ &= \Pr(\min_{i \in I_0} p_i \leq \delta) \leq \Pr(\min_{i \in I} p_i \leq \delta)\end{aligned}$$



min p_iの確率密度分布の模式図

Westfall-Young法の手順

検定1 A = {1,2,5,7} / B = {2,3,8,9} → 並べ替え → A = {1,2,2,5} / B = {3,7,8,9} → P値 0.042

検定2 A = {5,8,9,9} / B = {3,3,4,5} → 並べ替え → A = {5,8,9,9} / B = {3,3,4,5} → 0.163

…

繰り返し

最小値

0.042

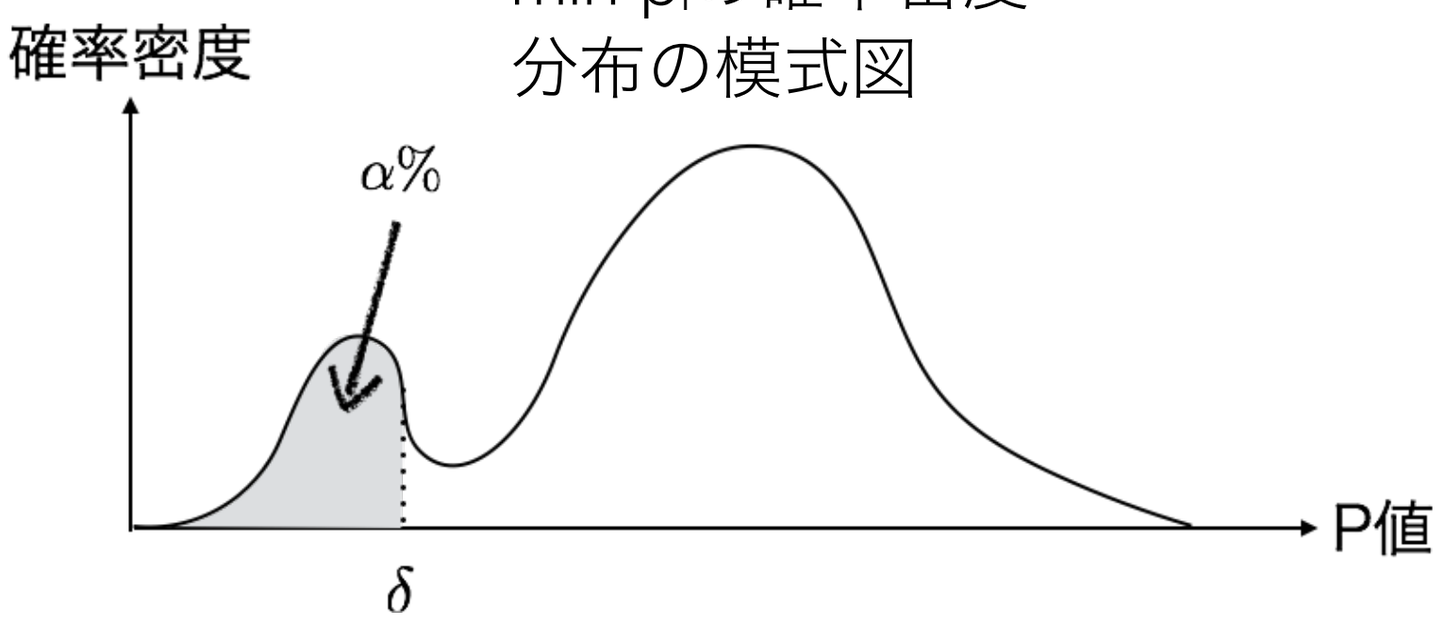
繰り返し2 0.008

繰り返し3 0.031

繰り返し4 0.028

…

min p_iの確率密度
分布の模式図



FWER制御法のまとめと課題

- Bonferroni
  - 手計算でできるシンプルさ
  - 分布に仮定をおかず，広く利用可能
- Holm
  - Bonferroni補正に準じ，広く利用可能
- Westfall-Young
  - 計算機を積極的に用いる手法
  - 検定間の従属性を扱える
  - GPUを利用した高速化 [Terada, Kim, Sese, ’15]
- FWERは，「ひとつの偽陽性も起こさない確率」を考えている．この条件がキツすぎる場合もあるので，「k個の偽陽性まで許す」k-FWERを考える場合もある．また，基準としてFDRを用いる場合もある（次頁以降）

目次

FDRによる制御

- 棄却された仮説I_Rの中に含まれる「帰無仮説に従っているにも関わらず，棄却した」仮説（$I_R \cap I_0$）の**割合**を制御する
  - 幾つか誤って棄却する場合があるだろうと考える
- もともと「複数の仮説が棄却されるだろう」という仮説集合に対して有用
  - 全3万遺伝子中，薬で発現量の変わる遺伝子を選びたい.
    （1割程度は変わるだろう，という予想のもとの検定）
- 補正後の値は，q値と呼ばれる.

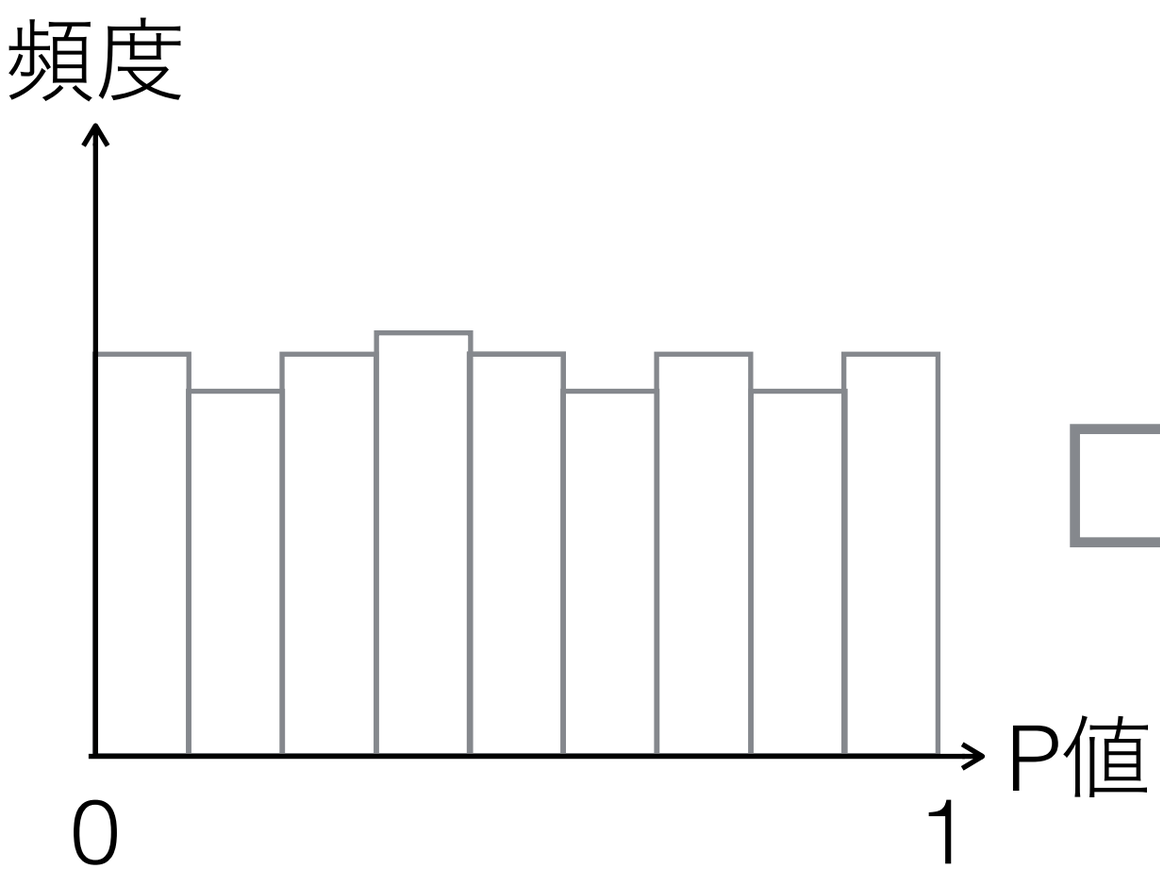
$$\mathrm{FDR} = E\left(\frac{|I_R \cap I_0|}{|I_R|}\right)$$

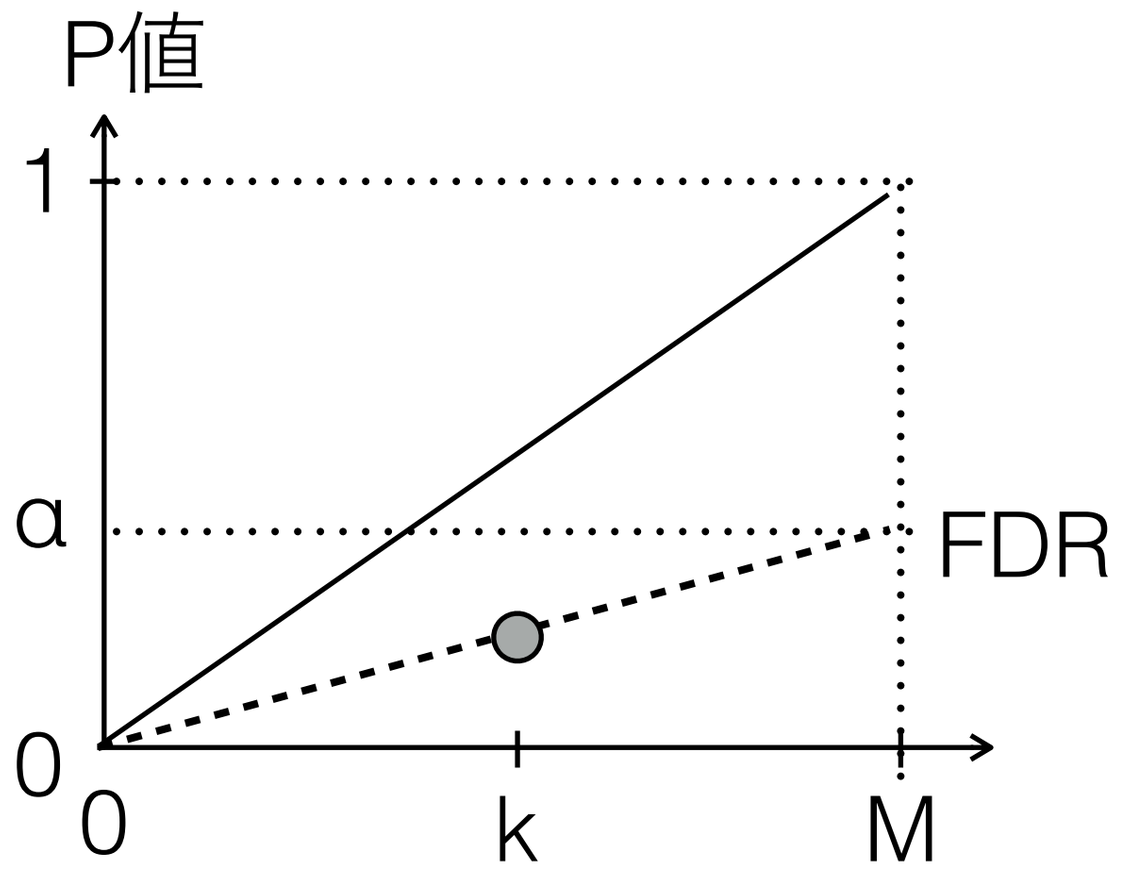
I_R:$P(p_i \leqq \delta)$となる仮説集合

I_0:帰無仮説に従う仮説集合

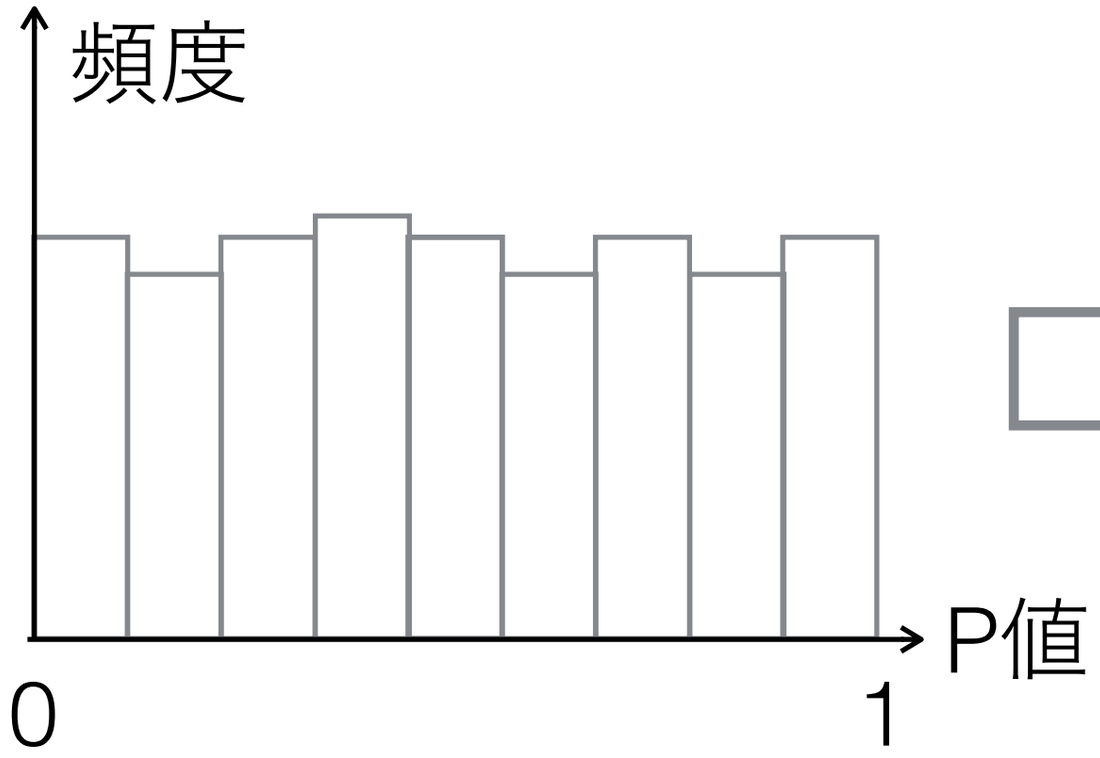
Benjamini-Hochberg法 (BH法)

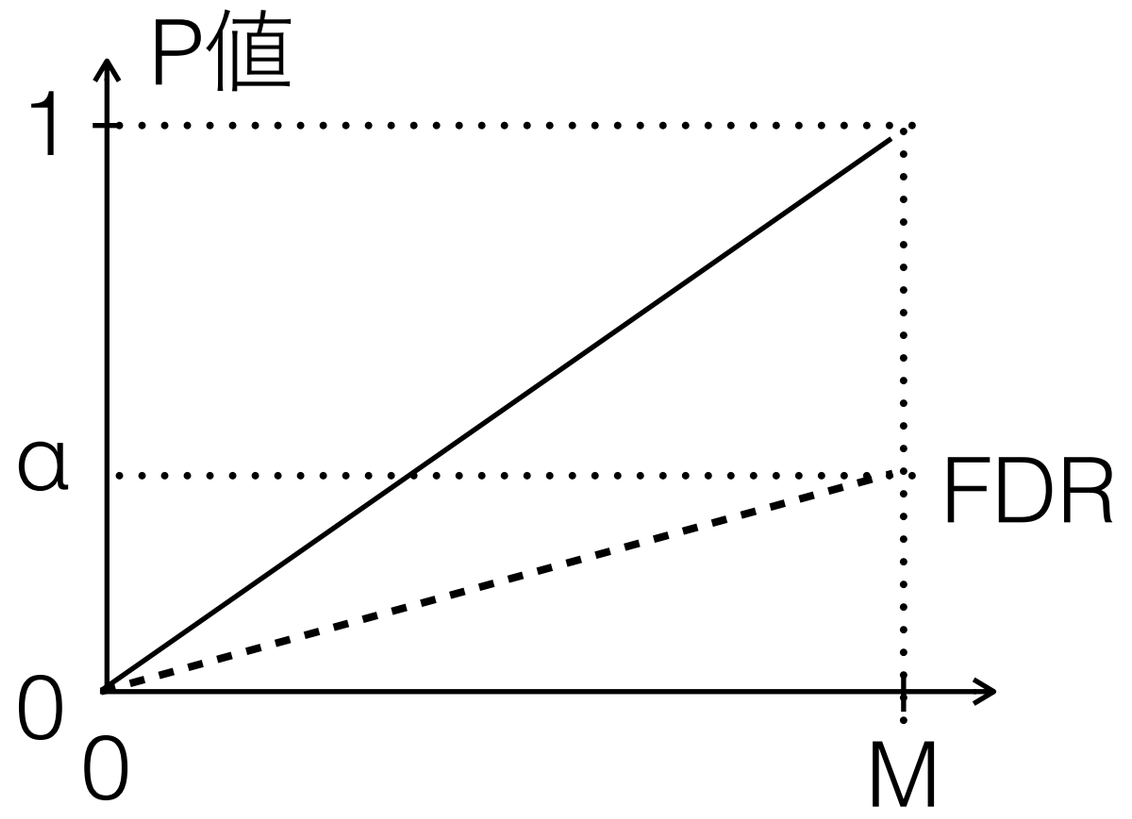
- 仮説のP値は0から1の間で一様に分布しているとする
  - P値の定義から，ほとんどの仮説が，帰無仮説に従っている場合は，正しい.
- P値を小さい順に仮説を並べ，下からk番目の仮説におけるFDRの期待値は α・k/M



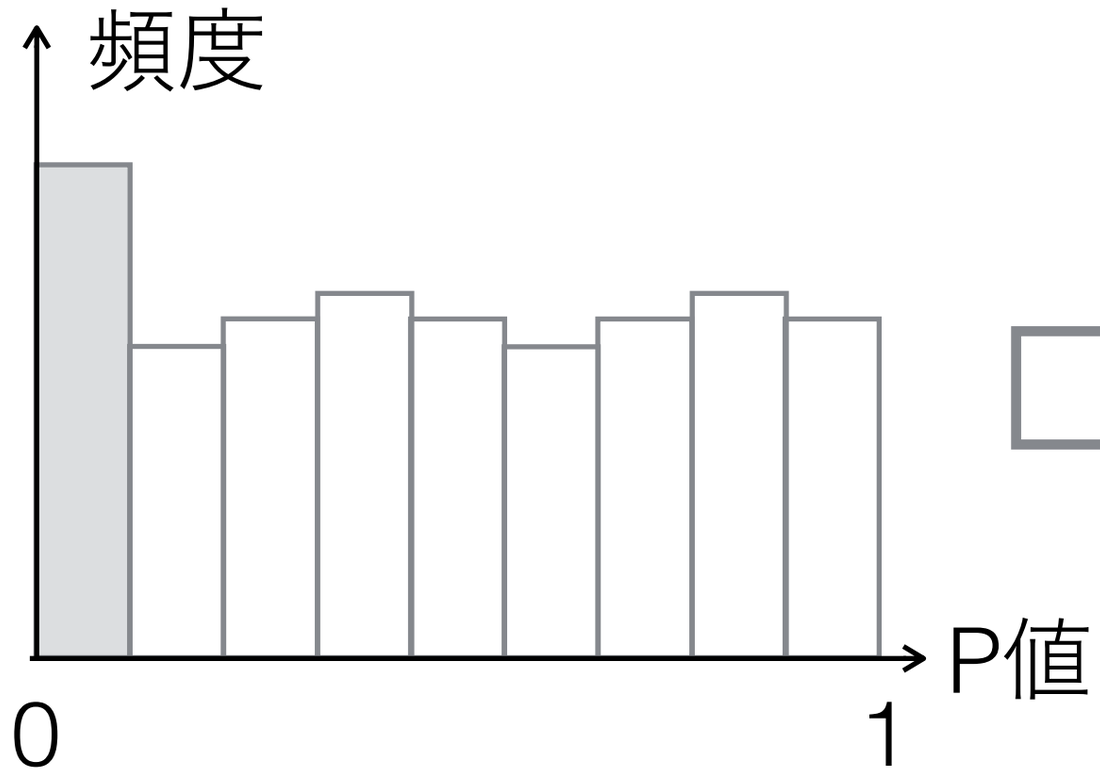


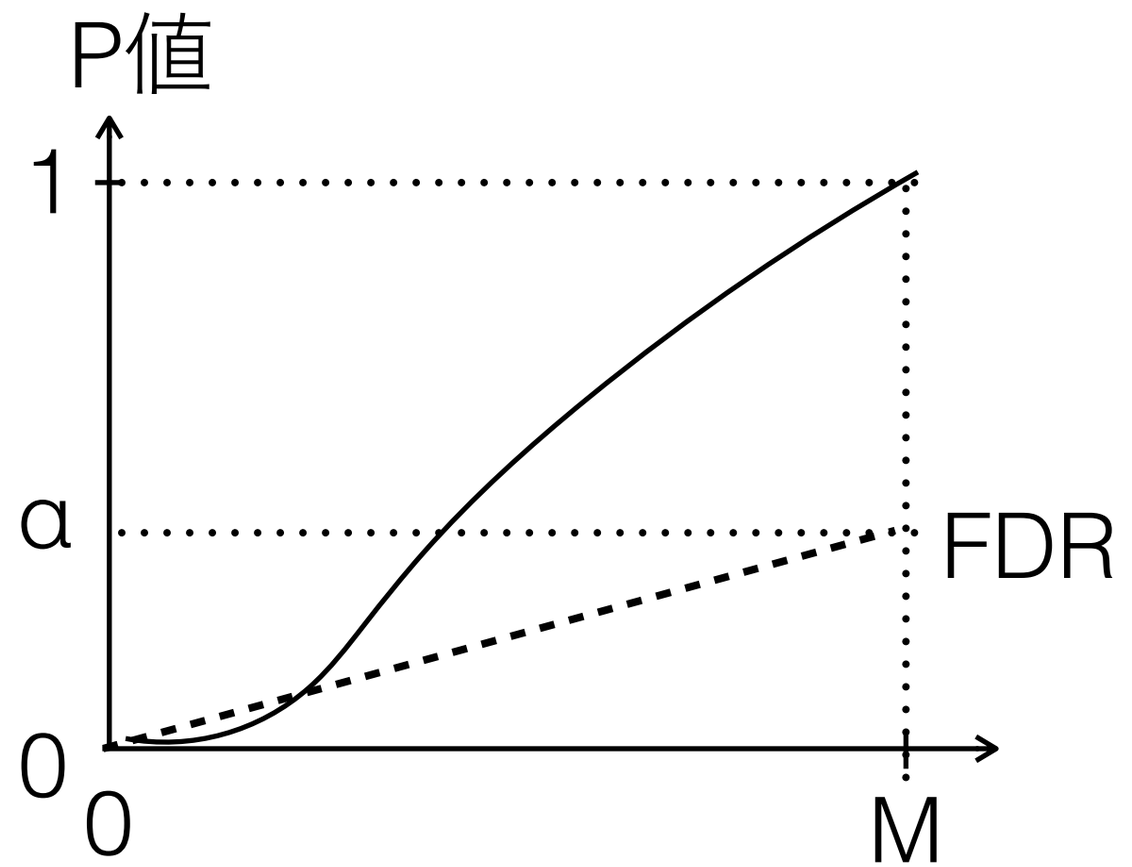
有意な差のある結果が無い時





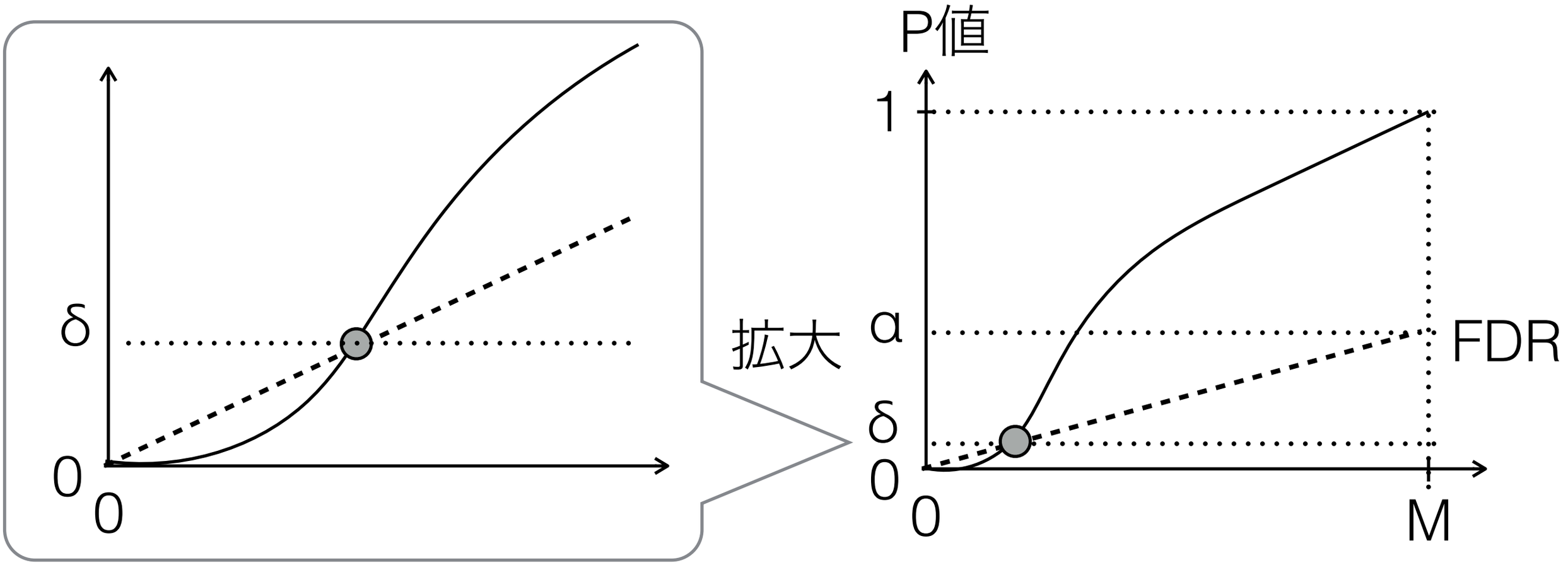
有意な差のある結果がある時





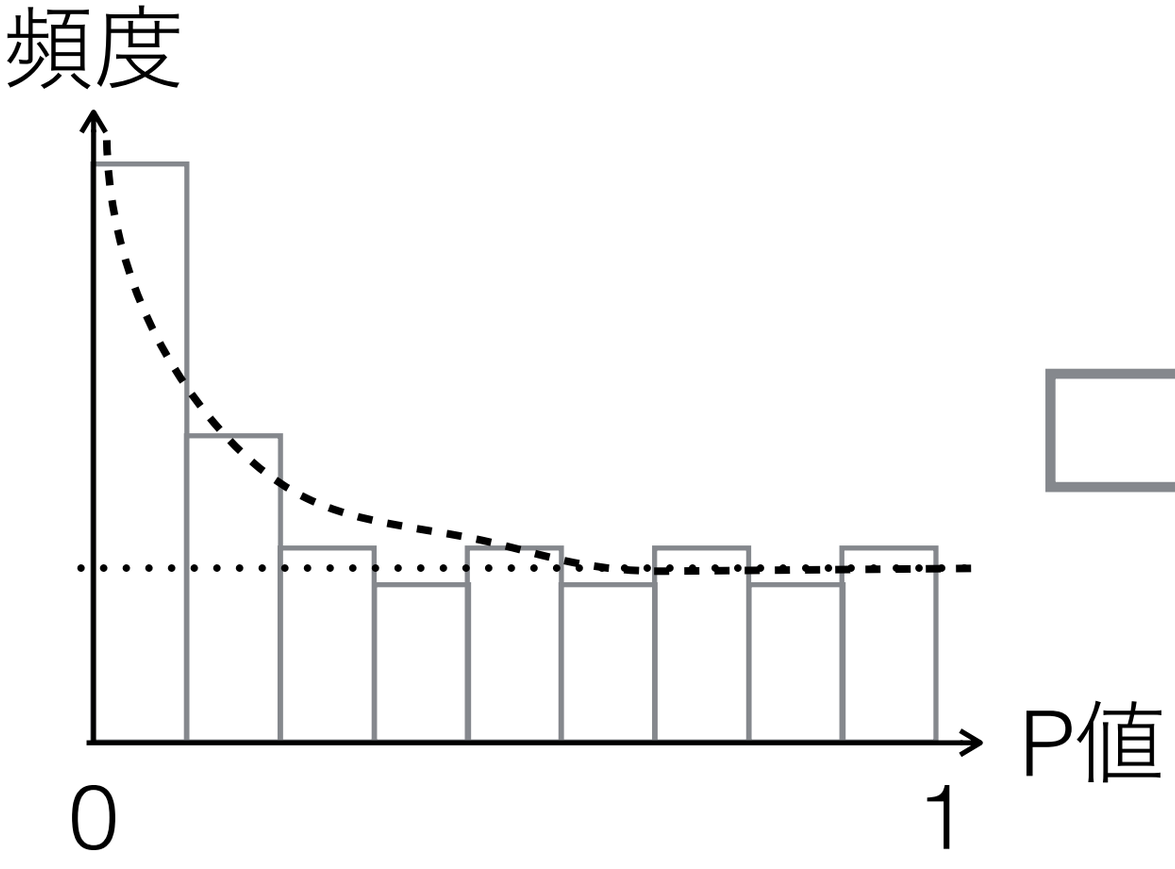
Benjamini-Hochberg法 (BH法)

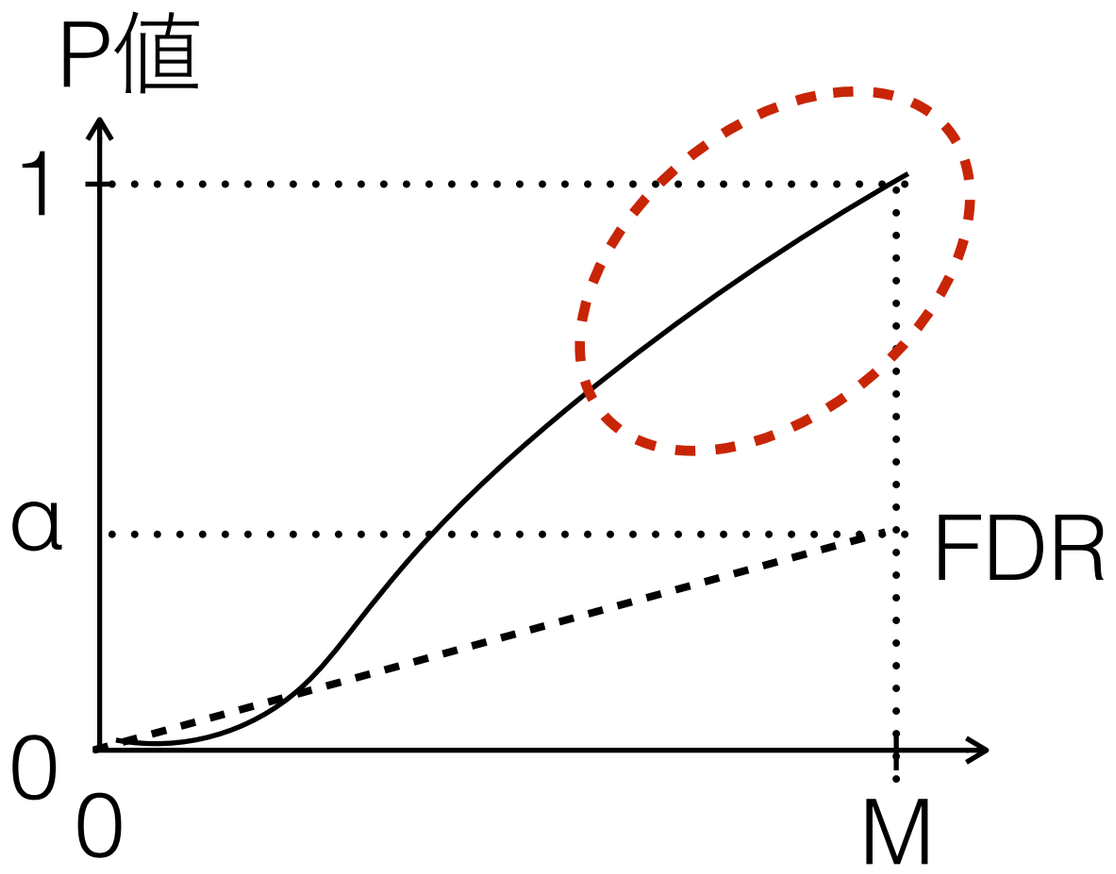
- P値が α・k/Mを下回るようにδを制御する.
- BH法では, Pkがα・k/Mを下回る, 最も大きなδを選ぶ



Storey and Tibshirani (ST) 法

- 仮説のP値の分布が一様でない事も多い
- 特に，ある程度の遺伝子や化合物に反応が認められる時
  - 投薬後は，数百の遺伝子発現が変わるとかんがえられる.
- P値が1に近い仮説に限れば，帰無仮説に従う＝一様分布をするだろう.
- 近似曲線をつかって，推定





目次

- なぜIBISで検定なのか
- 多重検定補正
- FWER
  - Bonferroni, Holm, Westfall-Young
- FDR
  - Benjamini-Hochberg, Storey-Tibshirani
- Bonferroni法の更なる発展
  - Tarone法
  - 組合せを考慮したアルゴリズム：LAMP
  - 酵母，ヒトのデータへの適用
- まとめ

Tarone法 [1990]

- 偽陽性を生まない検定はBonferroniの補正項から除外可能
  - 検定i が　$\Pr(p_i \leq \delta) = 0$ となる事が分かるなら，iを除ける
  - それ以外は，一律に　$\Pr(p_i \leq \delta) = \delta$ とする (=Bonferroni).

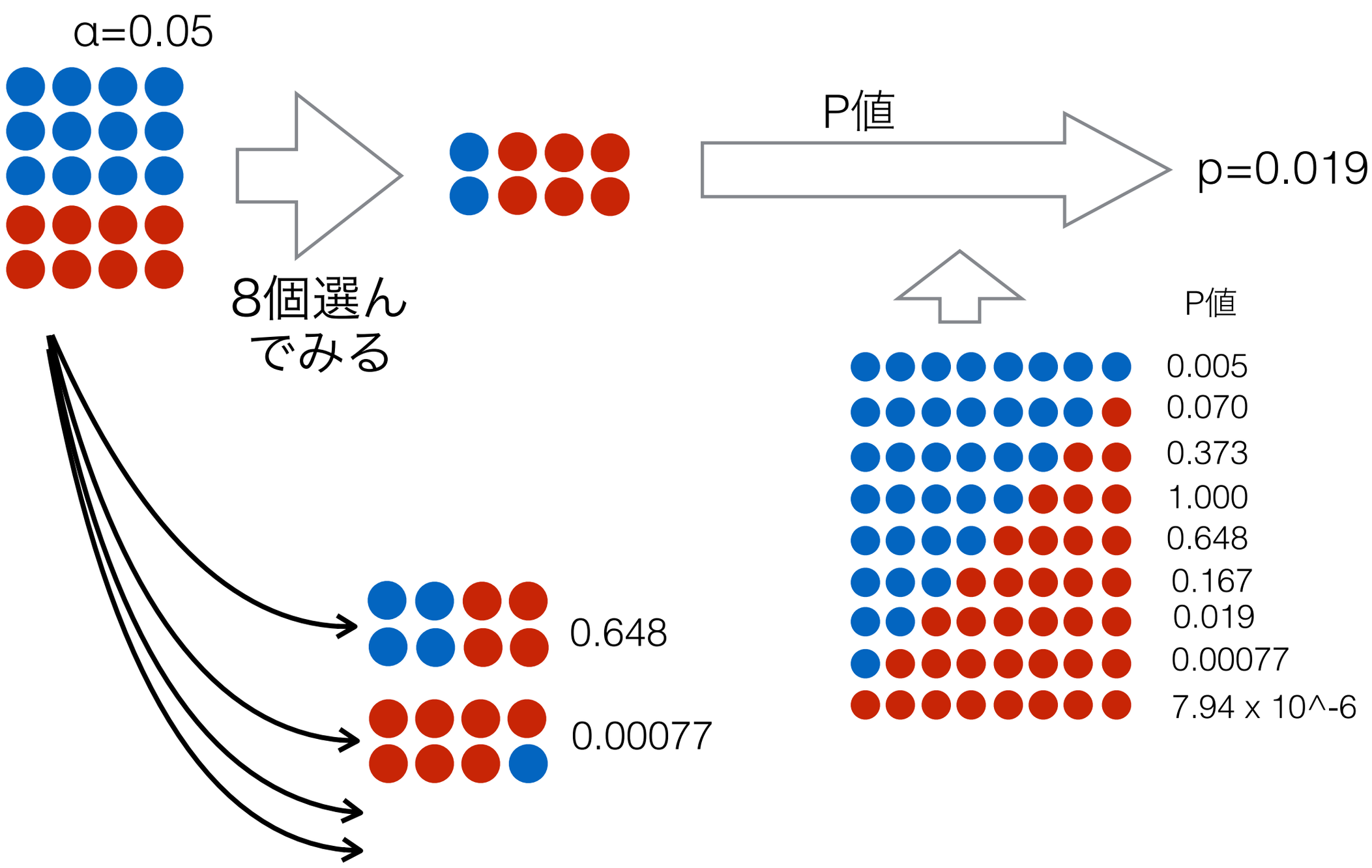
$$\text{FWER} = \Pr\left(\bigcup_{i\in\{1,\ldots,N\}} \{p_i \leq \delta\}\right) \leq \sum_{i\in\{1,\ldots,N\}} \Pr(p_i \leq \delta)$$

$\Pr(p_i \leq \delta) = 0$ となる条件 $\Leftrightarrow$ p_iが常に $> \delta$

$\Leftrightarrow$ p_i の最小値がδより大きい

p_i の最小値を $f(x_i)$ とすると

$$= \sum_{\{i|f(x_i)\leq\delta\}} \Pr(p_i \leq \delta) \leq |\{i|f(x_i) \leq \delta\}| \cdot \delta$$

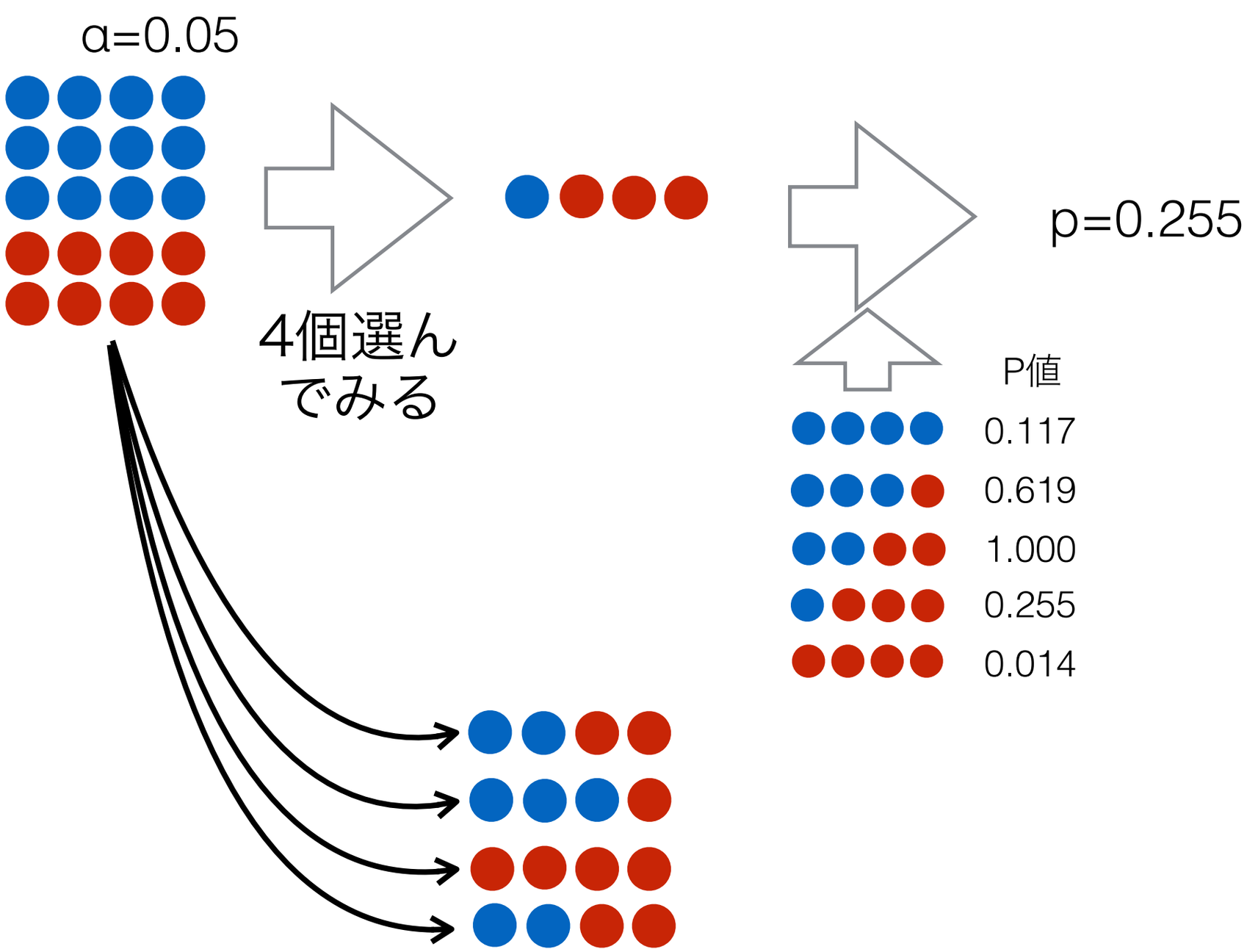


Bonferroni法：

偶然の有意差出現を防ぐために α / N (検定数) を有意の基準とする

5回検定すると 0.05 / 5 = 0.01 が基準

0.01を切るような場合が，存在する→検定に値する



Bonferroni法：

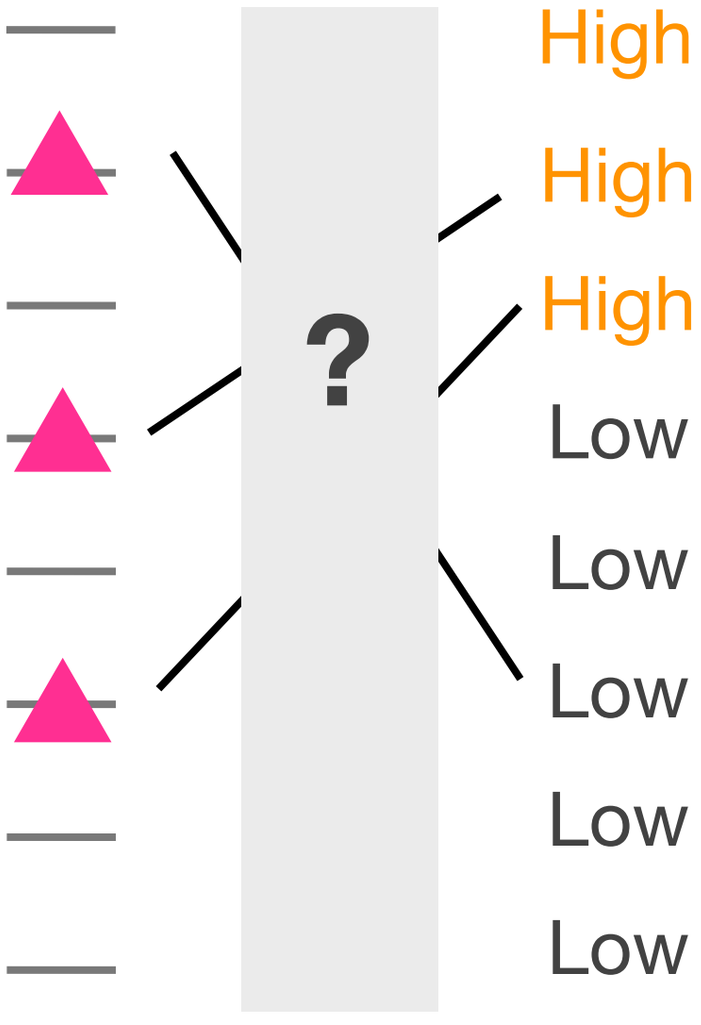
　偶然の有意差出現を防ぐためにαの代わりに α / N (検定数) を基準とする

5回検定すると 0.05 / 5 = 0.01 が基準

上の例だと，0.01以下になるような場合が無い

検定をしても，有意なものは現れない→検定として有効ではない

	▲	✕	Total
High	?	?	3
Low	?	?	5
Total	3	5	8



	▲	✕	Total
High	3	0	3
Low	0	5	5
Total	3	5	8

p=0.017

	▲	✕	Total
High	2	1	3
Low	1	4	5
Total	3	5	8

p=0.244

	▲	✕	Total
High	1	2	3
Low	2	3	5
Total	3	5	8

p=0.762

	▲	✕	Total
High	0	3	3
Low	3	2	5
Total	3	5	8

p=1

モチーフと発現の関係が分からない場合（＝周辺分布のみ分かっている）に,
どのようなp値を取りうるか考える.

最小値は 0.017

	▲	✕	Total
High	?	?	3
Low	?	?	5
Total	3	5	8

	▲	✕	Total
High	3	0	3
Low	0	5	5
Total	3	5	8

p=0.017

	▲	✕	Total
High	2	1	3
Low	1	4	5
Total	3	5	8

p=0.244

	▲	✕	Total
High	1	2	3
Low	2	3	5
Total	3	5	8

p=0.762

	▲	✕	Total
High	0	3	3
Low	3	2	5
Total	3	5	8

p=1

- もし，補正有意水準 $\delta < 0.017$，$\Pr(p_{▲} \leq \delta) = 0$

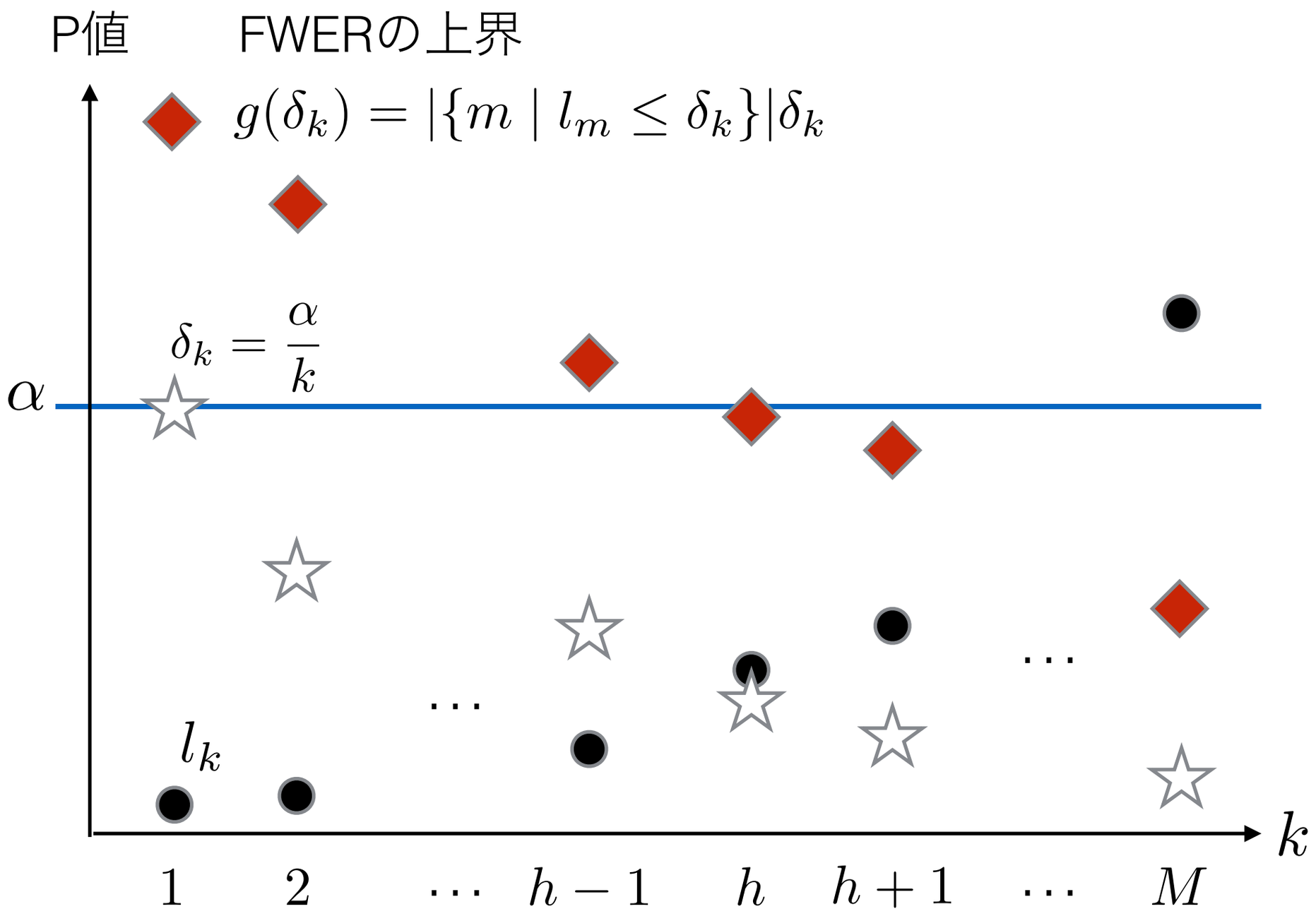
- 一般に，Fisher’s exact testの場合，piの最小値は

	▲	✕	Total
High	?	?	n_u
Low	?	?	N-n_u
Total	x	N-x	N

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

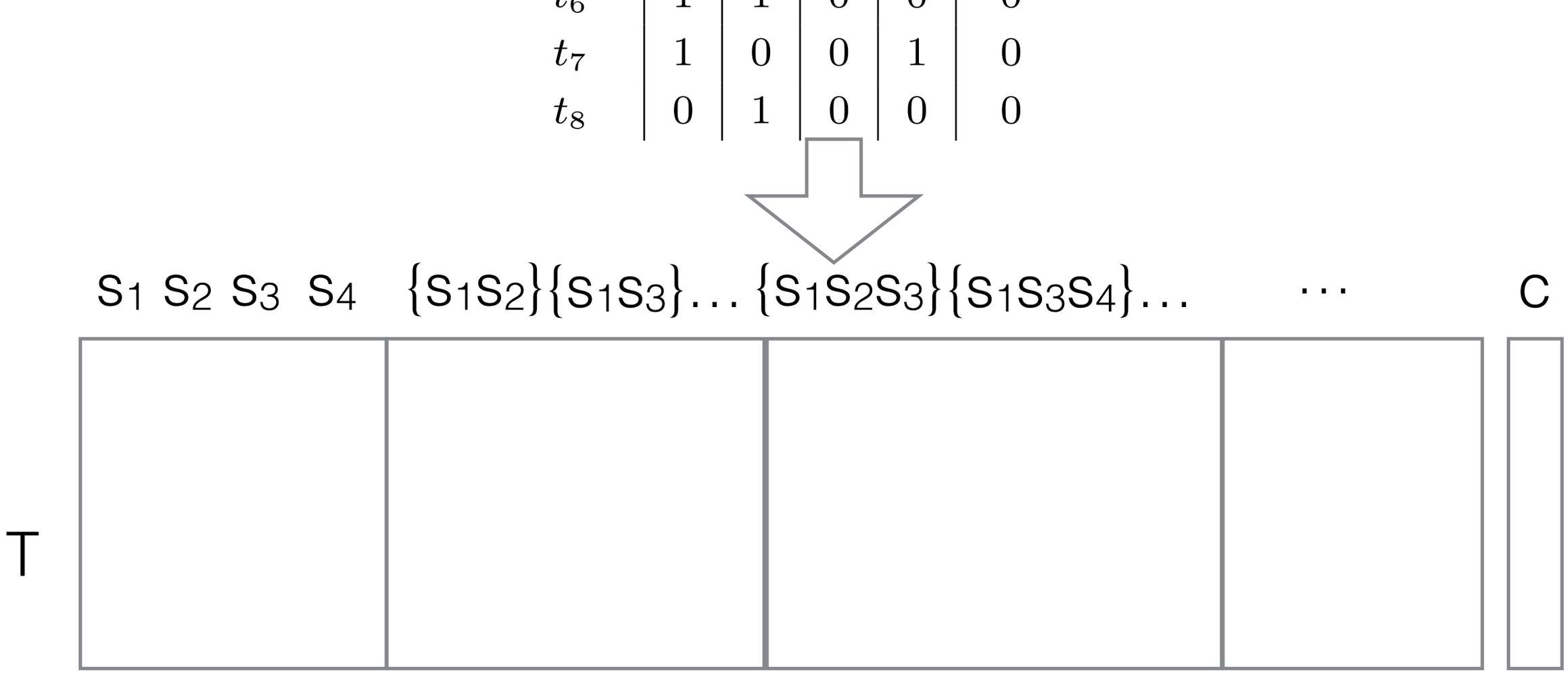
偽陽性が生まれないのは

$f(x) > \delta$ の時.


P値
FWERの上界
$g(\delta_k) = |\{m \mid l_m \leq \delta_k\}|\delta_k$
$\delta_k = \frac{\alpha}{k}$
α
l_k
k
1
2
$\cdots$
$h-1$
h
$h+1$
$\cdots$
M

組み合わせ問題への展開

	変異				疾患
被験者	s_1	s_2	s_3	s_4	c
t_1	0	1	1	0	1
t_2	1	1	1	0	1
t_3	0	1	1	1	1
t_4	1	0	0	0	0
t_5	0	1	1	1	0
t_6	1	1	0	0	0
t_7	1	0	0	1	0
t_8	0	1	0	0	0

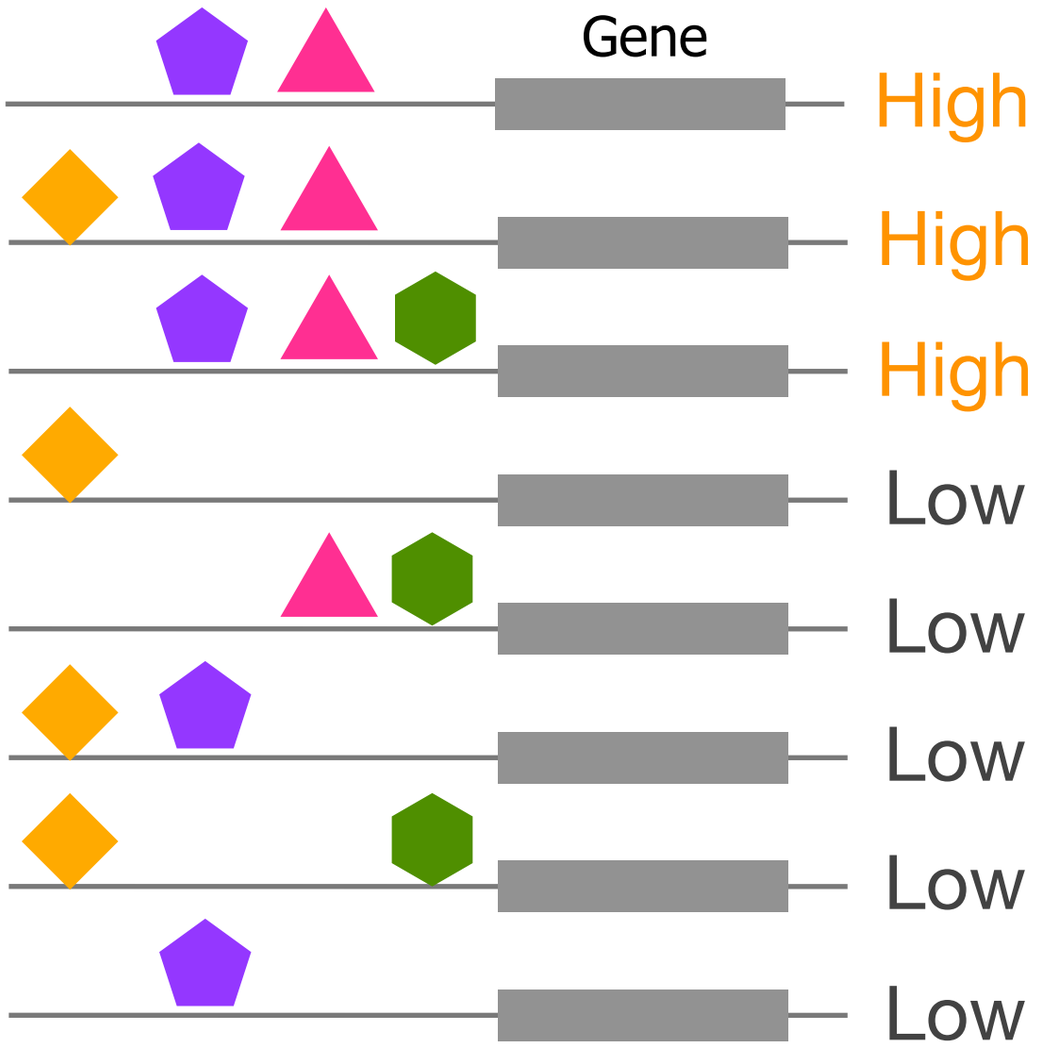


多重検定補正法の概観

	FWER		FDR	
	Theoretical	Empirical	Theoretical	Empirical
Single element	Bonferroni (1959), Holm (1979), etc.	Westfall-Young (1993) **HWY[Terada, Kim] (2015)**	Benjamini-Hochberg (1995), Efron (2005), etc.	Romano *et al.* (2008)
Combinatorial element	**LAMP [Terada et al.] (PNAS. 2013)**	FastANOVA (only pairs) (2008) **FastWY [Terada et al.] (BIBM 2013)**		

働いているモチーフの発見

- 遺伝子発現を制御しているモチーフを発見したい
  - 簡単のため，発現は高低の２種類とする



分割表

			Total
High	3	0	3
Low	0	5	5
Total	3	5	8

Fisher’s exact test
p=0.018 < 0.05→有意？

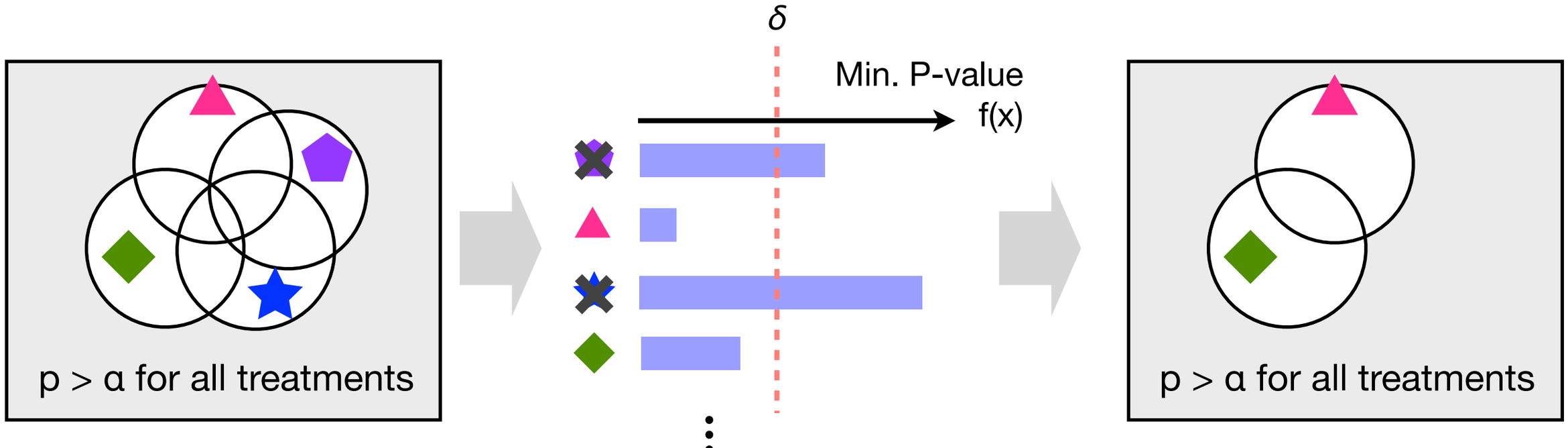
No！
多重検定補正が必要.
2個の組合せを考えるなら
4*3/2=6回の検定が必要

Taroneの補正

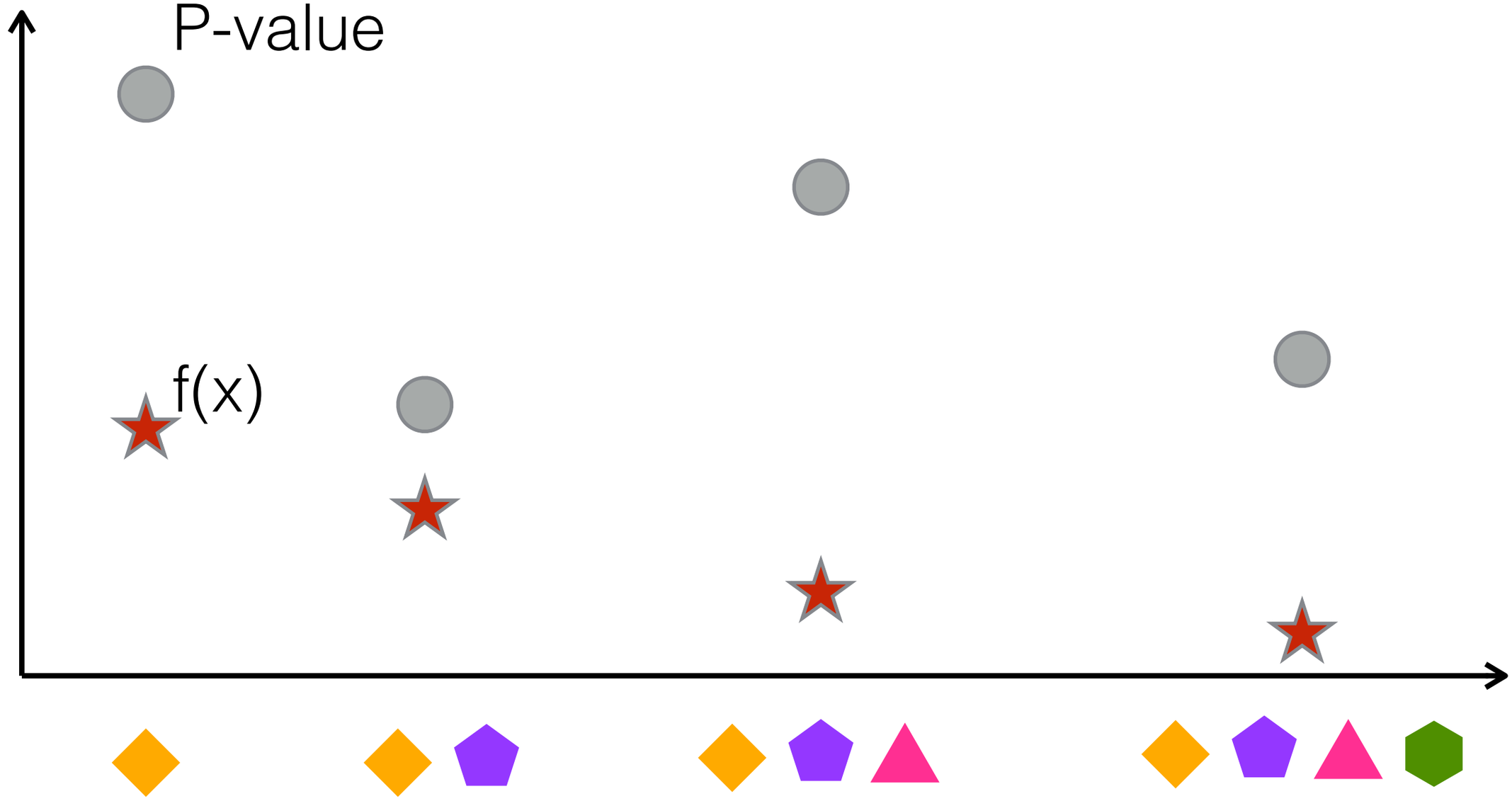
$$\alpha' = \Pr\left(\bigcup_{i\in\{1,\ldots,N\}} \{p_i \le \delta\}\right) \le \sum_{i\in\{1,\ldots,N\}} \Pr(p_i \le \delta)$$

$$= \sum_{\{i|f(x_i)\le\delta\}} \Pr(p_i \le \delta) \le |\{i|f(x_i) \le \delta\}| \cdot \delta$$

- FWERの上界が α以上になるような最大のδを取る
  - δが大きい＝鋭敏な補正＝発見が多くなる
  - Taroneは，全検定をf(x)でソートした後，探索を行なっている
  - 組合せを考えた場合には，計算時間的に困難

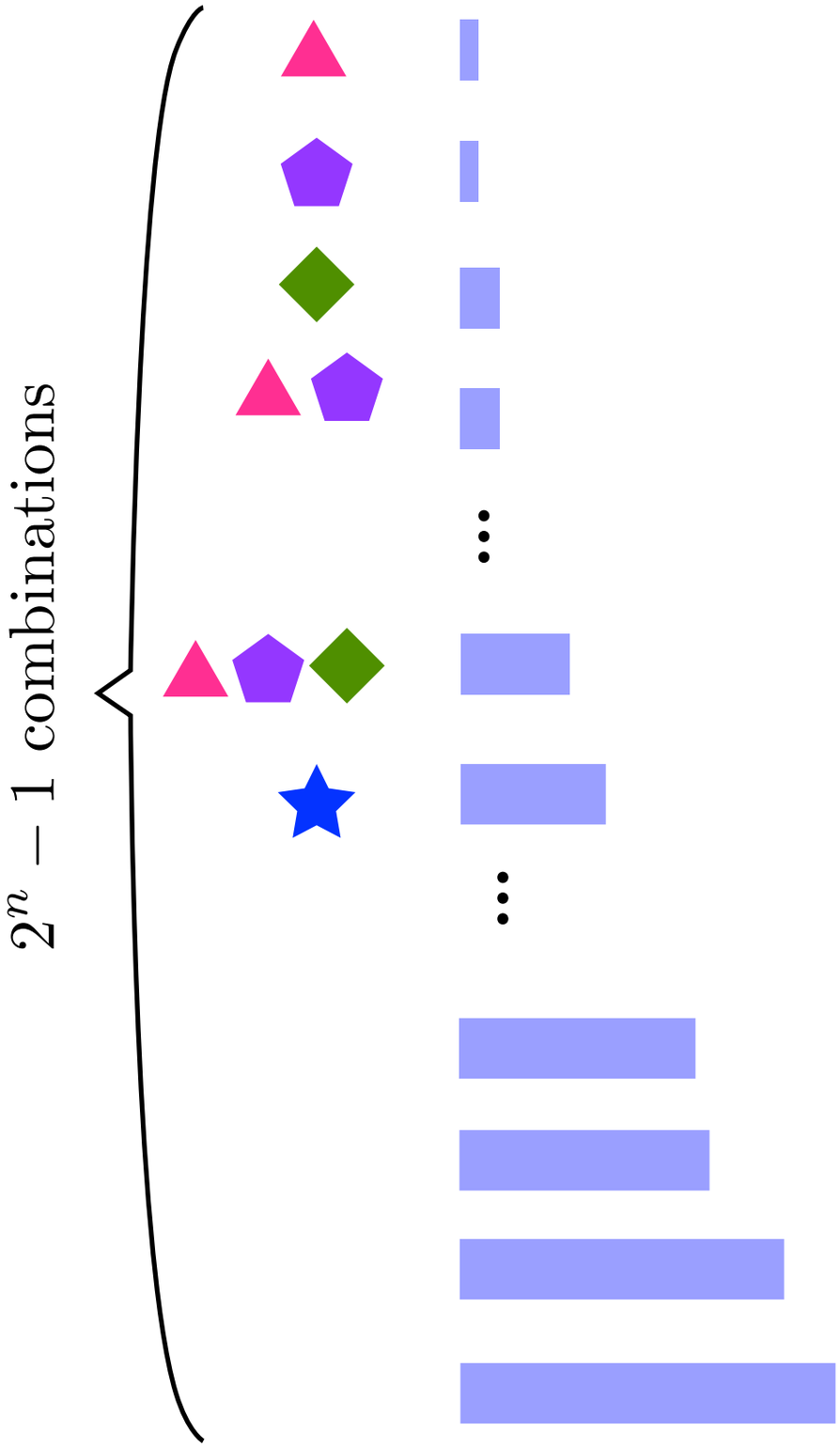


$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$



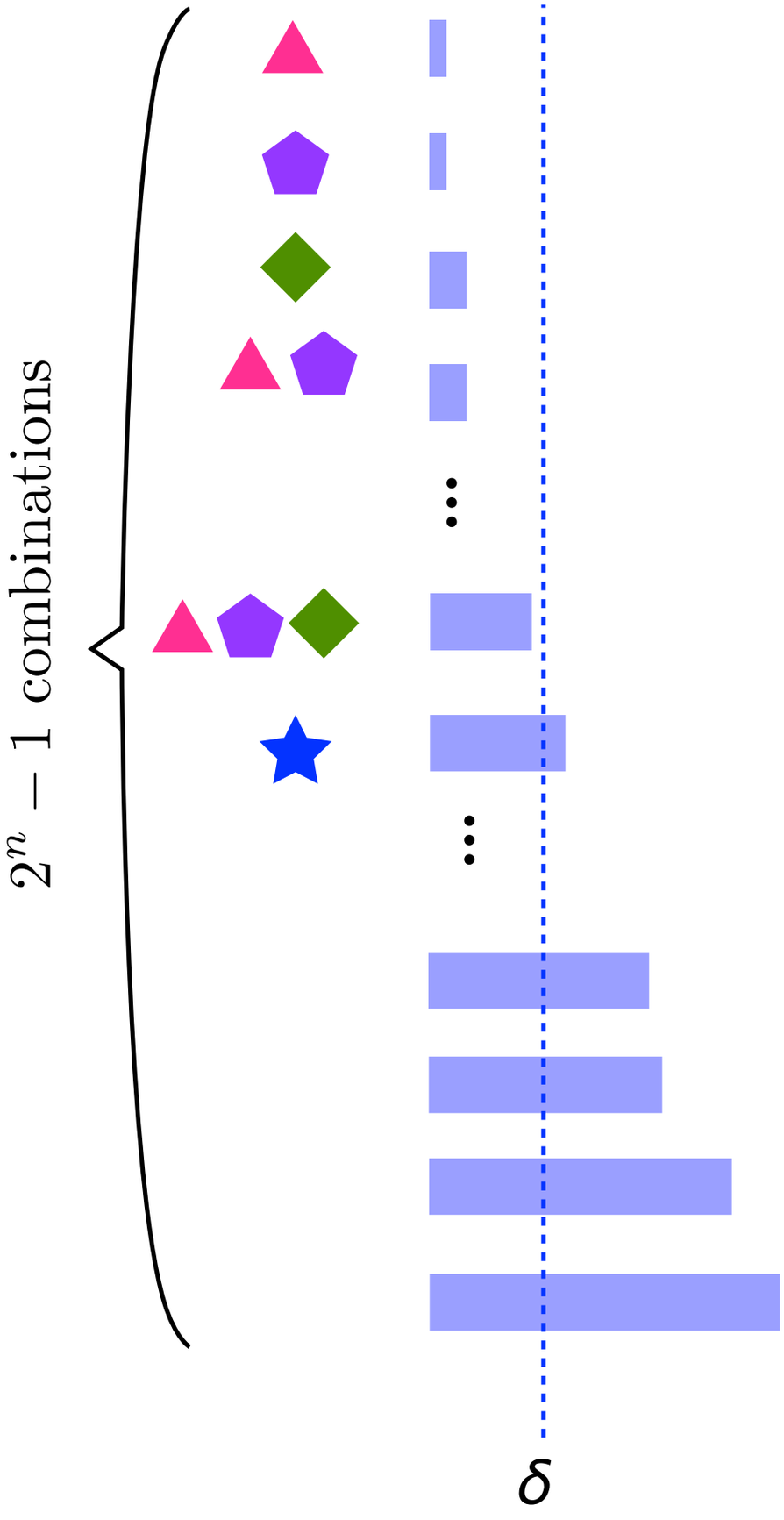
Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$



Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$



Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

$2^n - 1$ combinations

$$k = |\{i | f(x_i) \leq \delta\}|$$

δ

Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

$2^n - 1$ combinations

$$k = |\{i | f(x_i) \leq \delta\}|$$

$$\alpha \leq \sum_{\{i|f(x_i)\leq\delta\}} \Pr(p_i \leq \delta) \leq |\{i|f(x_i) \leq \delta\}| \cdot \delta$$

FWER の上界

$$g(x) = k\delta$$

FWER の上界 $g(x)$が, α以上である範囲内で, 最小のδを求める.

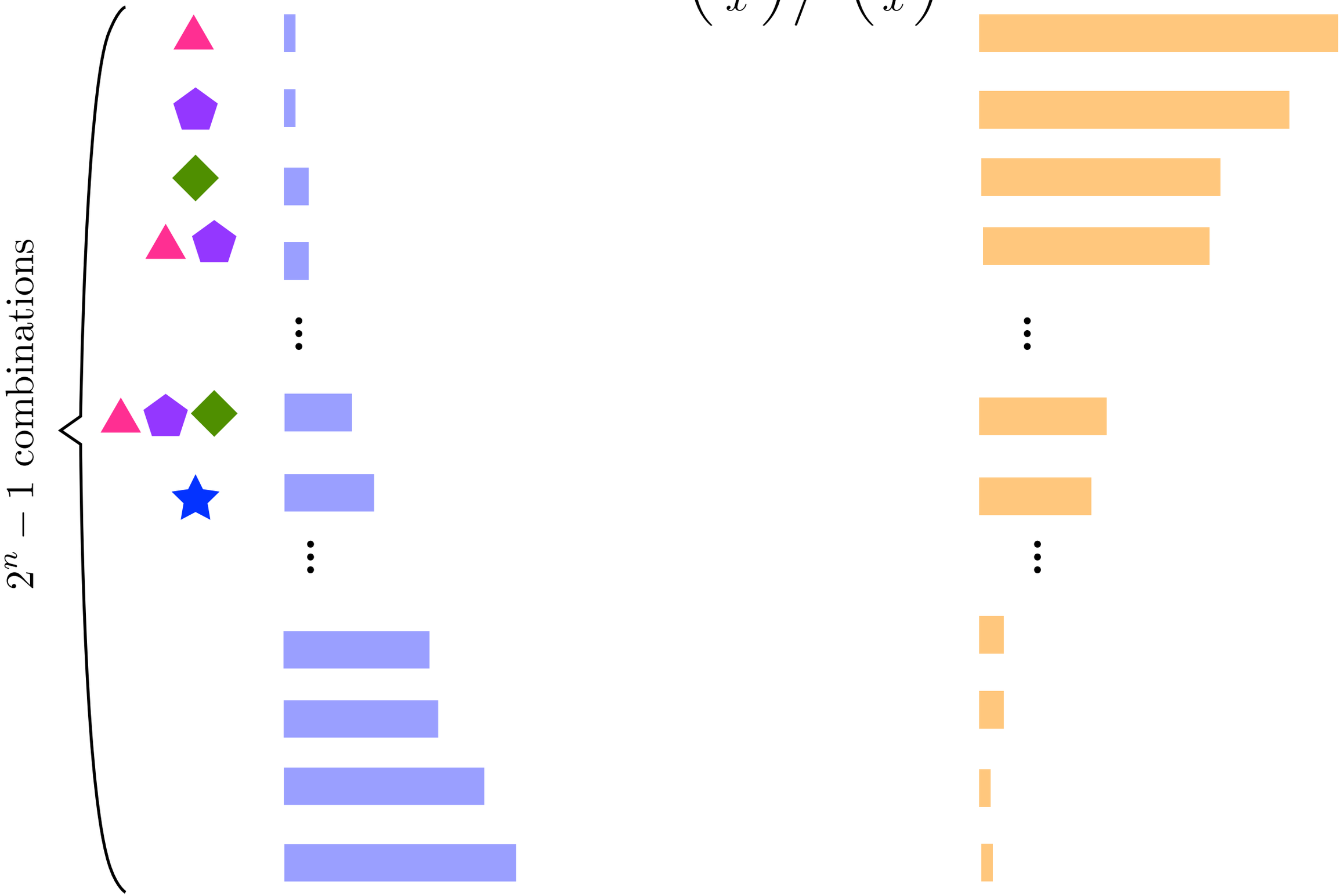
δ

Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

x: 関連している遺伝子数

$2^n - 1$ combinations



Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

x: 関連している遺伝子数

$2^n - 1$ combinations

λ

Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

x: 関連している遺伝子数

頻出パタンマイニング
[Agrawalら など]

$2^n - 1$ combinations

λ

Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

x: 関連している遺伝子数

頻出パタンマイニング
[Agrawalら など]

$2^n - 1$ combinations

$$k = |\{i | f(x_i) \leq \delta\}|$$

λ

δ

Min. P-value f(x)

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

x: 関連している遺伝子数

頻出パタンマイニング
[Agrawalら など]

$2^n - 1$ combinations

$$k = |\{i | f(x_i) \leq \delta\}|$$

λ

δ

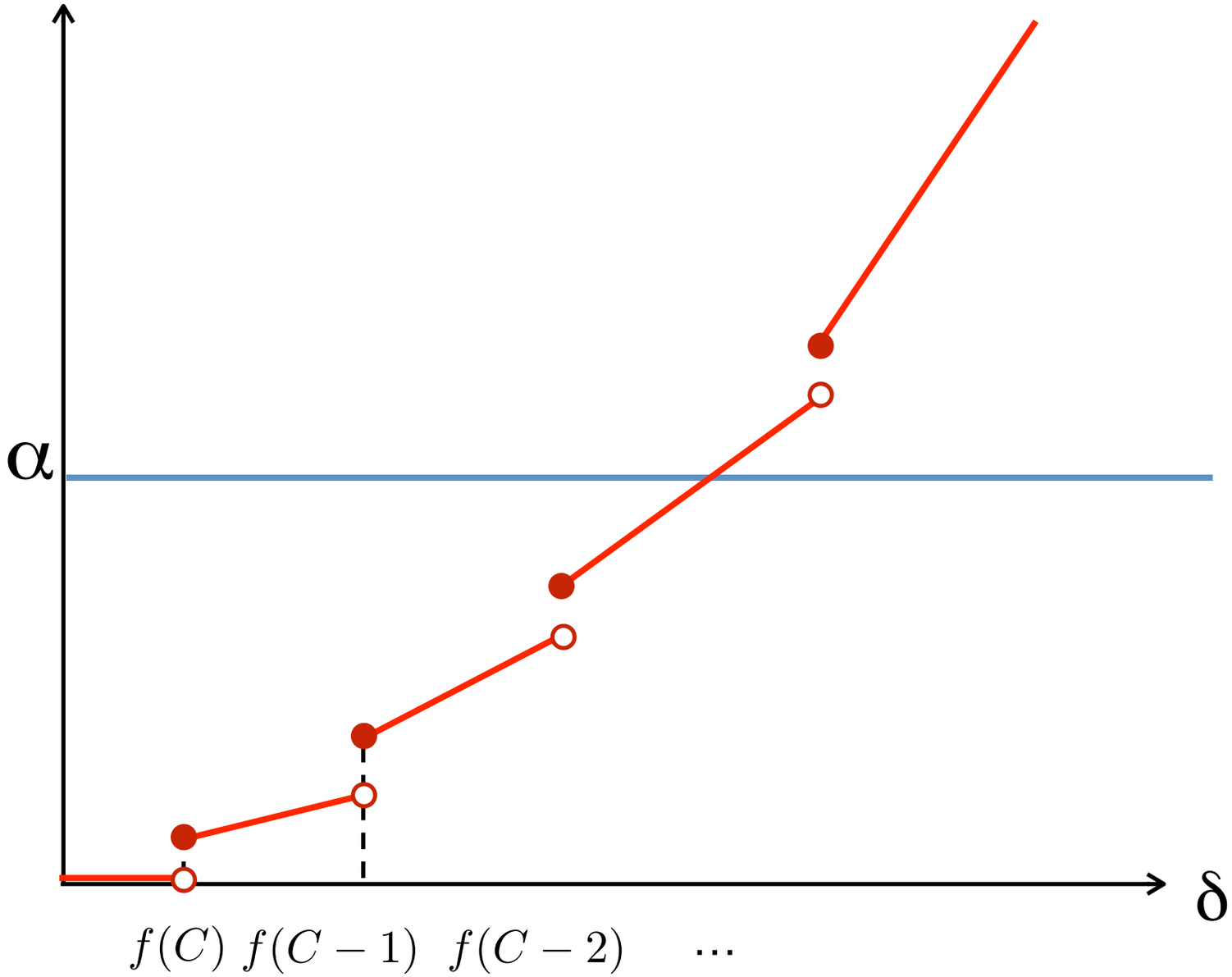
FWER の上界

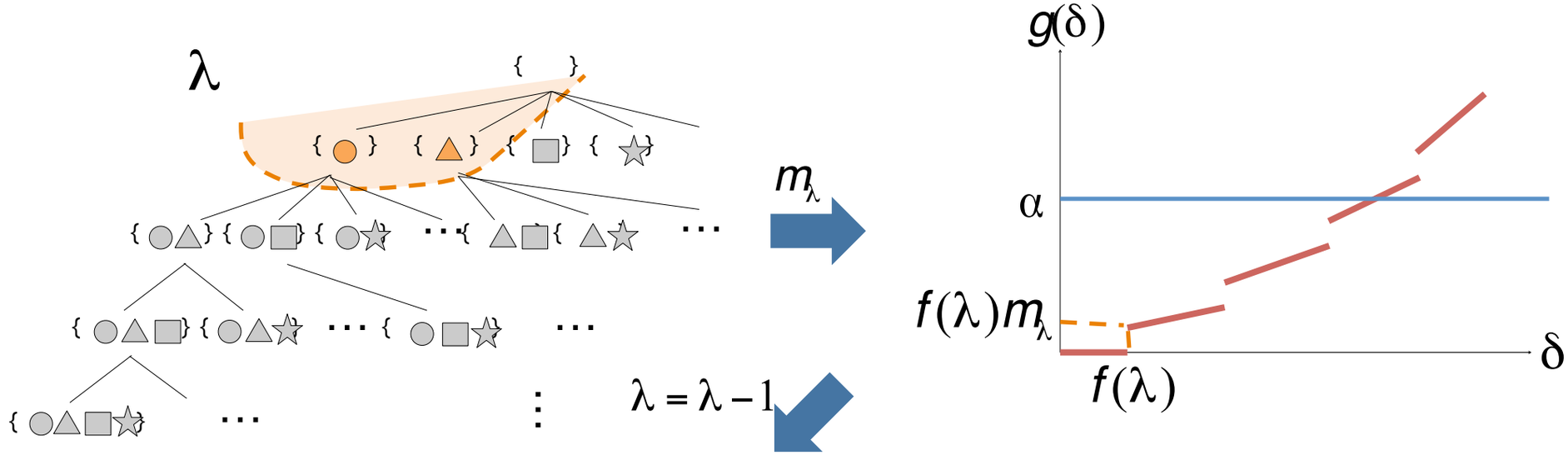
$$g(x) = k\delta$$

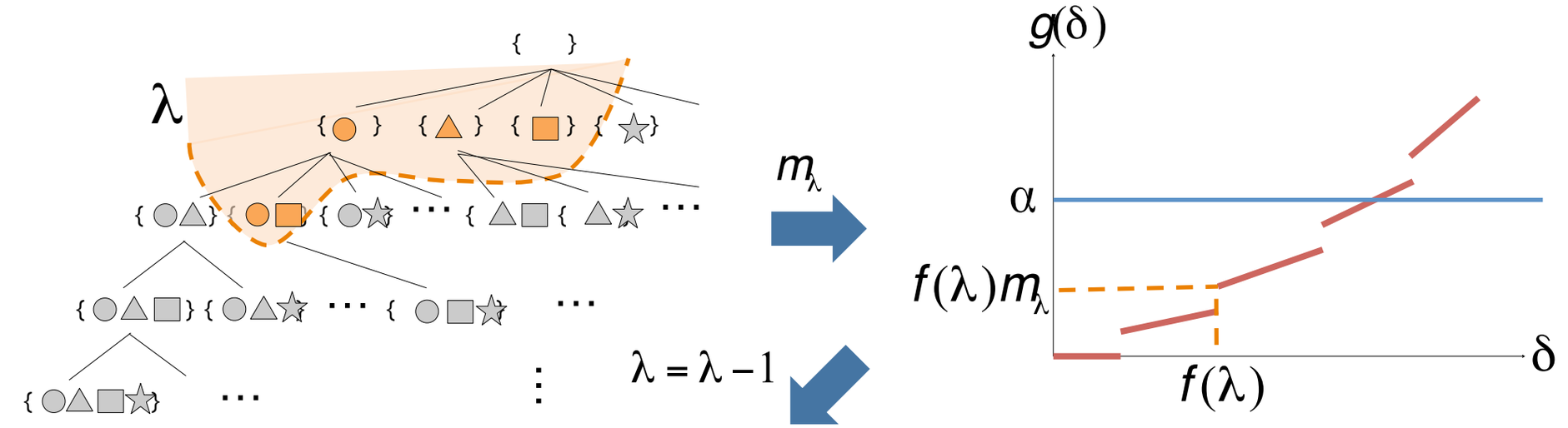
FWER の上界 $g(x)$が, α以上である範囲内で, 最小のδを求める.

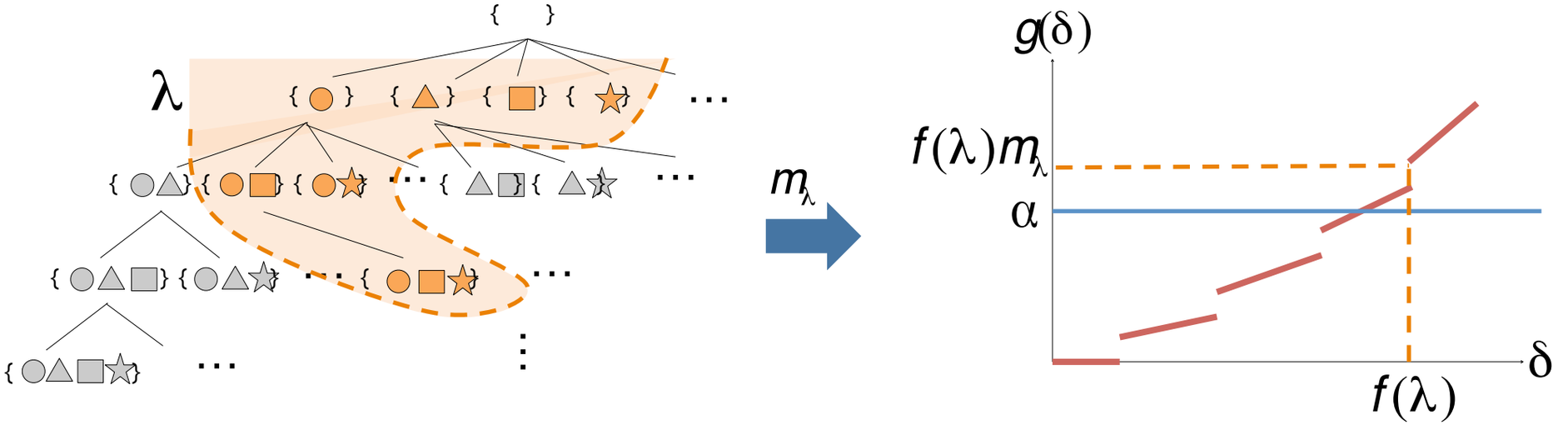
FWER の上界

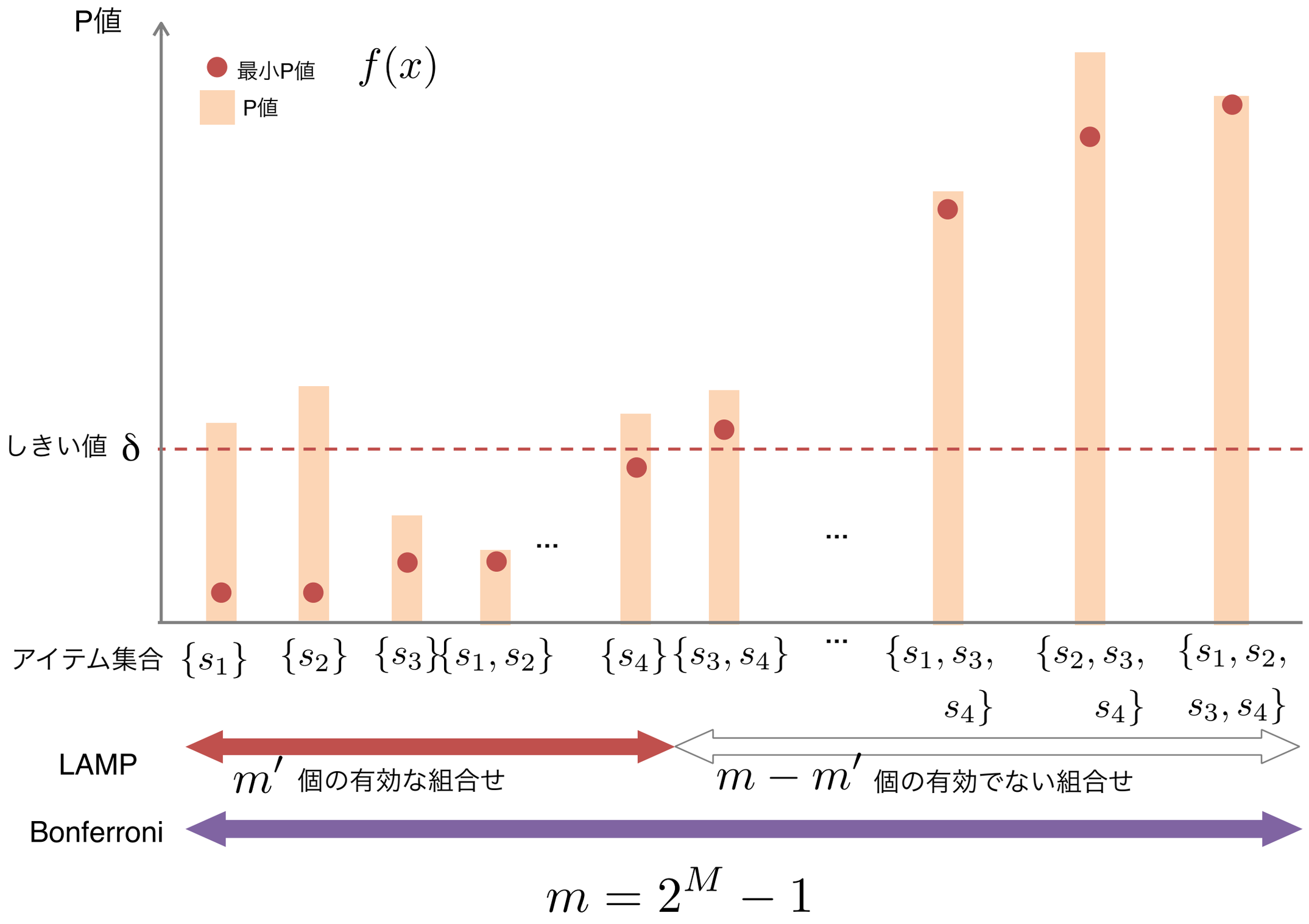
$$g(\delta) = |\{I \mid f(x(I)) \leq \delta\}|\delta$$




λ
m_λ
$\lambda = \lambda - 1$
$g(\delta)$
α
$f(\lambda)m_\lambda$
$f(\lambda)$
δ

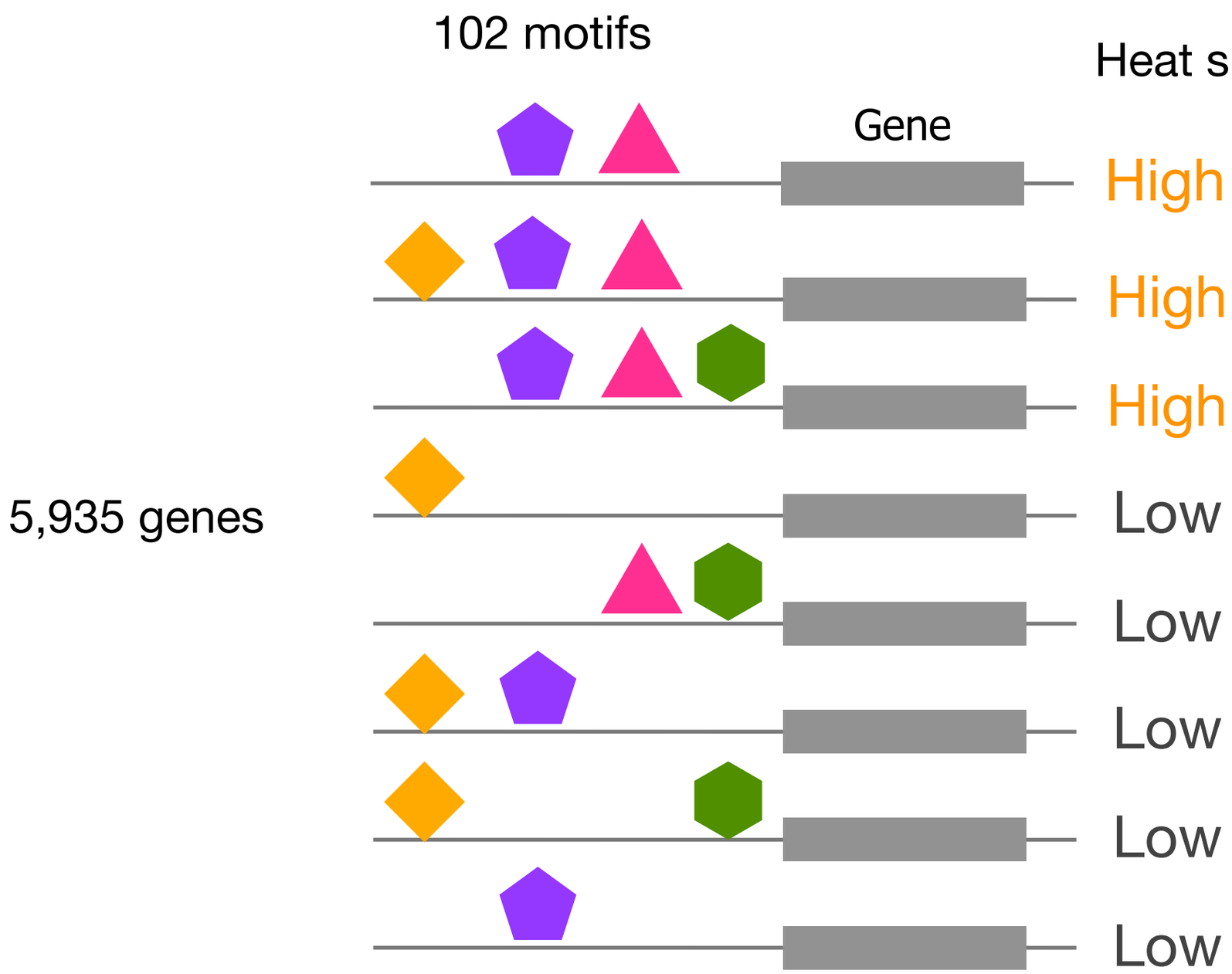

λ
m_λ
$\lambda = \lambda - 1$
$g(\delta)$
α
$f(\lambda)m_\lambda$
$f(\lambda)$
δ


λ
m_λ
$g(\delta)$
$f(\lambda)m_\lambda$
α
δ
$f(\lambda)$


P値
最小P値
P値
$f(x)$
しきい値 δ
...
...
アイテム集合
$\{s_1\}$ $\{s_2\}$ $\{s_3\}$ $\{s_1, s_2\}$ $\{s_4\}$ $\{s_3, s_4\}$ ... $\{s_1, s_3, s_4\}$ $\{s_2, s_3, s_4\}$ $\{s_1, s_2, s_3, s_4\}$
LAMP
m' 個の有効な組合せ
$m - m'$ 個の有効でない組合せ
Bonferroni

$$m = 2^M - 1$$

酵母の遺伝子制御の例

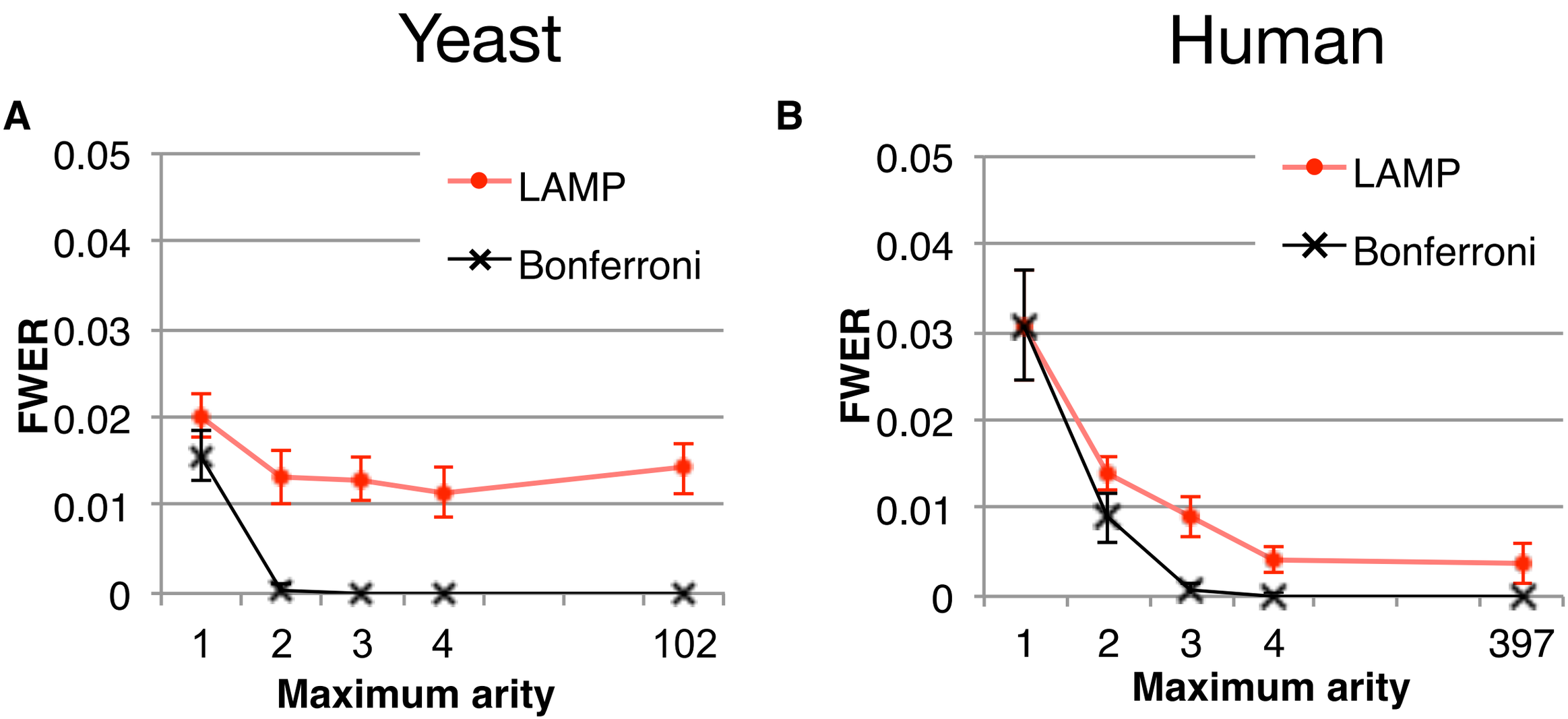


Expression: Gasch *et al.*
ChIP-Chip: Harbison *et al.*

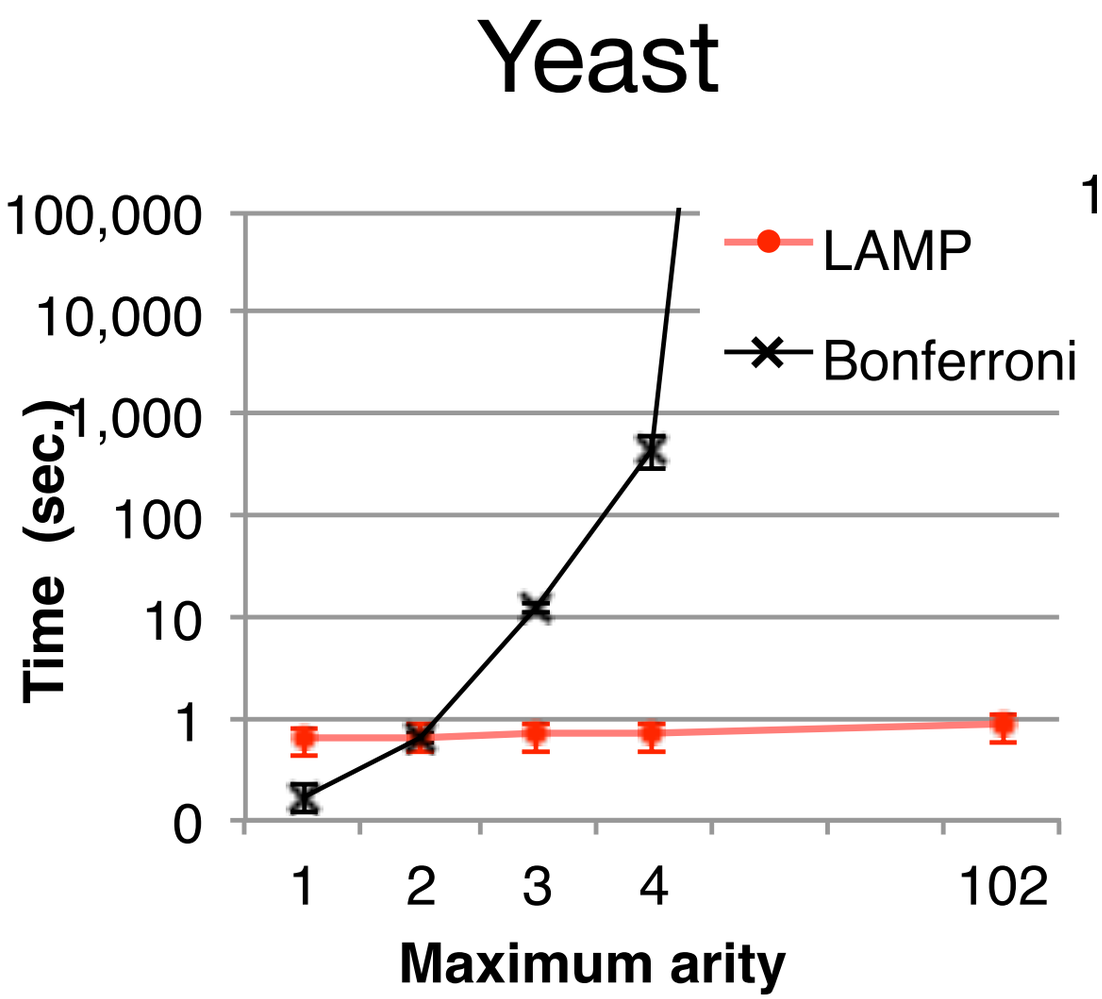
ヒト乳がん細胞での
組合せ制御 発見例

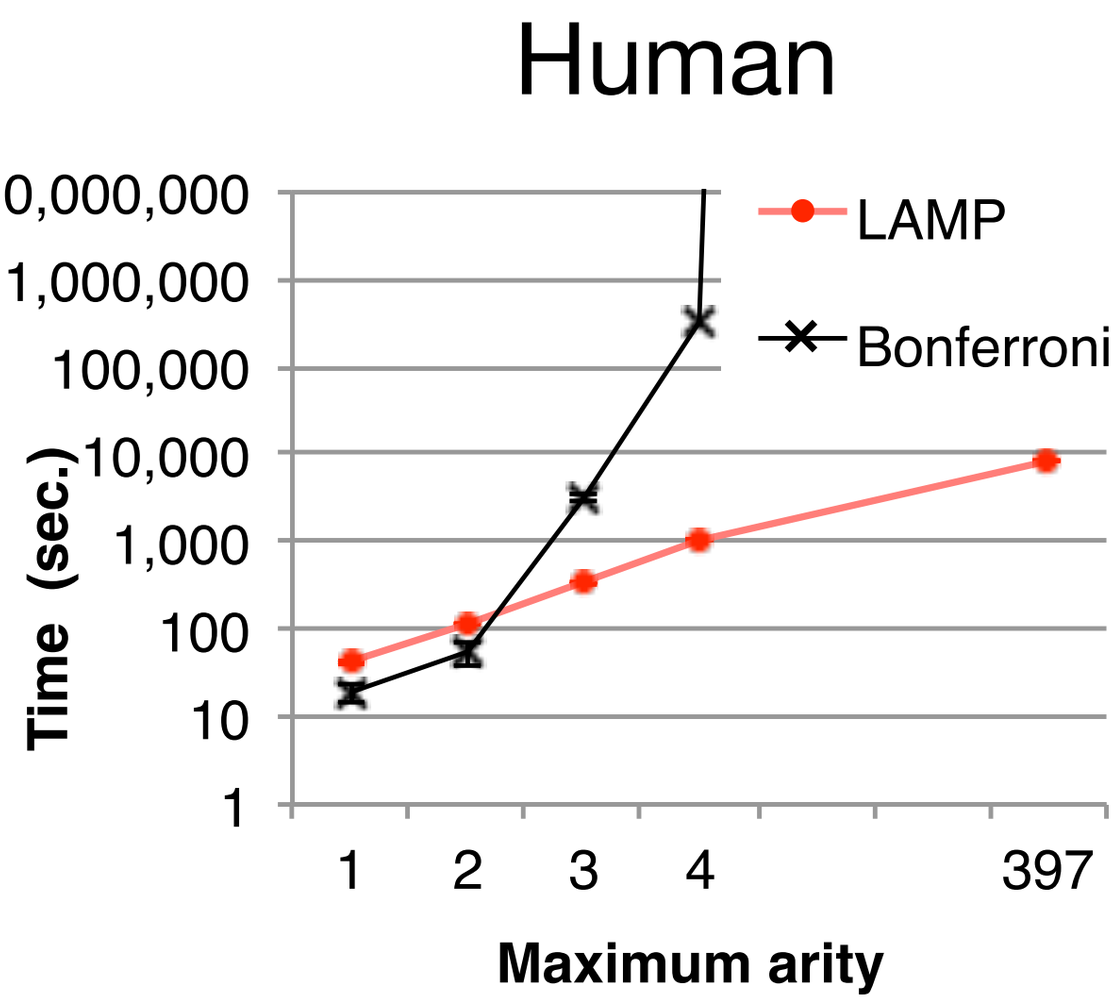
- ヒト乳がん細胞MCF-7の遺伝子発現プロファイル
- 成長因子EGF 投与後のプロファイルを利用
- モチーフ–遺伝子の関係は，MsigDBを利用
  - 397モチーフ．約12,000遺伝子が対象
  - LAMPの補正項K = 1,174,108 - 2,750,336
    - もしBonferroni法を使ったら，上限を8個までに限定しても，1.4・10^16

シミュレーションデータでの実際のFWER



実行時間


Yeast
100,000
10,000
1,000
100
10
1
0
Time (sec.)
LAMP
Bonferroni
1
2
3
4
102
Maximum arity

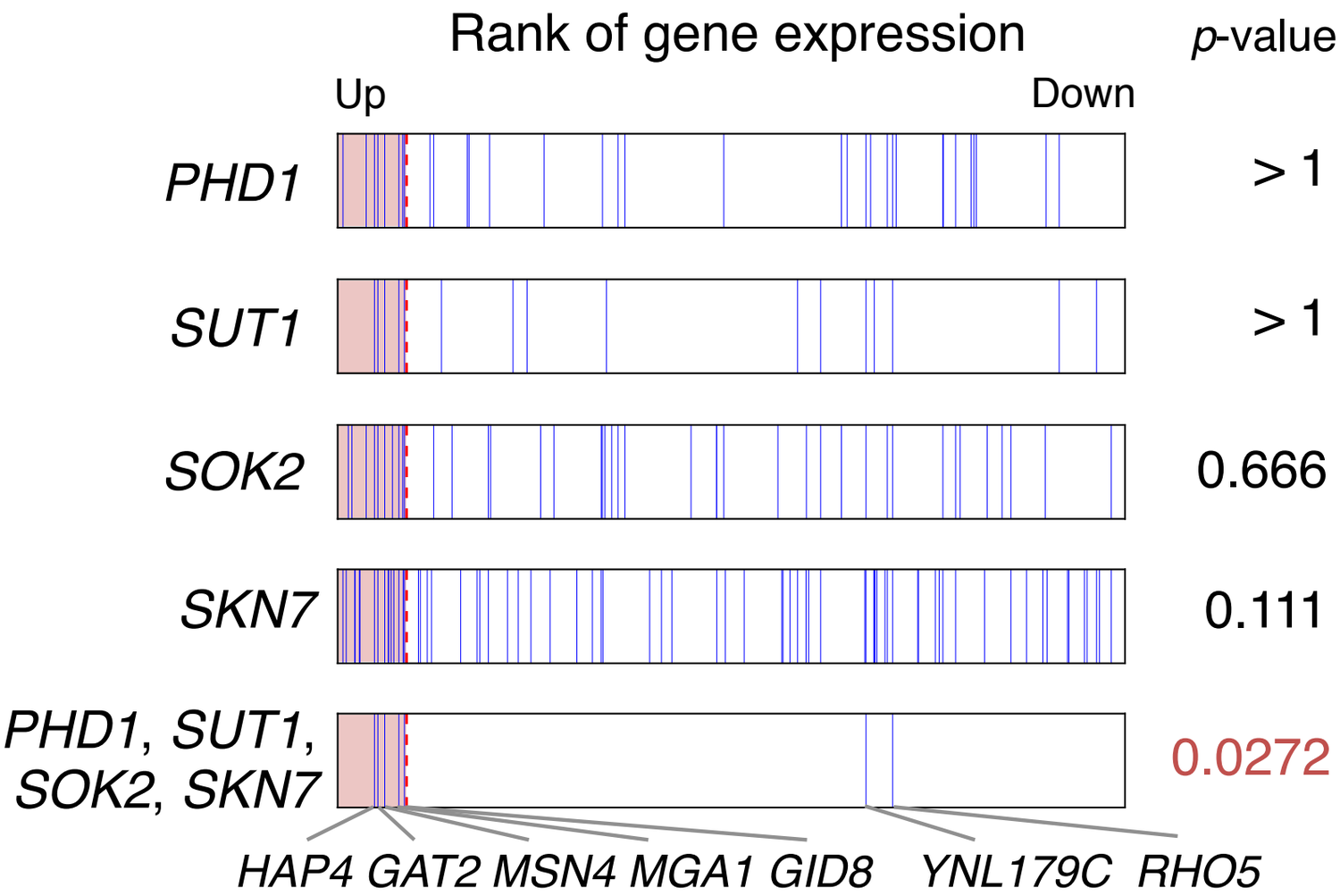

Human
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
Time (sec.)
LAMP
Bonferroni
1
2
3
4
397
Maximum arity

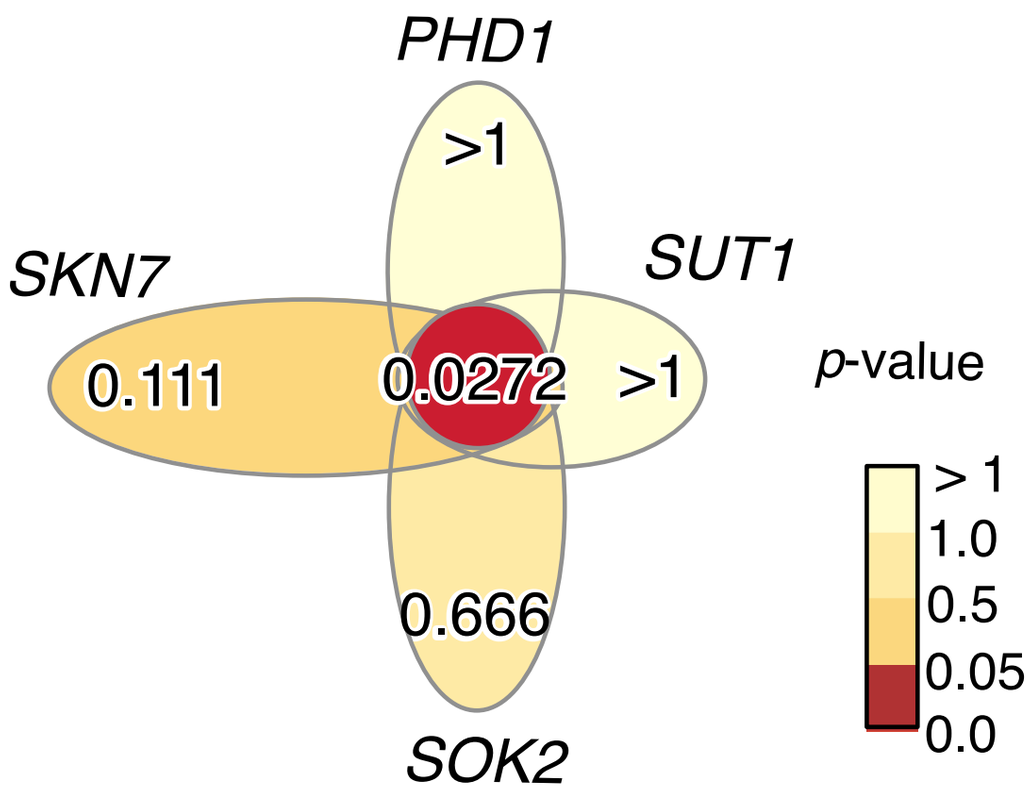
組合せを考慮した有意な
モチーフ一覧

Under heat shock condition

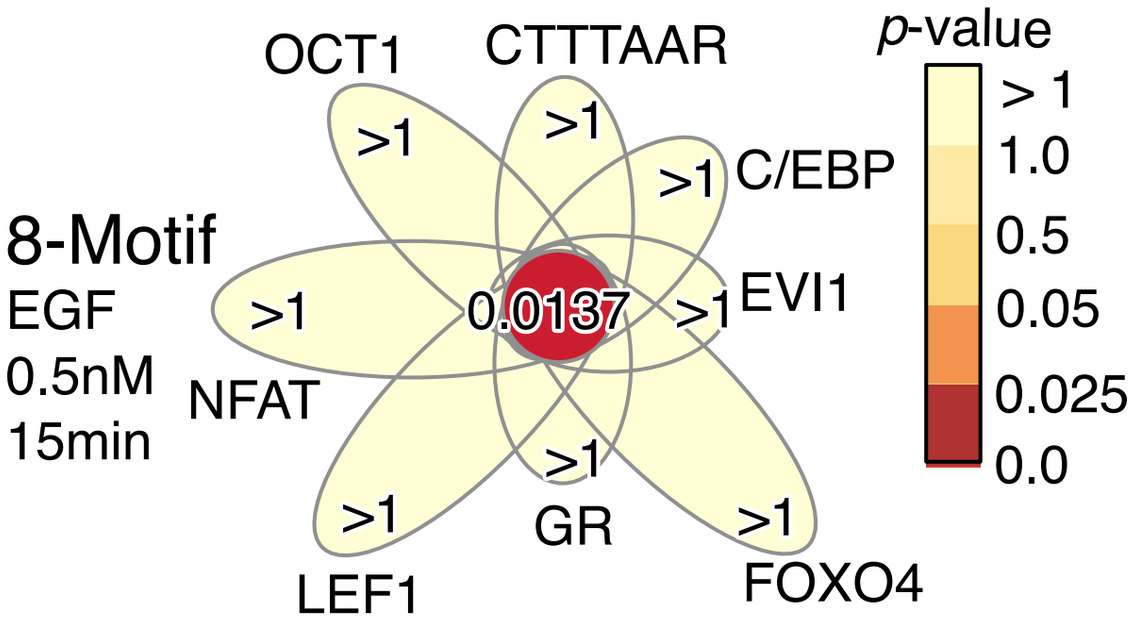
Corrected p-value. Red: significant

Motif combination	LAMP (≦ 102)	Bonferroni (≦ 4)
	K= 303	K = 4,426,528
HSF1	4.41E-24	6.44E-20
MSN2	3.73E-11	5.45E-07
MSN4	0.000532	>1
SKO1	0.00839	>1
SNT2	0.0192	>1
PHD1, SUT1, SOK2, SKN7	0.0272	>1


Rank of gene expression
p-value
Up
Down
PHD1
> 1
SUT1
> 1
SOK2
0.666
SKN7
0.111
PHD1, SUT1,
SOK2, SKN7
0.0272
HAP4 GAT2 MSN4 MGA1 GID8
YNL179C
RHO5

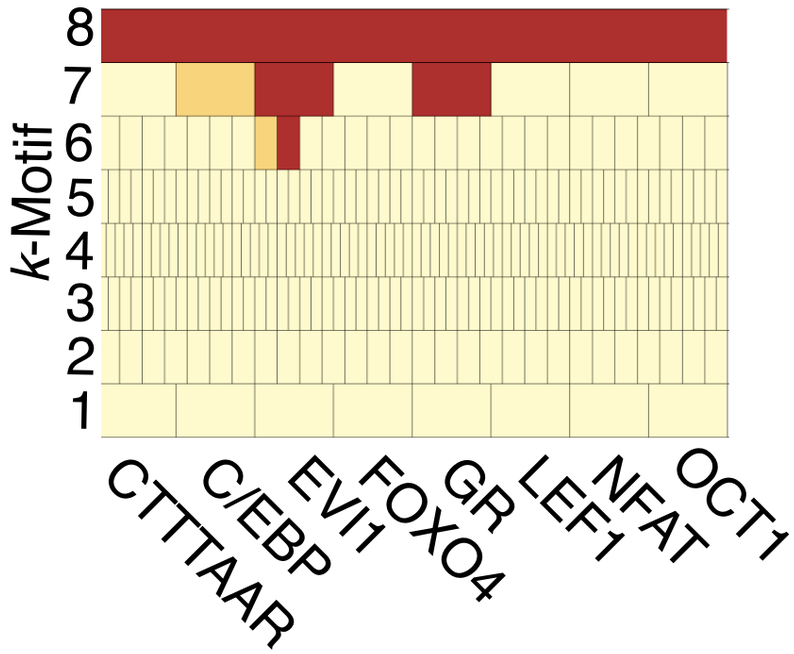

PHD1
>1
SUT1
SKN7
0.111
0.0272
>1
0.666
SOK2
p-value
> 1
1.0
0.5
0.05
0.0

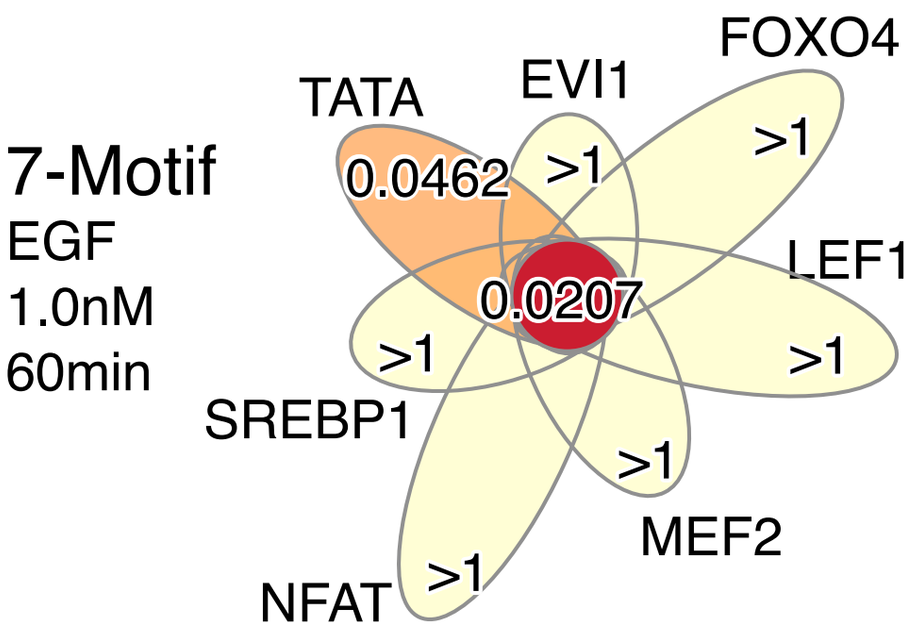
A

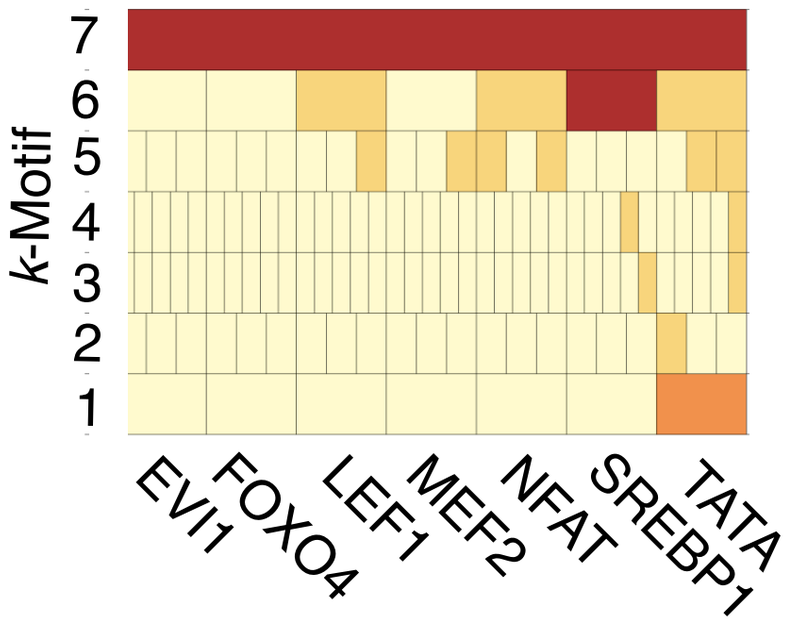


B

Expression: Nagashima *et al.*
Motif sites: Xie *et al.*







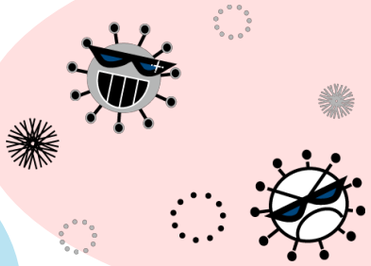
組合せを求める問題全般に利用可能

SNPsの組合せ

糖尿病

Sladek, R. *et al.*,
Nature (2007)

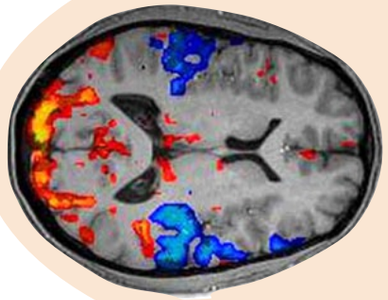
変異の組合せ



HIVウイルスの
薬物耐性

Zhang, J., et al.,
PNAS (2010).

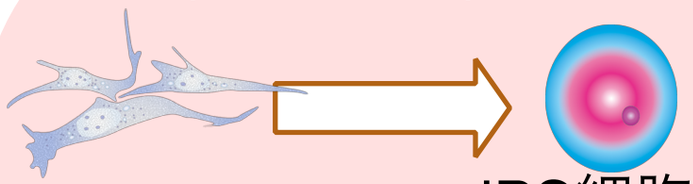
脳の機能解析



機能部位の
組合せ

転写因子の組合せ

細胞のiPS化



K. Okita *et al.*,
Nature (2007)

iPS細胞

遺伝子の相互作用

植物の開花時期

Atwell, S. *et al.*,
Nature (2010).

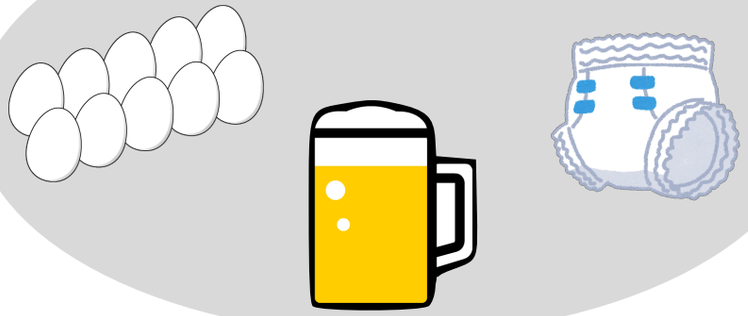
薬の組合せ

薬の飲みあわせ

アンケートの解析

アンケート項目
の組合せ

購買履歴の解析



まとめ

- 実験結果の再現性が大きな問題になっている．その中で，統計的有意性は重要な指標.
- 多次元データに対する検定として，多重検定補正が行われている
- 指標は主に２つ：FWERとFDR
  - それぞれ，複数の手法が開発されている
- LAMPについて
  - 組合せ的な要素の発見に対して，FWERを保証する方法を開発
  - 理論的には，どんな大きさの組合せも考えられる.
    - 乳がん細胞の例だと，8個より大きな組合せで有意なものが無いことも保証されている.
  - 組み合わせを考慮した，特徴選択と考えることもできる.
    - 合成変数では説明が難しいが，選択したものであれば解釈は容易