

August 10, 2016

大阪大学 基礎セミナー

知能とコンピュータ (2日目)




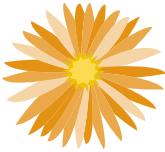
計算機による学習と発見

—統計的な解析—

大阪大学産業科学研究所 助教 杉山磨人

異常の原因となるDNAを見つける

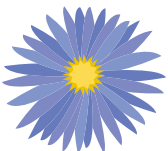
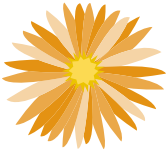
ゲノム塩基配列 (SNPs)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
ケース (病気あり) 	サンプル	1:	0	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1	0	1	0
	サンプル	2:	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
	サンプル	3:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
	サンプル	4:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
	サンプル	5:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
<hr/>																						
コントロール (病気なし) 	サンプル	6:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0
	サンプル	7:	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0
	サンプル	8:	1	0	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0
	サンプル	9:	1	1	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1

0: 通常, 1: レア

異常の原因となるDNAを見つける

ゲノム塩基配列 (SNPs)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
ケース (病気あり) 	サンプル	1:	0	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1	0	1	0	
		2:	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
		3:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1	
		4:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1	
		5:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0	
コントロール (病気なし) 	サンプル	6:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0	
		7:	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0	
		8:	1	0	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	
		9:	1	1	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1	

0: 通常, 1: レア

統計解析で結果を保証する

- **問題例：**
大量のDNAデータを解析して、病気の人と、そうでない人のDNAに、差があるかどうかを調べる
- 計算機がデータを解析して、それっぽいDNAを発見！
- このDNAが本当に重要かどうかを、**統計 (statistics)** を使って検証する
 - 「それっぽさ」を、みんなが納得できるようにする

1つのDNAに着目したときの例

- 病気ありの人：70人，そのなかで，
 - レアなDNAの人：46人
 - 通常のDNAの人：24人
- 病気なしの人：210人，そのなかで，
 - レアなDNAの人：50人
 - 通常のDNAの人：160人

データを分割表で表す

- 病気ありの人：70人，そのなかで，
 - レアなDNAの人：46人
 - 通常のDNAの人：24人
- 病気なしの人：210人，そのなかで，
 - レアなDNAの人：50人
 - 通常のDNAの人：160人

	DNAレア	DNA通常	合計
病気あり	46	24	70
病気なし	50	160	210
合計	96	184	280

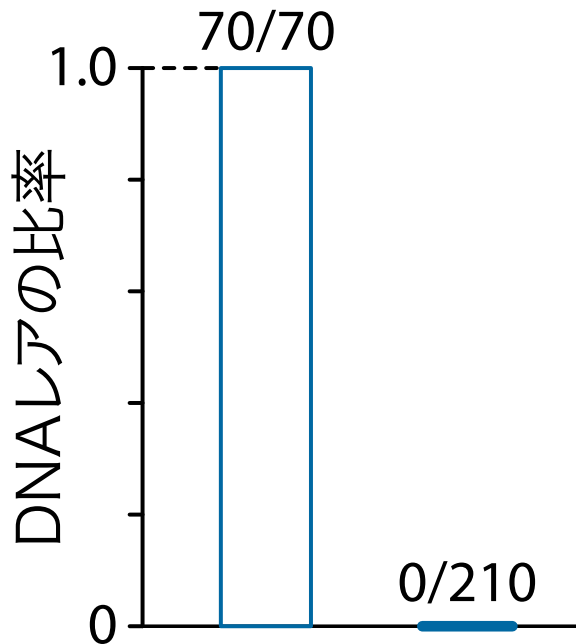
極端な場合はよいが...

OK!	DNAレア	DNA通常	合計
病気あり	70	0	70
病気なし	0	210	210
合計	70	210	280

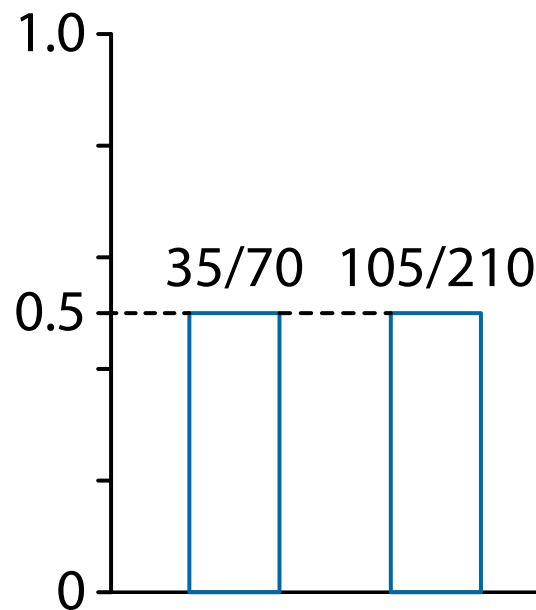
No...	DNAレア	DNA通常	合計
病気あり	35	35	70
病気なし	105	105	210
合計	140	140	280

棒グラフで書いてみる

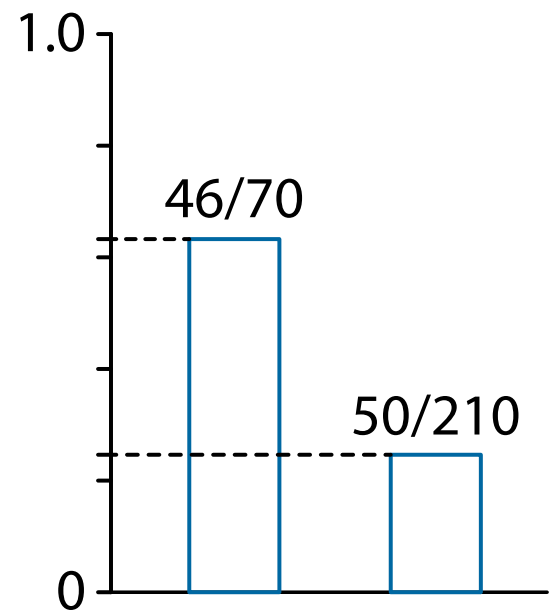
OK!



No...



??



データが「どのくらい珍しいか」を測る

	DNAレア	DNA通常	合計
病気あり	46	24	70
病気なし	50	160	210
合計	96	184	280

- 組合せを使って、確率を計算する。

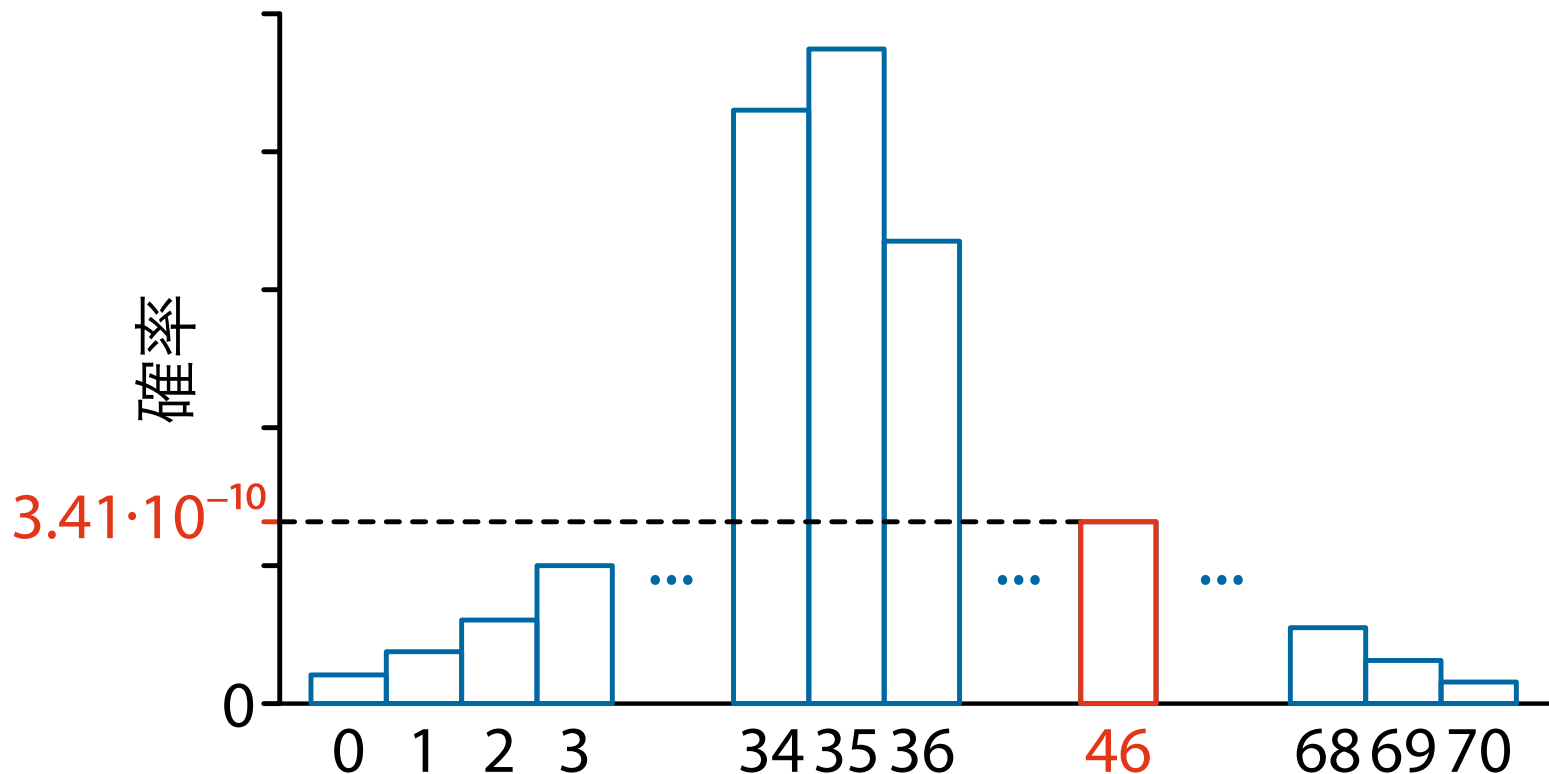
「病気のあるなし」と「このDNA」が独立のとき、

$$\text{この分割表が出てくる確率} = \frac{\binom{70}{46} \cdot \binom{210}{50}}{\binom{280}{96}} = 3.41 \cdot 10^{-10}$$

- 注意： $\binom{70}{46} = {}_{70}C_{46}$ だが、 ${}_{70}C_{46}$ という書き方は使わない

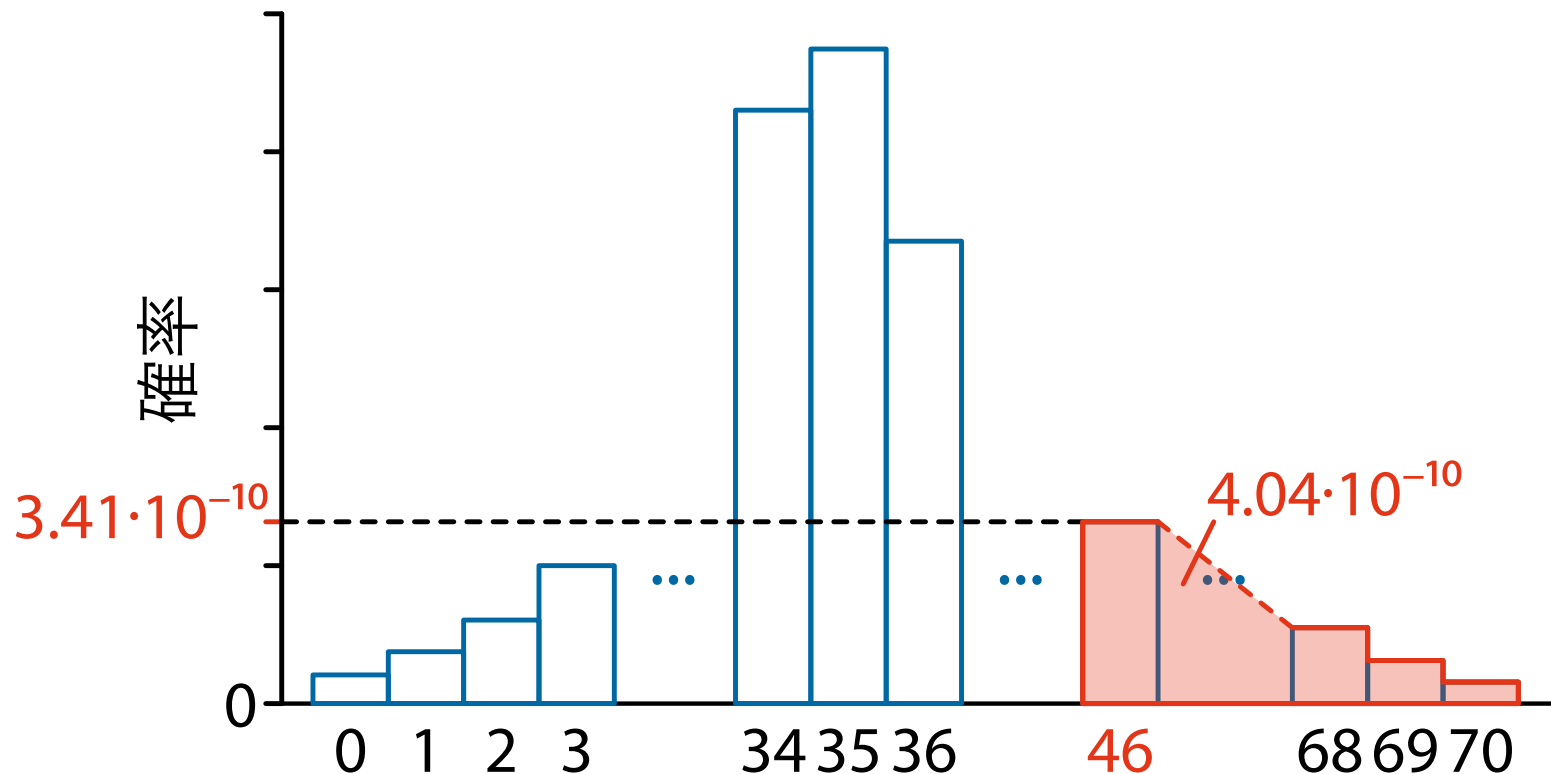
分割表から導かれる確率分布

- 「合計」の値を固定したまま、「病気あり×DNAレア」の値を0から70まで変えてみると、**確率分布（総和が1）**になる



データから得られる確率

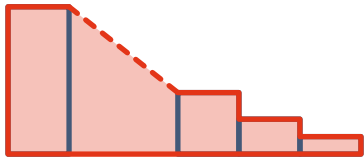
- データを基準にして、それよりも極端な場合の確率を合計。
この値が小さければ、このDNAは病気と関係があるのでは？



この確率 の意味とは？

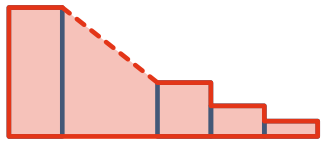
	本当は関係あり	本当は関係なし
ありと判定	真陽性 (true positive)	偽陽性 (false positive)
なしと判定	偽陰性 (false negative)	真陰性 (true negative)

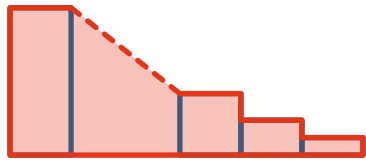
- このDNAが病気と本当は関係あるかないか、また、データからありと判定するかどうかで、4通りある



は偽陽性を表す

	本当は関係あり	本当は関係なし
ありと判定	真陽性 (true positive)	偽陽性 (false positive)
なしと判定	偽陰性 (false negative)	真陰性 (true negative)

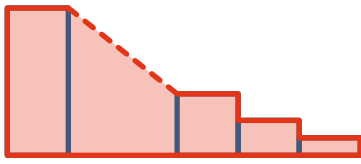
- データから計算した確率  は、「ありと判定」したDNAが「本当は関係なし」だったときの確率なので、偽陽性の割合になっている



は p 値と呼ばれる

1. あらかじめ, しきい値 α を自分で決める

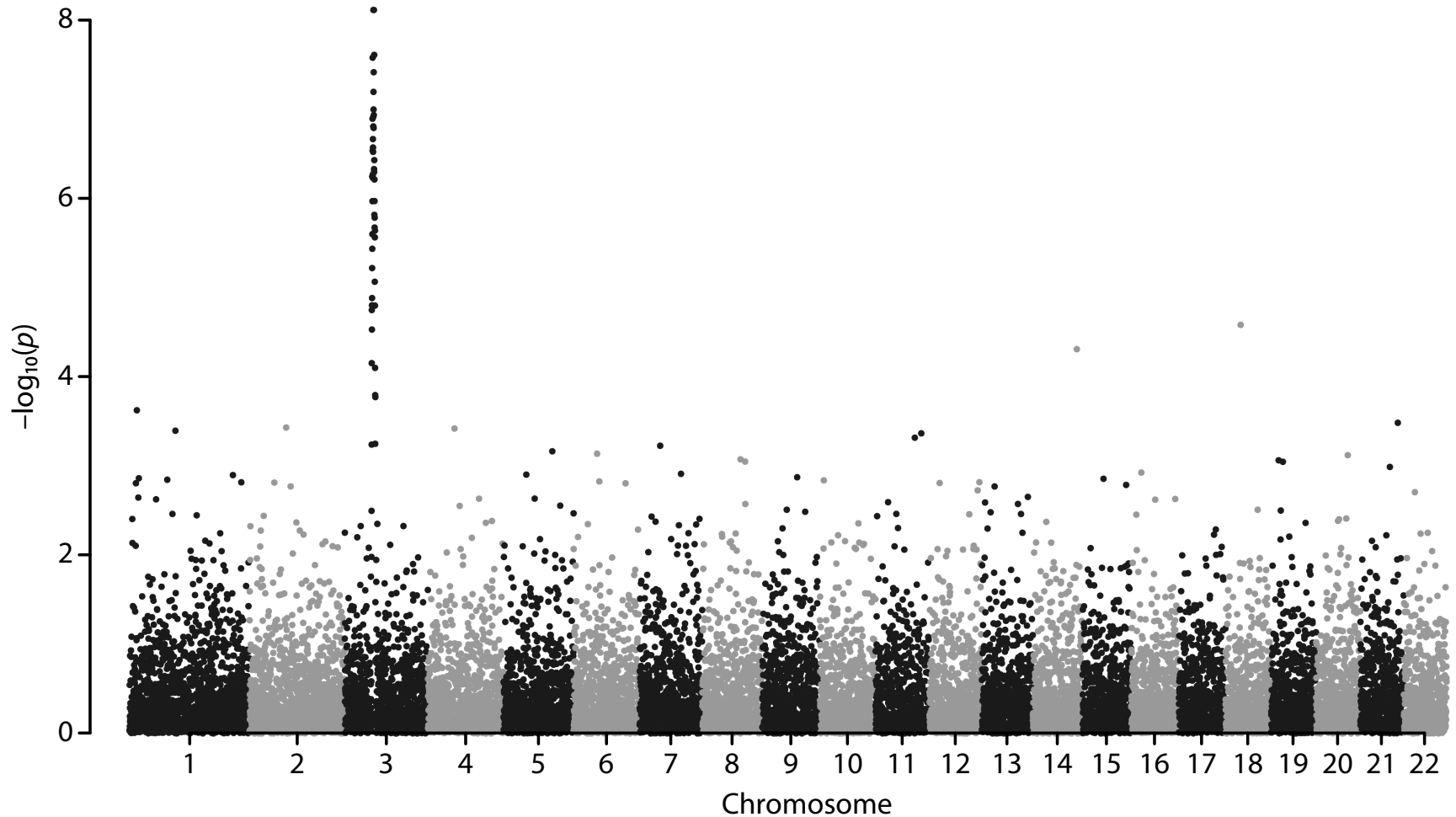
- α は, 慣習的に 0.05 や 0.01 がよく用いられる



2. p 値 が, α より小さいとき,
この DNA が病気と「関係あり」と結論づける


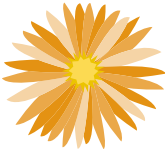
- この一連の手続きは, 仮説検定と呼ばれる
- この DNA が実は関係ない, というリスク (偽陽性) を α 以下に制御している
 - 直感的には, DNA の「それっぽさ」を, p 値という指標で統計的に測っている

DNA データでの例 (Manhattan plot)



次は、塩基のペアを見つける

ゲノム塩基配列 (SNPs)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
ケース (病気あり) 	ハンブル	1:	0	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1	0	1	0	
		2:	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
		3:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1	1
		4:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1	1
		5:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0
<hr/>																							
コントロール (病気なし) 	ハンブル	6:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0	
		7:	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0	
		8:	1	0	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0
		9:	1	1	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1	1

片方では似ているが、もう片方では似ていないペアを探す 0: 通常, 1: レア

そもそも全ペアチェックが難しい

- 塩基数が少ないときは、問題ない
- 塩基数が多くなると、全てのペアを確認することは不可能
 - 例えば、塩基が100万個あるとき、ペア数は
$$\binom{1000000}{2} = {}_{1000000}C_2 = \frac{1000000 \cdot 999999}{2} \approx 10^{12} !!$$
 - さらに、サンプル数が1000とかだと、 10^{15} のチェックが必要!!
- 解決策：「統計」と「アルゴリズム」両方の力を使う
 - ライトバルブアルゴリズム (lightbulb algorithm)
 - Paturi et al. (1995), Achlioptas et al. (2011)

似てるペアを探す

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1:	0	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1	0	1	0
2:	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
3:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
4:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
5:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
6:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
7:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
8:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
9:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0
10:	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1
11:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0

1. ランダムにサンプリングする (統計)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1:	0	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1	0	1	0
2:	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
3:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
4:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
5:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
6:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
7:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
8:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
9:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0
10:	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1
11:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0

1. ランダムにサンプリングする (統計)

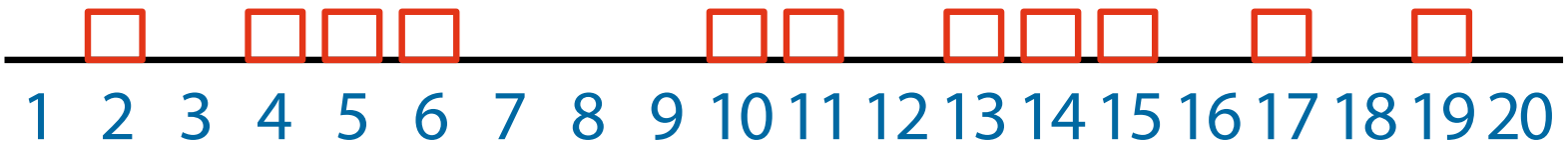
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
4:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
5:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
7:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
9:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0

2. 並べ替える (基数ソート; アルゴリズム)

	2	6	15	4	17	5	11	8	10	19	9	12	16	18	3	1	20	13	14	7
3:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
4:	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1
5:	0	0	0	0	0	1	1	1	0	0	0	1	1	0	0	1	0	0	0	1
7:	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1
9:	0	0	0	1	1	0	0	1	0	0	1	0	1	0	1	0	0	1	1	0

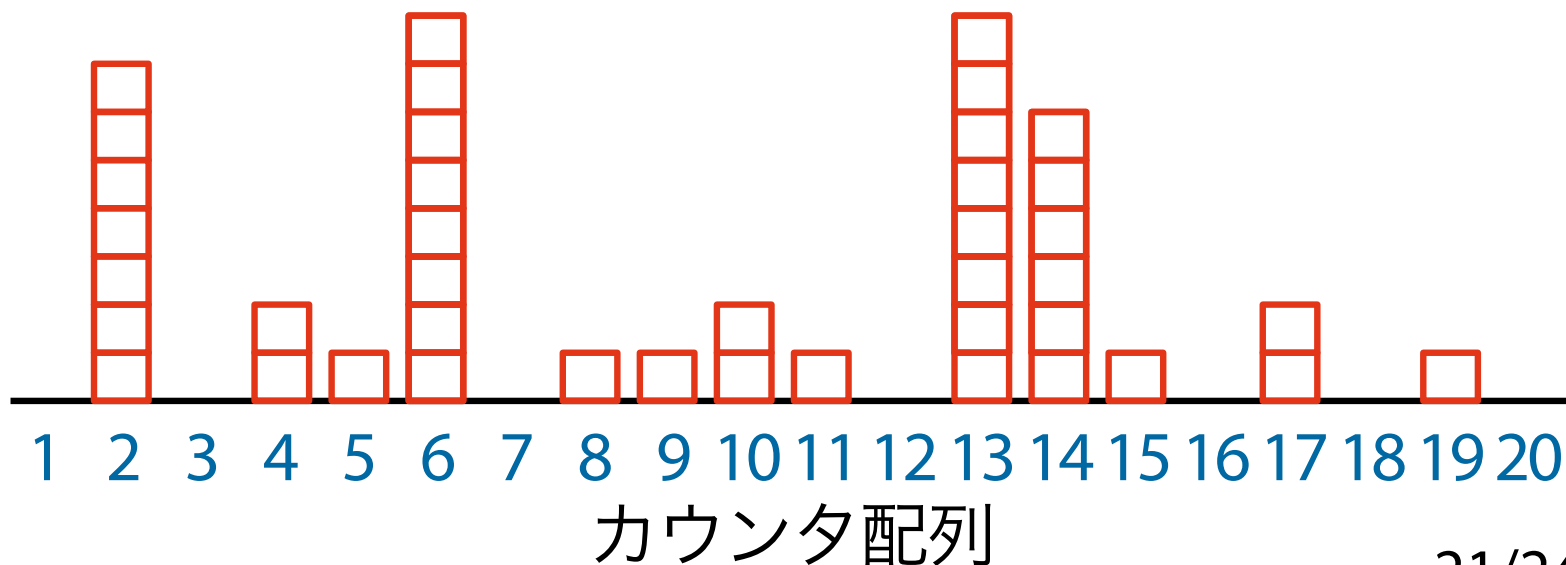
3. 同一のものをカウントする

	2	6	15	4	17	5	11	8	10	19	9	12	16	18	3	1	20	13	14	7
3:	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
4:	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1
5:	0	0	0	0	0	1	1	1	0	0	0	1	1	0	0	1	0	0	0	1
7:	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1
9:	0	0	0	1	1	0	0	1	0	0	1	0	1	0	1	0	0	1	1	0

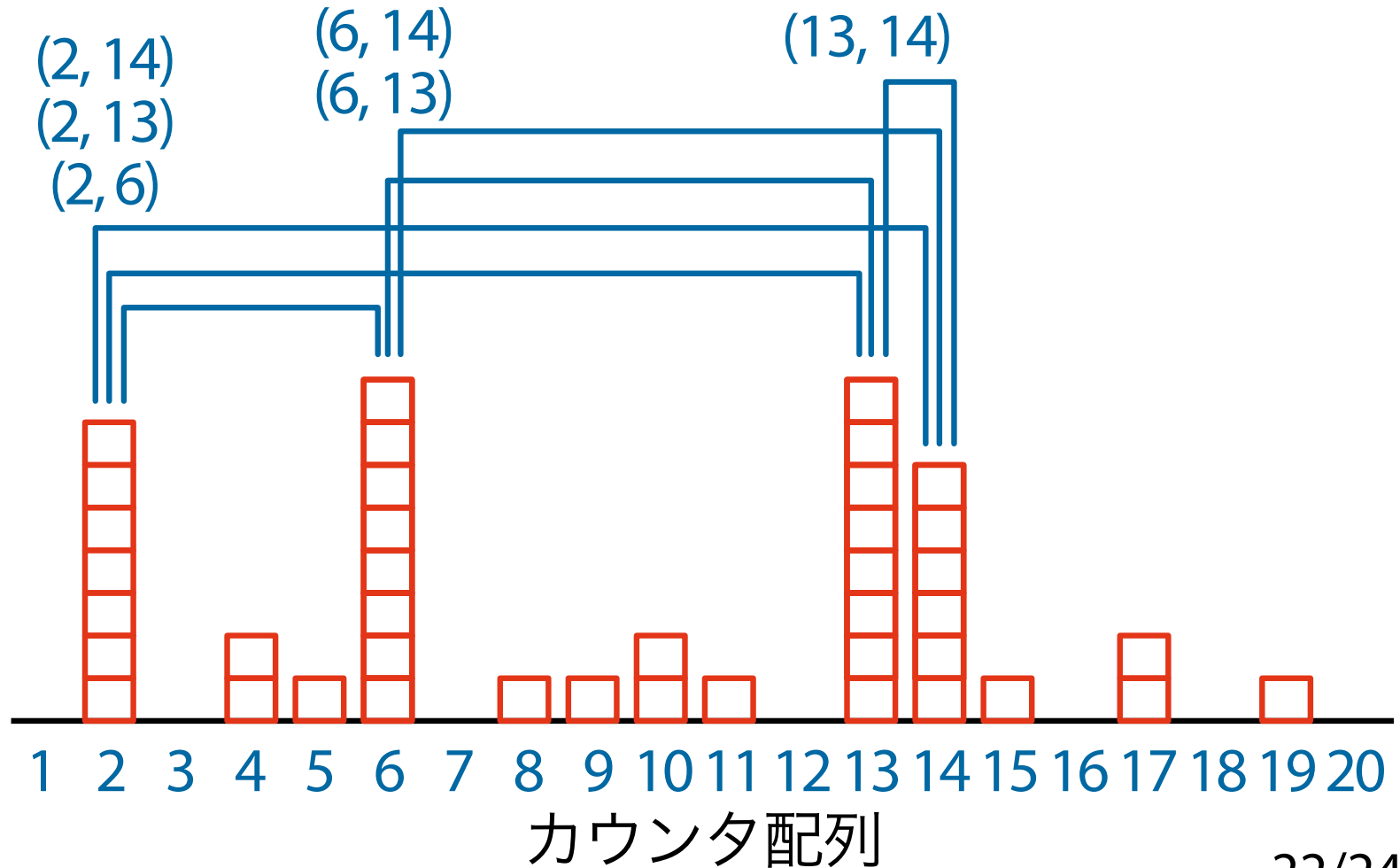


カウンタ配列

4. 以上の手順を繰り返す



5. カウンタの大きい同士のみを比較する



よく似ているペアが見つかる

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1:	0	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1	0	1	0
2:	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
3:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
4:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
5:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
6:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
7:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
8:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
9:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0
10:	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1
11:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0

まとめ

- 統計を使ったデータ解析について2つのトピックを紹介
1. どうやって正当性を担保するのか（「それっぽさ」を測る）
 - p 値を使った偽陽性の制御
 - 統計的な仮説検定
 2. どうやって計算を効率化するのか
 - 「統計」と「アルゴリズム」の融合