

December 13, 2016



Statistical Analysis

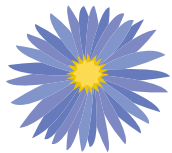
Data Mining Theory (データマイニング工学)

Mahito Sugiyama (杉山磨人)

Find Causal DNAs

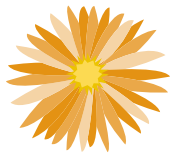
Genome sequence (SNPs)

Case
(disease)



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

Control
(health)



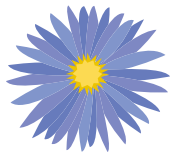
Samples	6:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0
	7:	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0
	8:	1	0	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0
	9:	1	1	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1

0: Normal, 1: Rare

Find Causal DNAs

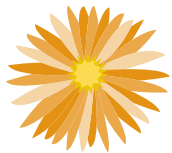
Genome sequence (SNPs)

Case
(disease)



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

Control
(health)



Samples	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: Normal, 1: Rare

Guarantee Results by Stat. Analysis

- **Example problem:**
Analyze DNA data and find the difference between people with a disease (**cases**) and people without (**controls**)
- Find some suspected DNAs using a computer!
- We need to validate the DNAs by **statistics**

Example of Single DNA

- The sample size of cases (with a disease): 70
 - Rare DNA: 46
 - Normal DNA: 24
- The sample size of control (without): 210
 - Rare DNA: 50
 - Normal DNA: 160

Represent Data by Contingency Table

- The sample size of cases (with a disease): 70
 - Rare DNA: 46
 - Normal DNA: 24
- The sample size of control (without): 210
 - Rare DNA: 50
 - Normal DNA: 160

	Rare DNA	Normal DNA	Total
Cases	46	24	70
Controls	50	160	210
Total	96	184	280

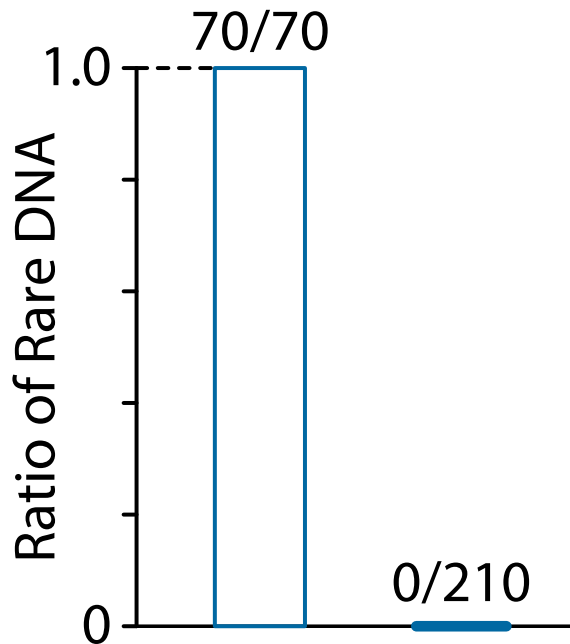
It is OK in an extreme case, but...

OK!	Rare DNA	Normal DNA	Total
Cases	70	0	70
Controls	0	210	210
Total	70	21	280

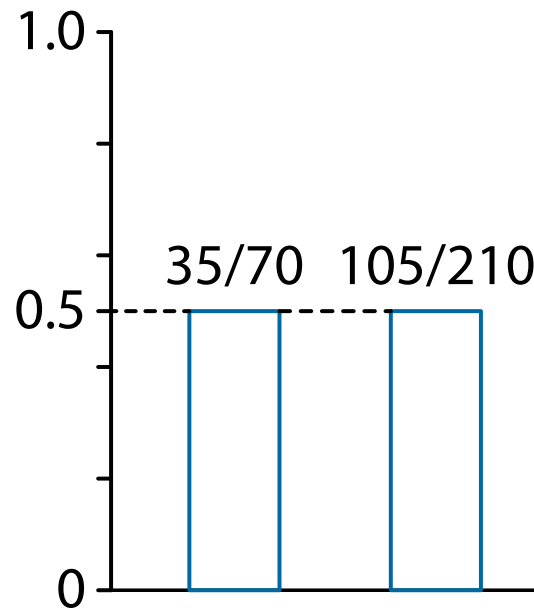
No...	Rare DNA	Normal DNA	Total
Cases	35	35	70
Controls	105	105	210
Total	140	140	280

Illustrate Barplots

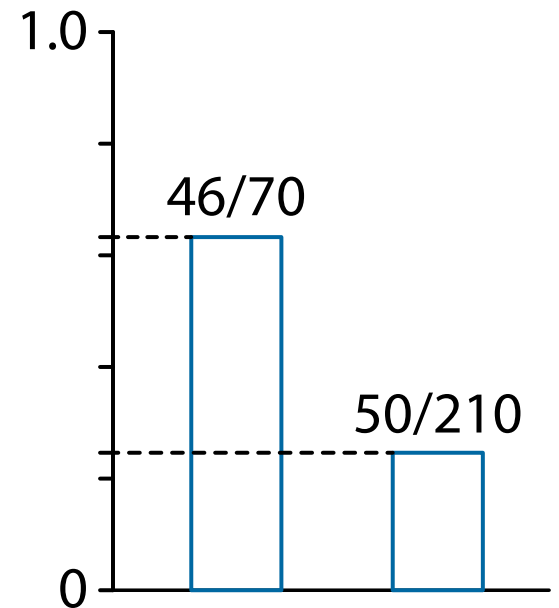
OK!



No...



??



Measure the Rareness of Data

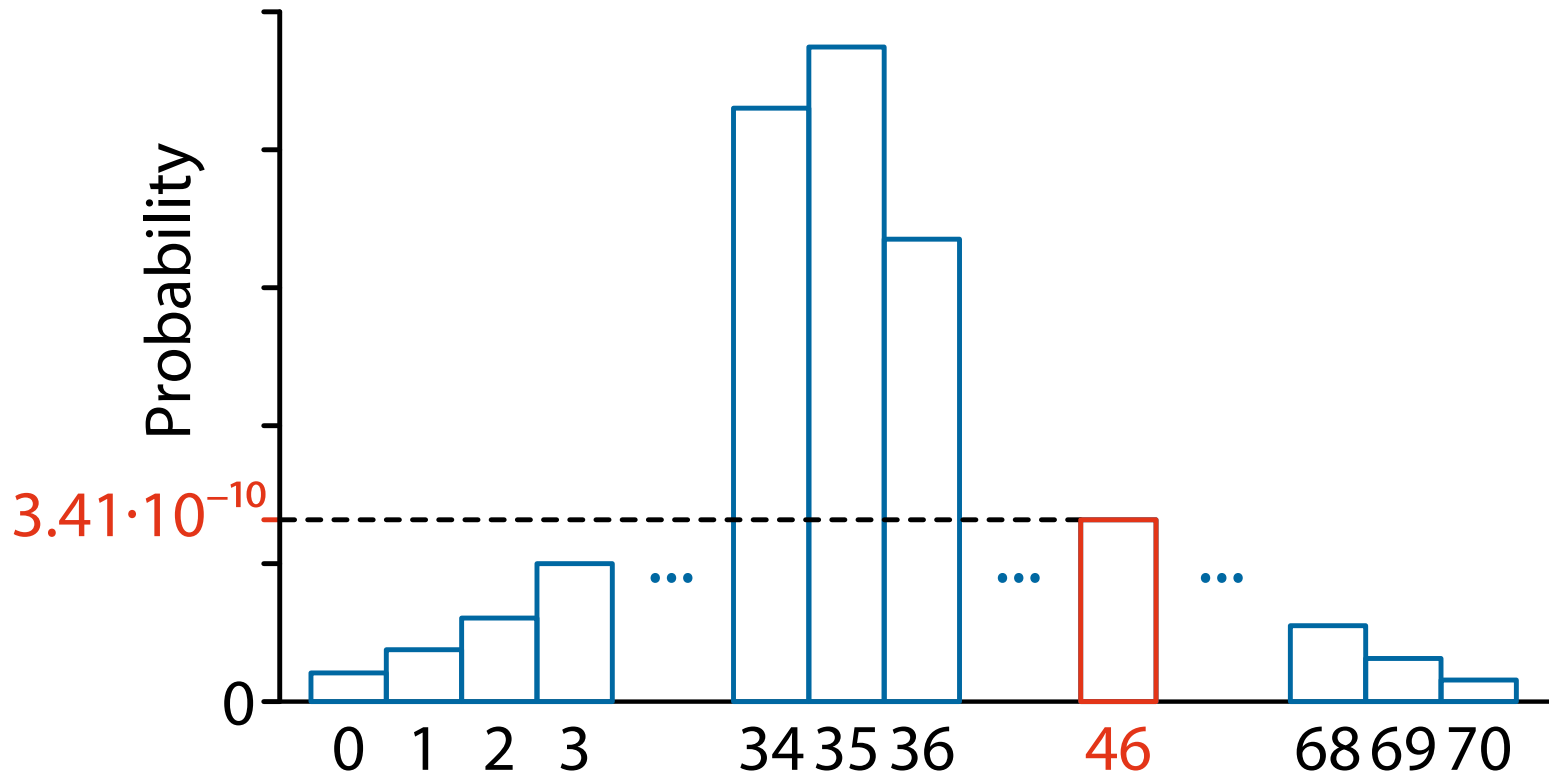
	Rare DNA	Normal DNA	Total
Cases	46	24	70
Controls	50	160	210
Total	96	184	280

- Compute probability using combinations.
If the disease and the DNA are independent,

$$\text{Prob. of observing this table} = \frac{\binom{70}{46} \cdot \binom{210}{50}}{\binom{280}{96}} = 3.41 \cdot 10^{-10}$$

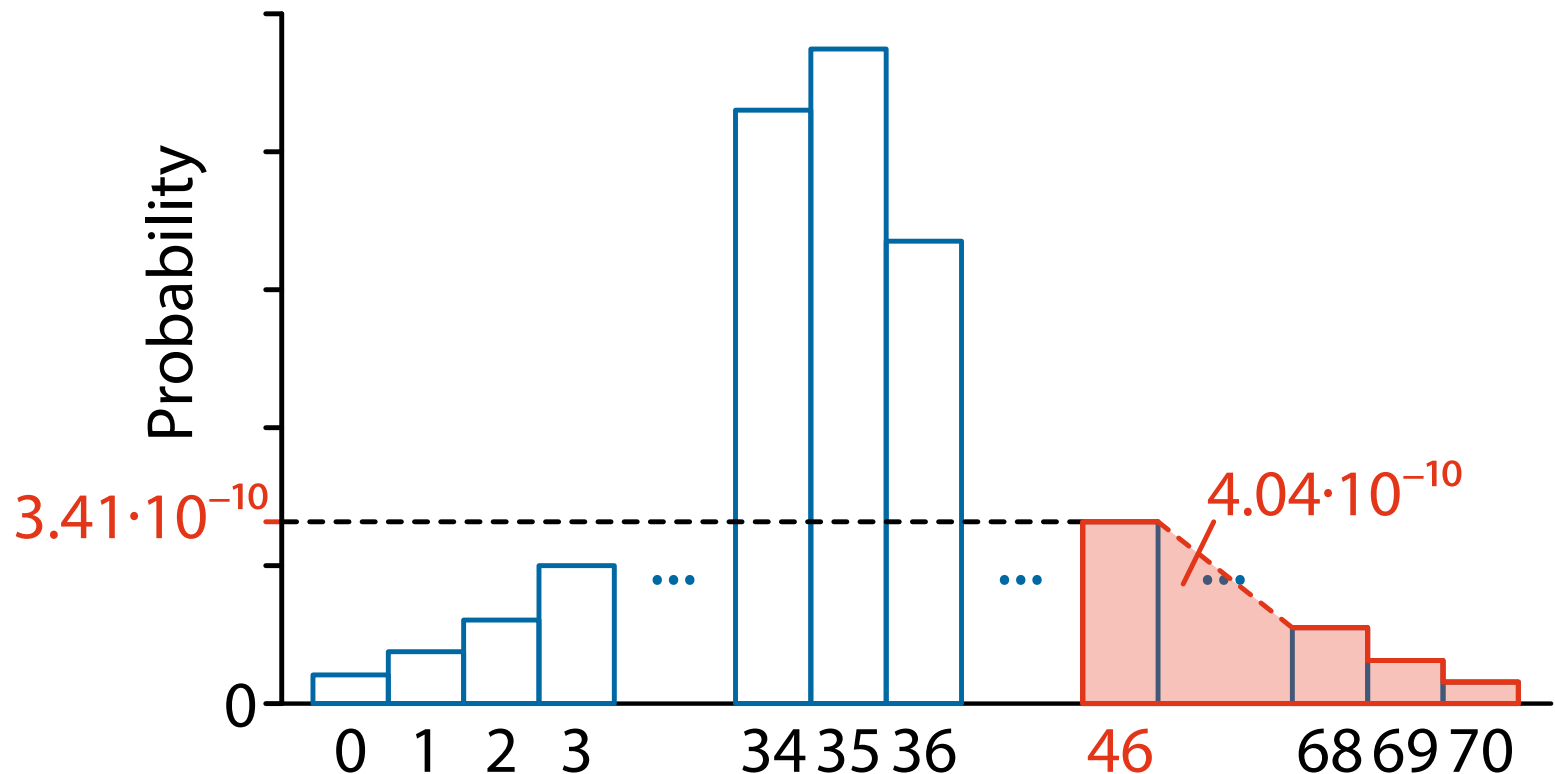
Contingency Table Induces Distribution

- Changing the value of “cases × Rare DNA” from 0 to 70 with fixing the marginal leads to a **probability distribution**

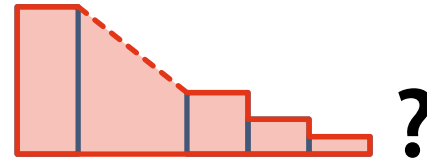


Probability from Data

- Sum up probabilities of more extreme cases than data. If the value is small, the DNA is associated with the disease?

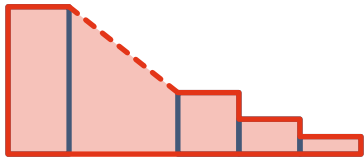


What is the Probability



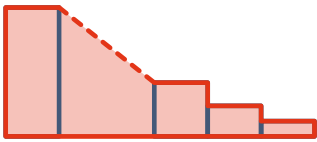
	Truly associated	Truly independent
Predicted associated	True positive	False positive
Predicted independent	False negative	True negative

- There are four cases whether the DNA is associated with the disease and whether we predict to be associated



is False Positive Rate

	Truly associated	Truly independent
Predicted associated	True positive	False positive
Predicted independent	False negative	True negative

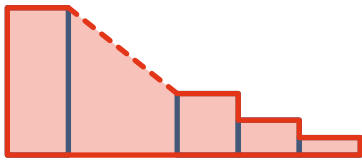
- The probability  computed from data is the **false positive rate** as it is the case in which the DNA is “predicted as associated” but is actually “independent”



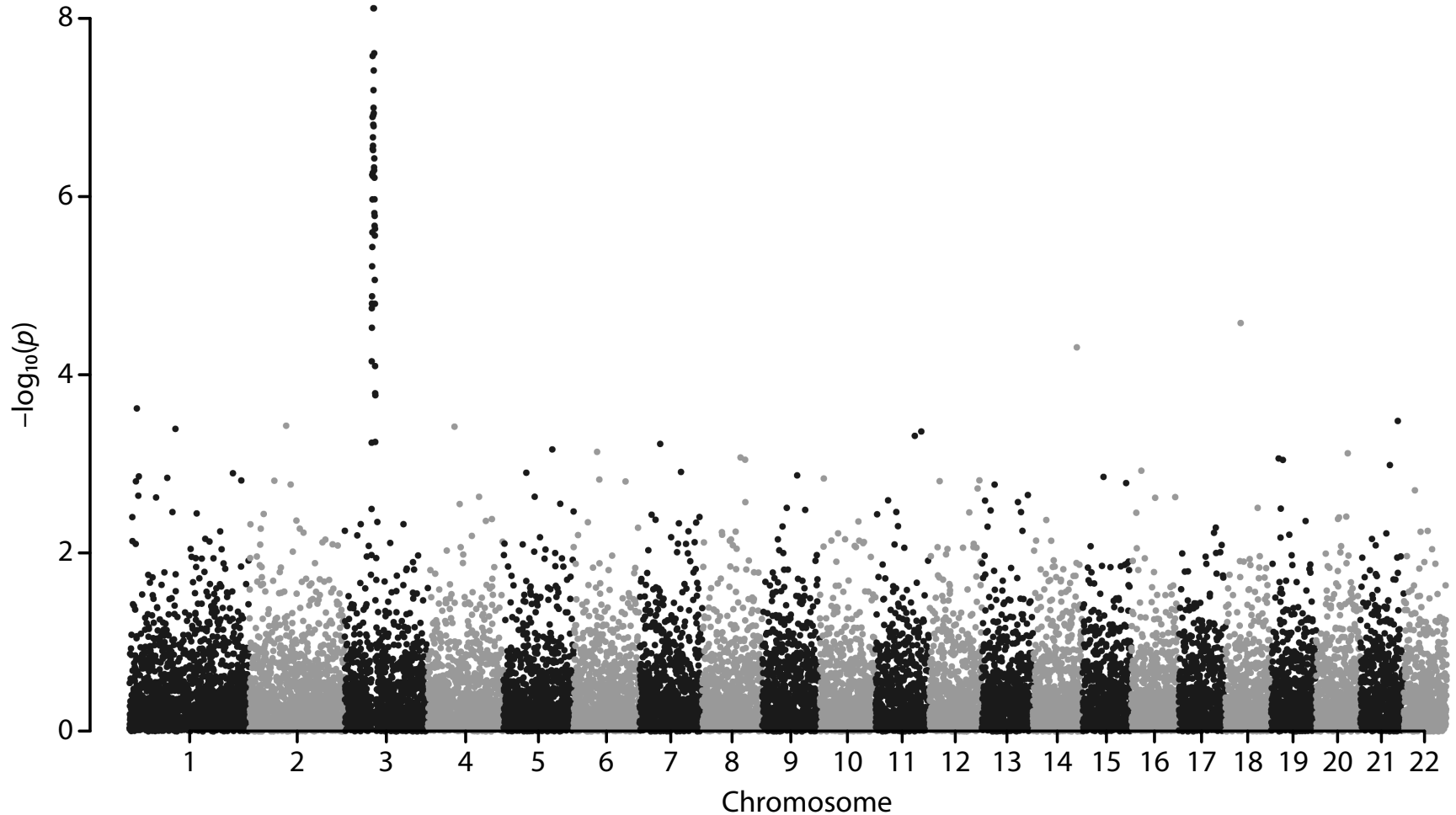
is called p -value

1. Set α (by the user)

- $\alpha = 0.05$ or 0.01 are commonly used

- 
- ## 2. If the p -value is smaller than α , we can conclude that the DNA is “associated” with the disease
- This procedure is called **hypothesis testing**
 - It controls the **risk (false positives)** under α
 - Intuitively, we measure the importance of the DNA by the p -value in a statistical manner

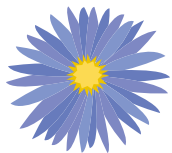
Example of DNA Data (Manhattan plot)



Next, let us find base (DNA) pairs

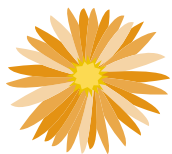
Genome sequence (SNPs)

Case
(disease)



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

Control
(health)



Samples	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

A pair of SNPs that are similar in a class 0: Normal, 1: Rare

It is difficult to check “all” pairs

- No problem if the number of bases is small
- It is impossible if the number of bases is large

- For example, if there are one million bases, the number of pairs becomes

$$\binom{1000000}{2} = \frac{1000000 \cdot 999999}{2} \approx 10^{12} !!$$

- If the samples size is 1000, 10^{15} checks are required!!
- **Solution:** Use the power of “statistics” and “algorithms”
 - The lightbulb algorithm
 - Paturi et al. (1995), Achlioptas et al. (2011)

Find Similar Pairs

	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. Random Sampling (Statistics)

	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. Random Sampling (Statistics)

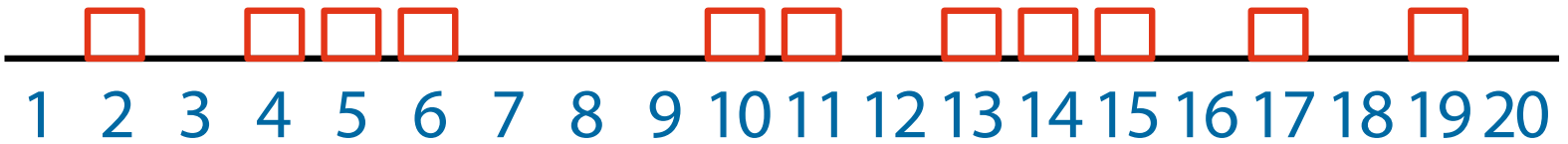
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3:	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1
4:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
5:	1	0	0	0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0
7:	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1
9:	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0

2. Sorting (Radix Sort; Algorithm)

	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

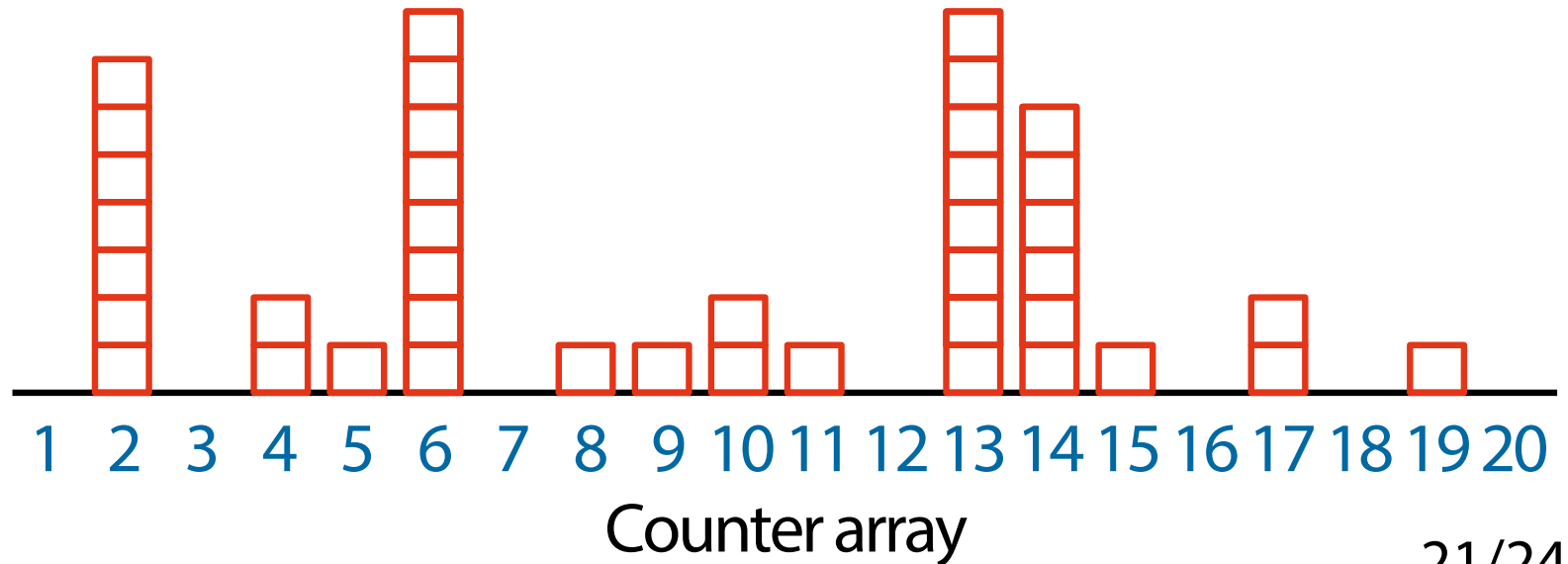
3. Count Exactly the Same Pairs

	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

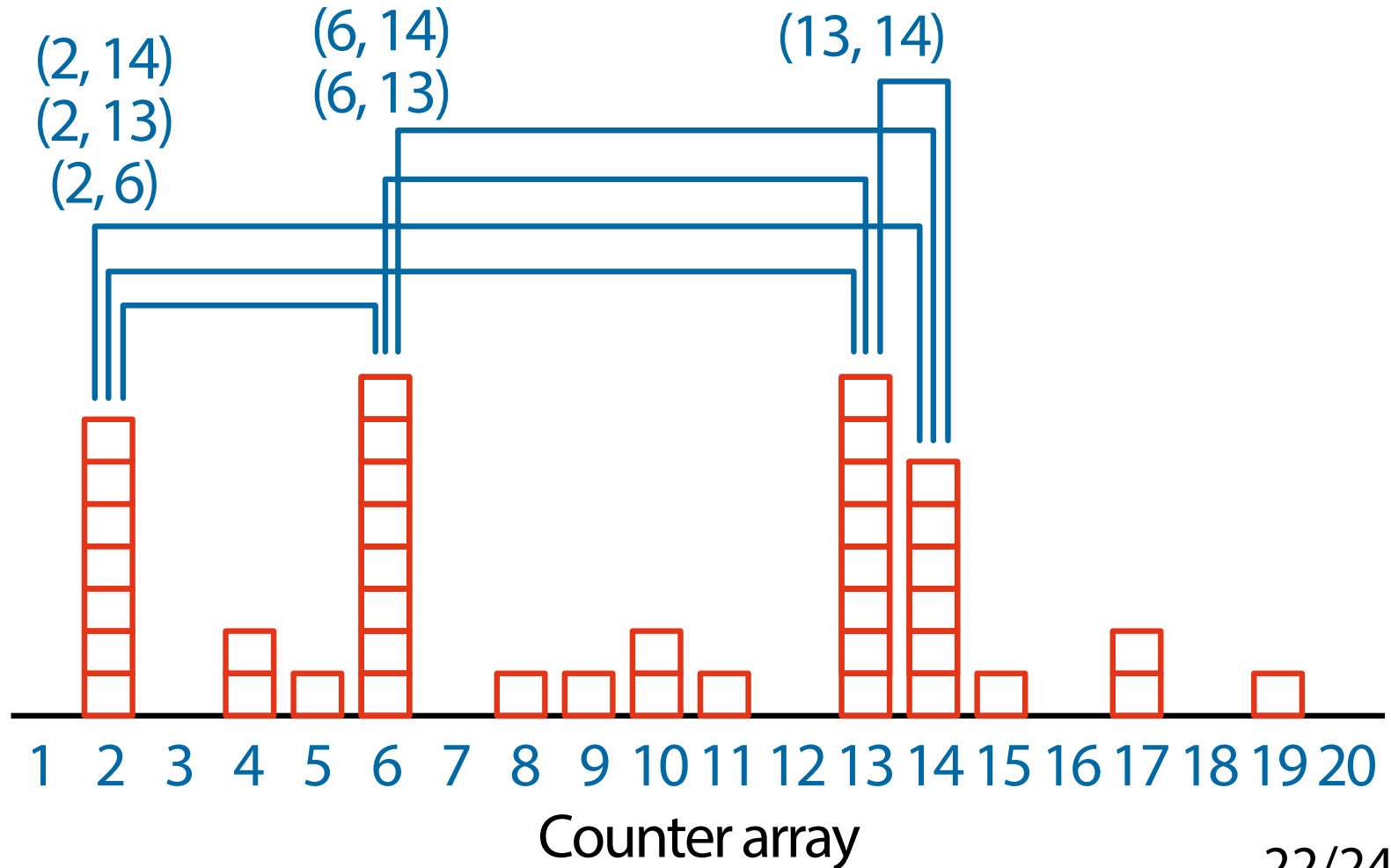


Counter array

4. Repeat the above Process



5. Compare Pairs with Large Values



We can find similar pairs

	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

Summary

- Two topics with statistical data analysis
 1. How to guarantee the correctness
 - Control false positives by the p -value
 - Hypothesis testing
 2. How to achieve efficient computation
 - Integrate “statistics” and “algorithms”