Hypothesis Testing and Statistically-sound Pattern Mining

Tutorial - KDD 2019

Leonardo Pellegrina[1] Matteo Riondato[2] Fabio Vandin[1]

[1]Dept. of Information Engineering, University of Padova (IT)

[2]Dept. of Computer Science, Amherst College (USA)

Slides available from `http://rionda.to/statdmtut`

Outline

1. **Introduction and Theoretical Foundations**
   1.1 **Introduction to Significant Pattern Mining**
   1.2 Statistical Hypothesis Testing
   1.3 Fundamental Tests
   1.4 Multiple Hypothesis Testing
   1.5 Selecting Hypothesis
   1.6 Hypotheses Testability
2. Mining Statistically-Sound Patterns
3. Recent developments and advanced topics
4. Final Remarks

Introduction

Data mining and (inferential) *statistics* have traditionally **two different point of views**

Introduction

Data mining and (inferential) *statistics* have traditionally **two different point of views**

- *data mining*: the data is the **complete representation of the world and of the phenomena** we are studying

Introduction

Data mining and (inferential) *statistics* have traditionally **two different point of views**

- *data mining*: the data is the **complete representation of the world and of the phenomena** we are studying
- *statistics*: the data is obtained from an **underlying generative process**, that is what we really care about

Data mining and (inferential) *statistics* have traditionally **two different point of views**

- *data mining*: the data is the **complete representation of the world and of the phenomena** we are studying
- *statistics*: the data is obtained from an **underlying generative process**, that is what we really care about

Similar questions but **different flavours**!

Example

Data: information from two online communities C_1 and C_2, regarding whether each post is in a given topic T.

Example

Data: information from two online communities C_1 and C_2, regarding whether each post is in a given topic T.

- Data mining: "what fraction of posts in C_1 are related to T? What fraction of posts in C_2 are related to T?"

Example

Data: information from two online communities C_1 and C_2, regarding whether each post is in a given topic T.

- Data mining: "what fraction of posts in C_1 are related to T? What fraction of posts in C_2 are related to T?"
- Statistics: "What is the probability that a post from C_1 is related to T? What is the probability that a post from C_2 is related to T?"

Example

Data: information from two online communities C_1 and C_2, regarding whether each post is in a given topic T.

- Data mining: "what fraction of posts in C_1 are related to T? What fraction of posts in C_2 are related to T?"
- Statistics: "What is the probability that a post from C_1 is related to T? What is the probability that a post from C_2 is related to T?"

Note: the two are **clearly related, but different**!

How do we **efficiently** identify patterns in data with **guarantees** on the **underlying generative process**?

Statistically-Sound Pattern Mining

How do we **efficiently** identify patterns in data with **guarantees** on the **underlying generative process**?

We use the **statistical hypothesis testing** framework

Outline

Statistical Hypothesis Testing

We are given:

- a **dataset** $\mathcal{D}$
- a **question** we want to answer

Statistical Hypothesis Testing

We are given:

- a **dataset** $\mathcal{D}$
- a **question** we want to answer $\Rightarrow$ a **pattern** $\mathcal{S}$

Statistical Hypothesis Testing

We are given:

- a **dataset** $\mathcal{D}$
- a **question** we want to answer $\Rightarrow$ a **pattern** $\mathcal{S}$

EXAMPLE

- $\mathcal{D} =$ for 1000 diseased individuals (*cases*), whether drug $\mathcal{S}$ had an effect (YES/NO); for 1000 healthy individuals (*controls*), whether drug $\mathcal{S}$ had an effect (YES/NO).

Statistical Hypothesis Testing

We are given:

- a **dataset** $\mathcal{D}$
- a **question** we want to answer $\Rightarrow$ a **pattern** $\mathcal{S}$

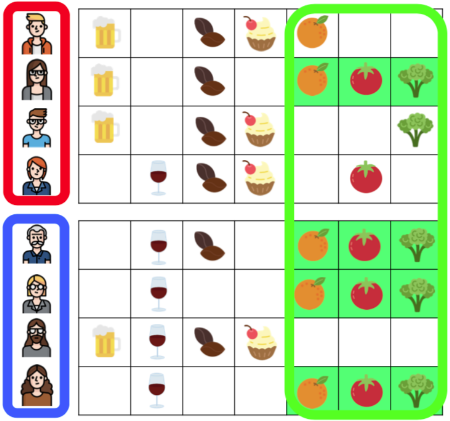
EXAMPLE

- $\mathcal{D} =$ for 1000 diseased individuals (*cases*), whether drug $\mathcal{S}$ had an effect (YES/NO); for 1000 healthy individuals (*controls*), whether drug $\mathcal{S}$ had an effect (YES/NO).
- does $\mathcal{S}$ have the same effect on diseased individuals (*cases*) and on healthy individuals (*controls*)?

Example: market basket analysis

Dataset $\mathcal{D}$: transactions = set of items, label (student/professor)

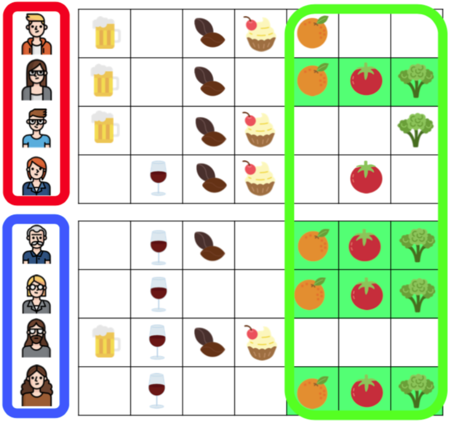
Pattern $\mathcal{S}$: subset of items (orange, tomato, broccoli)



Example: market basket analysis

Dataset $\mathcal{D}$: transactions = set of items, label (student/professor)

Pattern $\mathcal{S}$: subset of items (orange, tomato, broccoli)



Question: is $\mathcal{S}$ associated with one of the two labels?

Statistical Hypothesis Testing: Formalization

Frame the question in terms of a **null hypothesis**, describing the *default theory*, which corresponds to “nothing interesting” for pattern $\mathcal{S}$.

Statistical Hypothesis Testing: Formalization

Frame the question in terms of a **null hypothesis**, describing the *default theory*, which corresponds to “nothing interesting” for pattern $\mathcal{S}$.

The goal is to use the data to either **reject** H_0 (“$\mathcal{S}$ is interesting!”) **or not** (“$\mathcal{S}$ is not interesting).

Statistical Hypothesis Testing: Formalization

Frame the question in terms of a **null hypothesis**, describing the *default theory*, which corresponds to “nothing interesting” for pattern $\mathcal{S}$.

The goal is to use the data to either **reject** H_0 (“$\mathcal{S}$ is interesting!”) **or not** (“$\mathcal{S}$ is not interesting).

This is decided based on a **test statistic**, that is, a value $x_S = f_S(\mathcal{D})$ that describes $\mathcal{S}$ in $\mathcal{D}$

Statistical Hypothesis Testing: p-value

Let $x_S = f(\mathcal{D})$ the value of the *test statistic* for our dataset $\mathcal{D}$.

Statistical Hypothesis Testing: p-value

Let $x_S = f(\mathcal{D})$ the value of the *test statistic* for our dataset $\mathcal{D}$.

Let X_S be the *random variable* describing the value of the test statistic **under the null hypothesis** H_0 (i.e., when H_0 is true)

Statistical Hypothesis Testing: p-value

Let $x_S = f(\mathcal{D})$ the value of the *test statistic* for our dataset $\mathcal{D}$.

Let X_S be the *random variable* describing the value of the test statistic **under the null hypothesis** H_0 (i.e., when H_0 is true)

p-**value**: $p = \Pr[X_S \text{ more extreme than } x_S : H_0 \text{ is true}]$

Statistical Hypothesis Testing: p-value

Let $x_S = f(\mathcal{D})$ the value of the *test statistic* for our dataset $\mathcal{D}$.

Let X_S be the *random variable* describing the value of the test statistic **under the null hypothesis** H_0 (i.e., when H_0 is true)

p-**value**: $p = \Pr[X_S \text{ more extreme than } x_S : H_0 \text{ is true}]$

"X_S more extreme than x_S": depends on the test, may be $X_S \geqslant x_S$ or $X_S \leqslant x_S$ or something else...

Statistical Hypothesis Testing: p-value

Let $x_S = f(\mathcal{D})$ the value of the *test statistic* for our dataset $\mathcal{D}$.

Let X_S be the *random variable* describing the value of the test statistic **under the null hypothesis** H_0 (i.e., when H_0 is true)

p-**value**: $p = \Pr[X_S \text{ more extreme than } x_S : H_0 \text{ is true}]$

"X_S more extreme than x_S": depends on the test, may be $X_S \geqslant x_S$ or $X_S \leqslant x_S$ or something else...

Rejection rule:
Given a *statistical level* $\alpha \in (0, 1)$: **reject** H_0 iff $p \leqslant \alpha \Rightarrow \mathcal{S}$ **is significant**!

Statistical Hypothesis Testing: Errors

There are **two types of errors** we can make:

Statistical Hypothesis Testing: Errors

There are **two types of errors** we can make:

- **type I error**: reject H_0 when H_0 is true $\Rightarrow$ flag S as significant when it is not (*false discovery*)

Statistical Hypothesis Testing: Errors

There are **two types of errors** we can make:

- **type I error**: reject H_0 when H_0 is true $\Rightarrow$ flag S as significant when it is not (*false discovery*)
- **type II error**: do not reject H_0 when H_0 is false $\Rightarrow$ do not flag S as significant when it is

Statistical Hypothesis Testing: Errors

There are **two types of errors** we can make:

- **type I error**: reject H_0 when H_0 is true $\Rightarrow$ flag S as significant when it is not (*false discovery*)
- **type II error**: do not reject H_0 when H_0 is false $\Rightarrow$ do not flag S as significant when it is

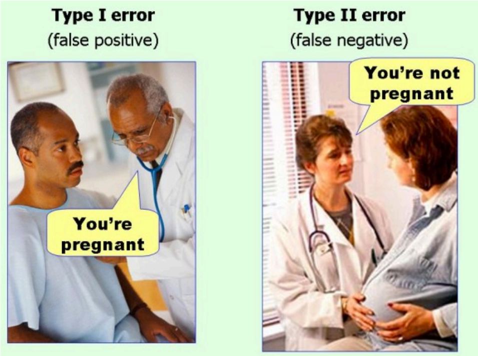
		REALITY	
		H_0 false	H_0 true
DECISION	reject H_0	Correct!	Type I error
	accept H_0	Type II error	Correct!

Statistical Hypothesis Testing: Errors

There are **two types of errors** we can make:

- **type I error**: reject H_0 when H_0 is true $\Rightarrow$ flag S as significant when it is not (*false discovery*)
- **type II error**: do not reject H_0 when H_0 is false $\Rightarrow$ do not flag S as significant when it is

		REALITY	
		H_0 false	H_0 true
DECISION	reject H_0	Correct!	Type I error
	accept H_0	Type II error	Correct!



Statistical Hypothesis Testing: Error Guarantees

There are **two types of errors** we can make:

- **type I error**: reject H_0 when H_0 is true $\Rightarrow$ flag S as significant when it is not (*false discovery*)
- **type II error**: do not reject H_0 when H_0 is false $\Rightarrow$ do not flag S as significant when it is

Statistical Hypothesis Testing: Error Guarantees

There are **two types of errors** we can make:

- **type I error**: reject H_0 when H_0 is true $\Rightarrow$ flag S as significant when it is not (*false discovery*)
- **type II error**: do not reject H_0 when H_0 is false $\Rightarrow$ do not flag S as significant when it is

Theorem

Using the **rejection rule**, *the probability of a type I error is* $\leqslant \alpha$

Avoiding type I errors is not everything!

Statistical Hypothesis Testing: Power

Avoiding type I errors is not everything!

If it was, it would be enough to *never* flag a pattern as significant. . .

Statistical Hypothesis Testing: Power

Avoiding type I errors is not everything!

If it was, it would be enough to *never* flag a pattern as significant...

Power:
A test has *power* β if $\Pr[H_0 \text{ is rejected} : H_0 \text{ is false}] = \beta$

Statistical Hypothesis Testing: Power

Avoiding type I errors is not everything!

If it was, it would be enough to *never* flag a pattern as significant. . .

Power:
A test has *power* β if $\Pr[H_0 \text{ is rejected} : H_0 \text{ is false}] = \beta$

Note: for a test with power β, we have $\Pr[\text{type II error}] = 1 - \beta$

Avoiding type I errors is not everything!

If it was, it would be enough to *never* flag a pattern as significant...

Power:
A test has *power* β if $\Pr[H_0 \text{ is rejected} : H_0 \text{ is false}] = \beta$

Note: for a test with power β, we have $\Pr[\text{type II error}] = 1 - \beta$

(Power is not everything: if it was, it would be enough to *always* flag all patterns as significant...)

Example: Testing for Independence

Given:

- transactional dataset $\mathcal{D} = \{t_1, \ldots, t_n\}$, each transaction t_i has a label $\ell(t_i) \in \{c_0, c_1\}$
- a pattern S

Example: Testing for Independence

Given:

- transactional dataset $\mathcal{D} = \{t_1, \ldots, t_n\}$, each transaction t_i has a label $\ell(t_i) \in \{c_0, c_1\}$
- a pattern S

Goal: understand if the appearance of S in transactions $(\mathcal{S} \subseteq t_i)$ and the transactions labels $(\ell(t_i))$ are *independent*.

Example: Testing for Independence

Given:

- transactional dataset $\mathcal{D} = \{t_1, \ldots, t_n\}$, each transaction t_i has a label $\ell(t_i) \in \{c_0, c_1\}$
- a pattern S

Goal: understand if the appearance of S in transactions ($\mathcal{S} \subseteq t_i$) and the transactions labels ($\ell(t_i)$) are *independent.*

Null hypothesis H_0: the events "$\mathcal{S} \subseteq t_i$" and "$\ell(t_i) = c_1$" are independent.

Example: Testing for Independence

Given:

- transactional dataset $\mathcal{D} = \{t_1, \ldots, t_n\}$, each transaction t_i has a label $\ell(t_i) \in \{c_0, c_1\}$
- a pattern S

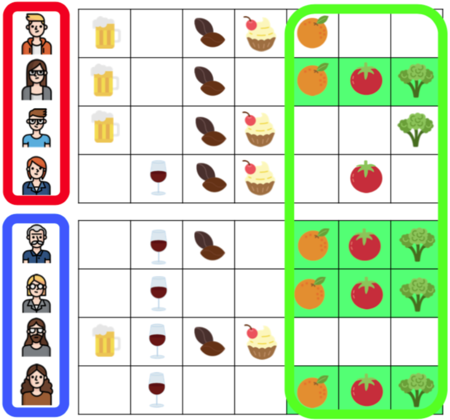
Goal: understand if the appearance of S in transactions ($\mathcal{S} \subseteq t_i$) and the transactions labels ($\ell(t_i)$) are *independent.*

Null hypothesis H_0: the events "$\mathcal{S} \subseteq t_i$" and "$\ell(t_i) = c_1$" are independent.

Alternative hypothesis: there is a dependency between "$\mathcal{S} \subseteq t_i$" and "$\ell(t_i) = c_1$"

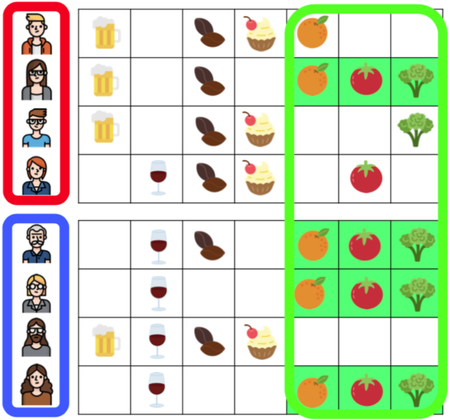
Example: market basket analysis

$\mathcal{S} = \{\text{orange, tomato, broccoli}\}$



Example: market basket analysis

$\mathcal{S}$ = {orange, tomato, broccoli}



H_0: presence of $\mathcal{S}$ is independent of (not associated with) label "professor"

Example: Testing for Independence (2)

Useful representation of the data: *contingency table*

Example: Testing for Independence (2)

Useful representation of the data: *contingency table*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Example: Testing for Independence (2)

Useful representation of the data: *contingency table*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

- $\sigma_1(\mathcal{S}) =$ number of transactions containing $\mathcal{S}$ ($=$*support* of $\mathcal{S}$) with label c_1

Example: Testing for Independence (2)

Useful representation of the data: *contingency table*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

- $\sigma_1(\mathcal{S})$ = number of transactions containing $\mathcal{S}$ (=*support* of $\mathcal{S}$) with label c_1
- $\sigma_0(\mathcal{S})$ = support of $\mathcal{S}$ with label c_0

Example: Testing for Independence (2)

Useful representation of the data: *contingency table*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

- $\sigma_1(\mathcal{S})$ = number of transactions containing $\mathcal{S}$ (=*support* of $\mathcal{S}$) with label c_1
- $\sigma_0(\mathcal{S})$ = support of $\mathcal{S}$ with label c_0
- $\sigma(\mathcal{S}) = \sigma_0(\mathcal{S}) + \sigma_1(\mathcal{S})$ = support of $\mathcal{S}$ in $\mathcal{D}$

Example: Testing for Independence (2)

Useful representation of the data: *contingency table*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

- $\sigma_1(\mathcal{S})$ = number of transactions containing $\mathcal{S}$ (=*support* of $\mathcal{S}$) with label c_1
- $\sigma_0(\mathcal{S})$ = support of $\mathcal{S}$ with label c_0
- $\sigma(\mathcal{S}) = \sigma_0(\mathcal{S}) + \sigma_1(\mathcal{S})$ = support of $\mathcal{S}$ in $\mathcal{D}$
- n_i = number transactions with label c_i

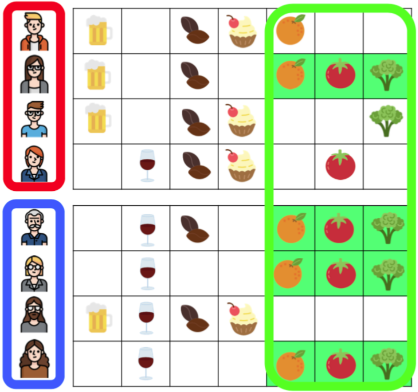
Example: Testing for Independence (3)

Useful representation of the data: *contingency table*

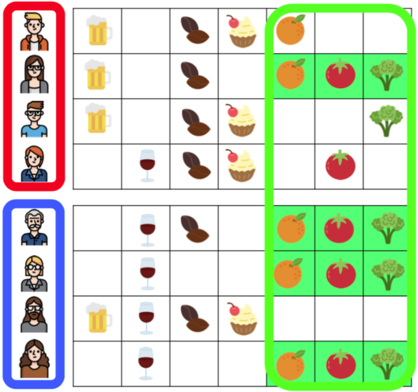
	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Test statistic $= \sigma_1(S)$

Example: market basket analysis

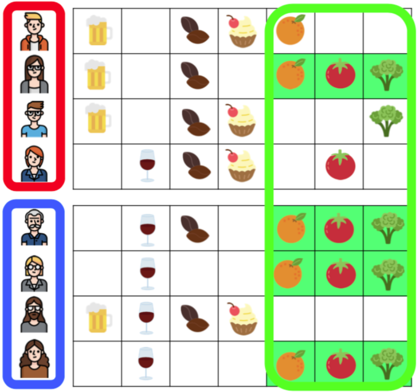


Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

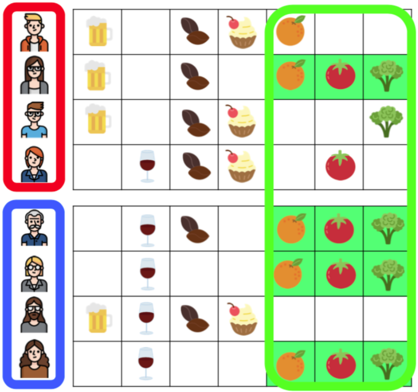
Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

Value of test statistic $= \sigma_1(\mathcal{S})$

Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

Value of test statistic $= \sigma_1(\mathcal{S}) = 3$

Example: Testing for Independence (3)

Useful representation of the data: *contingency table*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Test statistic $= \sigma_1(S)$

Example: Testing for Independence (3)

Useful representation of the data: *contingency table*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Test statistic $= \sigma_1(S)$

p-value: **how do we compute it**?

Example: Testing for Independence (3)

Useful representation of the data: *contingency table*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Test statistic $= \sigma_1(S)$

p-value: **how do we compute it**?

Most common method: **Fisher's exact test**

Outline

Fisher's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Fisher's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assumption: the column marginals ($\sigma(S)$, $n - \sigma(S)$ *and* the row marginals (n_0, n_1) are **fixed**.

Fisher's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assumption: the column marginals ($\sigma(S)$, $n - \sigma(S)$ *and* the row marginals (n_0, n_1) are **fixed**.

$\Rightarrow$ under the null hypothesis (*independence*), the support of S in class c_1 follows an hypergeometric distribution of parameters n, n_1, and $\sigma_{\mathcal{S}}$

Fisher's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assumption: the column marginals ($\sigma(S)$, $n - \sigma(S)$ *and* the row marginals (n_0, n_1) are **fixed**.

$\Rightarrow$ under the null hypothesis (*independence*), the support of S in class c_1 follows an hypergeometric distribution of parameters n, n_1, and $\sigma_{\mathcal{S}}$

$\Rightarrow$ the p-value is **easily computable!**

Fisher's exact test(2)

Let $X_{\mathcal{S}}$ be the r.v. describing the support of $\mathcal{S}$ in class c_1 when the null hypothesis holds

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Fisher's exact test(2)

Let $X_{\mathcal{S}}$ be the r.v. describing the support of $\mathcal{S}$ in class c_1 when the null hypothesis holds

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

$$\Pr(X_{\mathcal{S}} = k) = \frac{\binom{n_1}{k}\binom{n_0}{\sigma(\mathcal{S})-k}}{\binom{n}{\sigma(\mathcal{S})}}$$

Fisher's exact test(2)

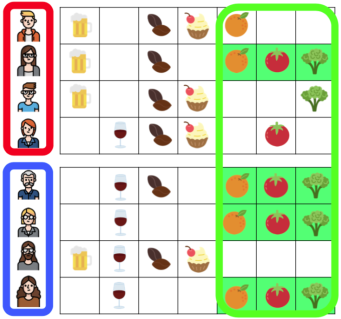
Let $X_{\mathcal{S}}$ be the r.v. describing the support of $\mathcal{S}$ in class c_1 when the null hypothesis holds

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

$$\Pr(X_{\mathcal{S}} = k) = \frac{\binom{n_1}{k}\binom{n_0}{\sigma(\mathcal{S})-k}}{\binom{n}{\sigma(\mathcal{S})}}$$

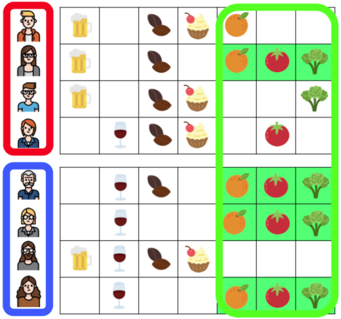
p-value for $\mathcal{S}$: $p_{\mathcal{S}} = \sum_{k \geqslant \sigma_1(\mathcal{S})} \Pr(X_{\mathcal{S}} = k)$

Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

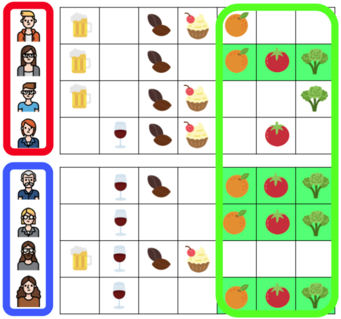
Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim$ hypergeometric of parameters 8, 4, 3

Example: market basket analysis

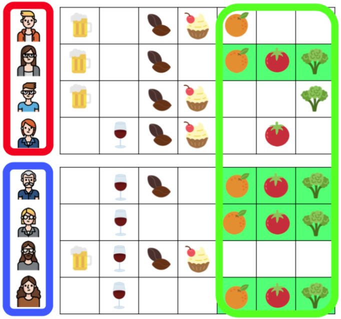


	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim$ hypergeometric of parameters 8, 4, 3

$\Rightarrow$ Probability of table $= \Pr(X_{\mathcal{S}} = 3) = 0.228$

Example: market basket analysis

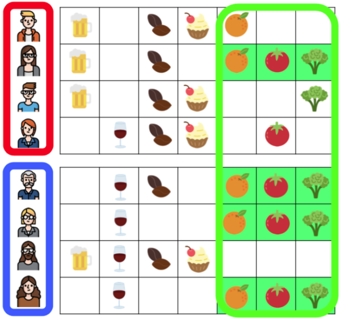


	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim$ hypergeometric of parameters 8, 4, 3
$\Rightarrow$ Probability of table $= \Pr(X_{\mathcal{S}} = 3) = 0.228$

p-value $= \Pr(X_{\mathcal{S}} \geqslant 3) = 0.243$

Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim$ hypergeometric of parameters 8, 4, 3
$\Rightarrow$ Probability of table $= \Pr(X_{\mathcal{S}} = 3) = 0.228$

p-value $= \Pr(X_{\mathcal{S}} \geqslant 3) = 0.243$

If $\alpha = 0.05 \Rightarrow \mathcal{S}$ is not associated with label "professor"

χ^2 test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

In the old days: "Fisher's exact test is computationally expensive..." 

χ^2 test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

In the old days: "Fisher's exact test is computationally expensive..." 

Random variables (r.v.) describing outcome under H_0 (H_0 is true)

χ^2 test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

In the old days: "Fisher's exact test is computationally expensive..." 

Random variables (r.v.) describing outcome under H_0 (H_0 is true)

- $X_{\mathcal{S},0}$ = r.v. describing the support of $\mathcal{S}$ in class c_0

χ^2 test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

In the old days: "Fisher's exact test is computationally expensive..." 

Random variables (r.v.) describing outcome under H_0 (H_0 is true)

- $X_{\mathcal{S},0}$ = r.v. describing the support of $\mathcal{S}$ in class c_0
- $X_{\mathcal{S},1}$ = r.v. describing the support $\mathcal{S}$ in class c_1

χ^2 test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

In the old days: "Fisher's exact test is computationally expensive..." 

Random variables (r.v.) describing outcome under H_0 (H_0 is true)

- $X_{\mathcal{S},0}$ = r.v. describing the support of $\mathcal{S}$ in class c_0
- $X_{\mathcal{S},1}$ = r.v. describing the support $\mathcal{S}$ in class c_1
- $X_{\bar{\mathcal{S}},0}$ = r.v. describing num. transactions without $\mathcal{S}$ in class c_0

χ^2 test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

In the old days: "Fisher's exact test is computationally expensive..." 

Random variables (r.v.) describing outcome under H_0 (H_0 is true)

- $X_{\mathcal{S},0}$ = r.v. describing the support of $\mathcal{S}$ in class c_0
- $X_{\mathcal{S},1}$ = r.v. describing the support $\mathcal{S}$ in class c_1
- $X_{\bar{\mathcal{S}},0}$ = r.v. describing num. transactions without $\mathcal{S}$ in class c_0
- $X_{\bar{\mathcal{S}},1}$ = r.v. describing num. transactions without $\mathcal{S}$ in class c_1

χ^2 test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

In the old days: "Fisher's exact test is computationally expensive..." 

Random variables (r.v.) describing outcome under H_0 (H_0 is true)

- $X_{\mathcal{S},0}$ = r.v. describing the support of $\mathcal{S}$ in class c_0
- $X_{\mathcal{S},1}$ = r.v. describing the support $\mathcal{S}$ in class c_1
- $X_{\bar{\mathcal{S}},0}$ = r.v. describing num. transactions without $\mathcal{S}$ in class c_0
- $X_{\bar{\mathcal{S}},1}$ = r.v. describing num. transactions without $\mathcal{S}$ in class c_1

Test statistic: $X = \sum_{i \in \{\mathcal{S},\bar{\mathcal{S}}\}, j \in \{0,1\}} (X_{i,j} - \mathbb{E}[X_{i,j}])^2 / \mathbb{E}[X_{i,j}]$

χ^2 test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

In the old days: "Fisher's exact test is computationally expensive…" 

Random variables (r.v.) describing outcome under H_0 (H_0 is true)

- $X_{\mathcal{S},0}$ = r.v. describing the support of $\mathcal{S}$ in class c_0
- $X_{\mathcal{S},1}$ = r.v. describing the support $\mathcal{S}$ in class c_1
- $X_{\bar{\mathcal{S}},0}$ = r.v. describing num. transactions without $\mathcal{S}$ in class c_0
- $X_{\bar{\mathcal{S}},1}$ = r.v. describing num. transactions without $\mathcal{S}$ in class c_1

Test statistic: $X = \sum_{i\in\{\mathcal{S},\bar{\mathcal{S}}\},j\in\{0,1\}} (X_{i,j} - \mathbb{E}[X_{i,j}])^2/\mathbb{E}[X_{i,j}]$

Note: $\mathbb{E}[X_{i,j}]$ are easily computable

χ^2 test

Theorem

When $n \to +\infty$, $X \to \chi^2$ distribution with 1 degree of freedom

Theorem

When $n \to +\infty$, $X \to \chi^2$ *distribution with 1 degree of freedom*

Why is this important? There are *tables* to compute probabilities for the χ^2 distribution

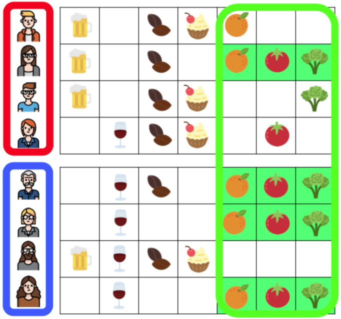
Theorem

When $n \to +\infty$*,* $X \to \chi^2$ *distribution with 1 degree of freedom*

Why is this important? There are *tables* to compute probabilities for the χ^2 distribution

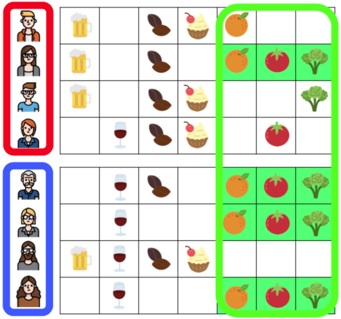
Note: the χ^2 test is the *asymptotic* version of Fisher's exact test.

Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

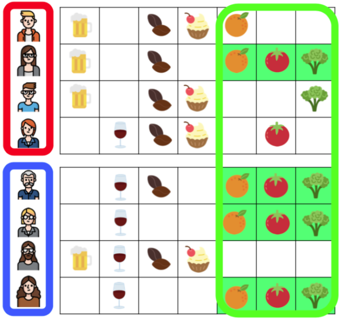
Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim \chi^2$ with 1 degree of freedom

Example: market basket analysis

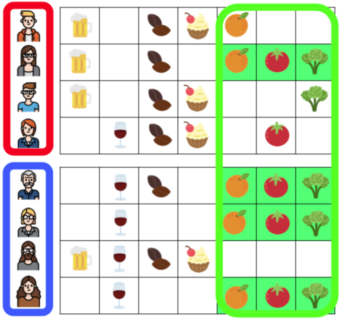


	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim \chi^2$ with 1 degree of freedom

Test statistic: 2

Example: market basket analysis



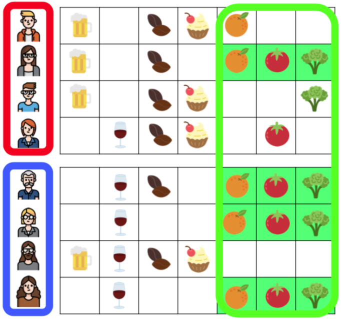
	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim \chi^2$ with 1 degree of freedom

Test statistic: 2

p-value $= \Pr(X_{\mathcal{S}} \geqslant 2) = 0.16$

Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim \chi^2$ with 1 degree of freedom

Test statistic: 2

p-value $= \Pr(X_{\mathcal{S}} \geqslant 2) = 0.16$

If $\alpha = 0.05 \Rightarrow \mathcal{S}$ is not associated with label "professor"

Barnard's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assumption: the row marginals (n_0, n_1) are fixed

Barnard's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assumption: the row marginals (n_0, n_1) are fixed **but the column marginals ($\sigma(S)$, $n - \sigma(S)$) are not!**

Barnard's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assumption: the row marginals (n_0, n_1) are fixed **but the column marginals ($\sigma(S)$, $n - \sigma(S)$) are not!**

$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_0] = \pi_0$$
$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_1] = \pi_1$$

Barnard's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assumption: the row marginals (n_0, n_1) are fixed **but the column marginals ($\sigma(S)$, $n - \sigma(S)$) are not!**

$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_0] = \pi_0$$
$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_1] = \pi_1$$

Null hypothesis H_0: $\pi_0 = \pi_1 = \pi$

Barnard's exact test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assumption: the row marginals (n_0, n_1) are fixed **but the column marginals ($\sigma(S)$, $n - \sigma(S)$) are not!**

$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_0] = \pi_0$$
$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_1] = \pi_1$$

Null hypothesis H_0: $\pi_0 = \pi_1 = \pi$

π is *nuisance parameter*, in the sense that we are not interested in its value, but its value *defines* the distribution of our observations

Bernard's exact test(2)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_0] = \pi_0$$
$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_1] = \pi_1$$

Null hypothesis H_0: $\pi_0 = \pi_1 = \pi$

Bernard's exact test(2)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_0] = \pi_0$$
$$\Pr[\mathcal{S} \subseteq t_i : \ell(t_i) = c_1] = \pi_1$$

Null hypothesis H_0: $\pi_0 = \pi_1 = \pi$

How do we compute the p-value?

Bernard's exact test(3)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assuming π is known, the probability depends only on

- X = r.v. describing the support of $\mathcal{S}$
- Y = r.v. describing the support $\mathcal{S}$ in class c_1

Bernard's exact test(3)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assuming π is known, the probability depends only on

- X = r.v. describing the support of $\mathcal{S}$
- Y = r.v. describing the support $\mathcal{S}$ in class c_1

Let x the observed value of X and y the observed value of Y

$$P(x, y|\pi) = \binom{n_0}{x-y}\binom{n_1}{y}(\pi)^x(1-\pi)^{n-x)}$$

Bernard's exact test(3)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Assuming π is known, the probability depends only on

- X = r.v. describing the support of $\mathcal{S}$
- Y = r.v. describing the support $\mathcal{S}$ in class c_1

Let x the observed value of X and y the observed value of Y

$$P(x, y|\pi) = \binom{n_0}{x-y}\binom{n_1}{y}(\pi)^x (1-\pi)^{n-x)}$$

Test statistic: **probability of the contingency table**.

Bernard's exact test(4)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Let x the observed value of X and y the observed value of Y

$$\Pr(x, y|\pi) = \binom{n_0}{x-y}\binom{n_1}{y}(\pi)^x(1-\pi)^{n-x)}$$

Bernard's exact test(4)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Let x the observed value of X and y the observed value of Y

$$\Pr(x, y|\pi) = \binom{n_0}{x-y}\binom{n_1}{y}(\pi)^x (1-\pi)^{n-x)}$$

Let $T(x,y)$= set of *more extreme tables* for a given π

$$T(x, y, \pi) = \{(x', y') : \Pr(x', y' \mid \pi) \leqslant \Pr(x, y|\pi)\}$$

Bernard's exact test(4)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Let x the observed value of X and y the observed value of Y

$$\Pr(x, y|\pi) = \binom{n_0}{x-y}\binom{n_1}{y}(\pi)^x(1-\pi)^{n-x)}$$

Let $T(x, y)$= set of *more extreme tables* for a given π

$$T(x, y, \pi) = \{(x', y') : \Pr(x', y' \mid \pi) \leqslant \Pr(x, y|\pi)\}$$

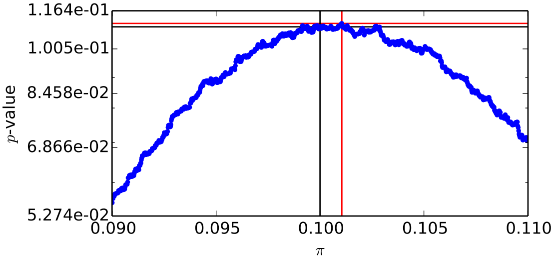
Then p-value: $p = \max_{\pi \in (0,1)} \sum_{(x,y) \in T(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}), \pi)} \Pr(x, y|\pi)$

Barnard's exact test(5)

p-value: $p = \max_{\pi \in (0,1)} \sum_{(x,y) \in T(\sigma(\mathcal{S}),\sigma_1(\mathcal{S}),\pi)} \Pr(x, y | \pi)$

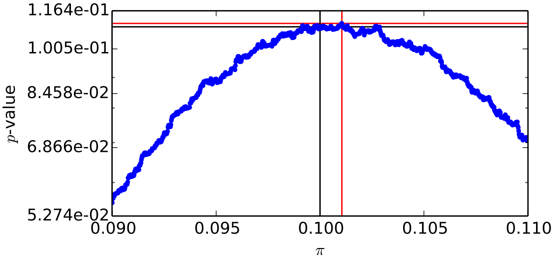
Barnard's exact test(5)

$$p\text{-value: } p = \max_{\pi\in(0,1)} \sum_{(x,y)\in T(\sigma(\mathcal{S}),\sigma_1(\mathcal{S}),\pi)} \Pr(x,y|\pi)$$



Barnard's exact test(5)

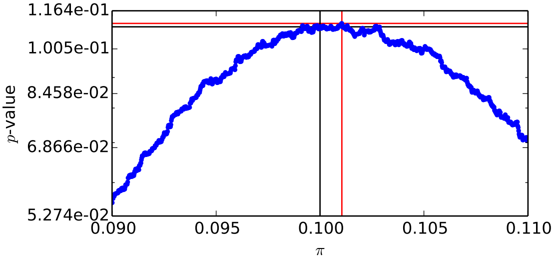
$$p\text{-value: } p = \max_{\pi \in (0,1)} \sum_{(x,y) \in T(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}), \pi)} \Pr(x, y | \pi)$$



Computing the p-value is computationally expensive!

Barnard's exact test(5)

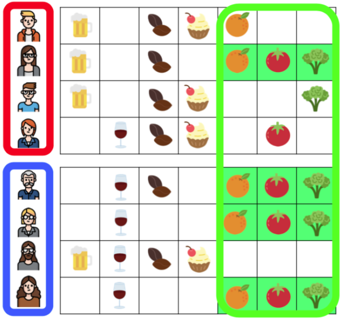
p-value: $p = \max_{\pi \in (0,1)} \sum_{(x,y) \in T(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}), \pi)} \Pr(x, y | \pi)$



Computing the p-value is computationally expensive!

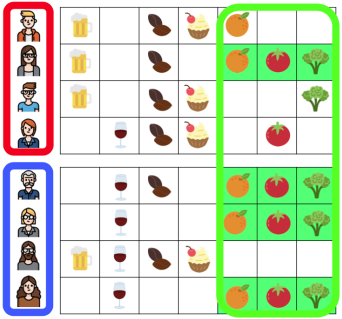
- consider a grid of value for π
- enumerate all tables in $T(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}), \pi)$

Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

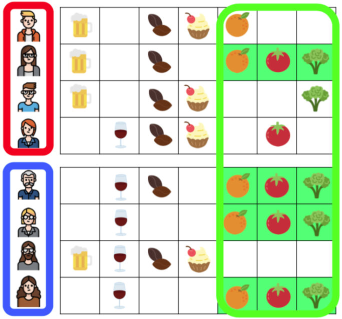
Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$$\Pr(4, 3|\pi) = \binom{4}{1}\binom{4}{3}(\pi)^4(1-\pi)^4$$

Example: market basket analysis

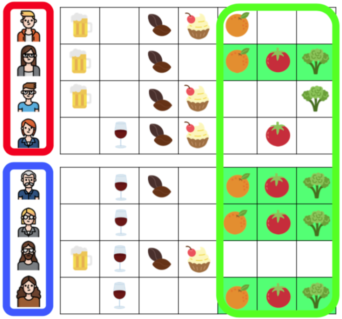


	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$$\Pr(4, 3|\pi) = \binom{4}{1}\binom{4}{3}(\pi)^4(1-\pi)^4$$

$$T(x, y, \pi) = \{(x', y') : \Pr(x', y' \mid \pi) \leqslant \Pr(4, 3|\pi)\}$$

Example: market basket analysis



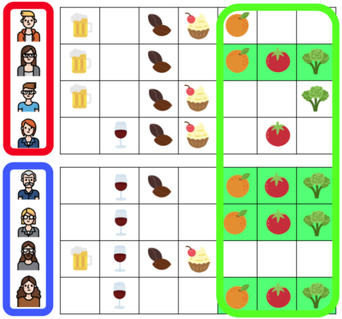
	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$$\Pr(4,3|\pi) = \binom{4}{1}\binom{4}{3}(\pi)^4(1-\pi)^4$$

$$T(x,y,\pi) = \{(x',y') : \Pr(x',y' \mid \pi) \leqslant \Pr(4,3|\pi)\}$$

p-value: $\max_{\pi\in(0,1)} \sum_{(x,y)\in T(\sigma(\mathcal{S}),\sigma_1(\mathcal{S}),\pi)} \Pr(x,y|\pi)$

Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$$\Pr(4, 3|\pi) = \binom{4}{1}\binom{4}{3}(\pi)^4(1-\pi)^4$$

$$T(x, y, \pi) = \{(x', y') : \Pr(x', y' \mid \pi) \leqslant \Pr(4, 3|\pi)\}$$

p-value: $\max_{\pi \in (0,1)} \sum_{(x,y) \in T(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}), \pi)} \Pr(x, y|\pi) = 0.50$ (for $\pi = 0.4$)

Fisher's exact text vs Barnard's exact test

Fisher's test: assumes the frequency $\sigma(S)$ of the pattern is fixed
Barnard's test: does not assume the frequency $\sigma(S)$ of the pattern is fixed

Fisher's exact text vs Barnard's exact test

Fisher's test: assumes the frequency $\sigma(S)$ of the pattern is fixed
Barnard's test: does not assume the frequency $\sigma(S)$ of the pattern is fixed

Note: Barnard's exact test depends on (unknown) nuisance parameter $\pi =$ probability that pattern $\mathcal{S}$ appears in a transaction.

Fisher's exact text vs Barnard's exact test

Fisher's test: assumes the frequency $\sigma(S)$ of the pattern is fixed
Barnard's test: does not assume the frequency $\sigma(S)$ of the pattern is fixed

Note: Barnard's exact test depends on (unknown) nuisance parameter $\pi =$ probability that pattern $\mathcal{S}$ appears in a transaction.

What about Fisher's exact test?

Fisher's exact text vs Barnard's exact test

Fisher's test: assumes the frequency $\sigma(S)$ of the pattern is fixed
Barnard's test: does not assume the frequency $\sigma(S)$ of the pattern is fixed

Note: Barnard's exact test depends on (unknown) nuisance parameter $\pi =$ probability that pattern $\mathcal{S}$ appears in a transaction.

What about Fisher's exact test?

Fixing the frequency $\sigma(S)$ of $\mathcal{S} \approx$ fixing the probability that $\mathcal{S}$ appears in a transaction

Fisher's exact text vs Barnard's exact test (2)

Fisher's test: assumes the frequency $\sigma(S)$ of the pattern is fixed
Barnard's test: does not assume the frequency $\sigma(S)$ of the pattern is fixed

Fisher's exact text vs Barnard's exact test (2)

Fisher's test: assumes the frequency $\sigma(S)$ of the pattern is fixed
Barnard's test: does not assume the frequency $\sigma(S)$ of the pattern is fixed

Which one is more appropriate?

Fisher's exact text vs Barnard's exact test (2)

Fisher's test: assumes the frequency $\sigma(S)$ of the pattern is fixed
Barnard's test: does not assume the frequency $\sigma(S)$ of the pattern is fixed

Which one is more appropriate?

Depends on how the data is collected!

Fisher's exact text vs Barnard's exact test (2)

Fisher's test: assumes the frequency $\sigma(S)$ of the pattern is fixed
Barnard's test: does not assume the frequency $\sigma(S)$ of the pattern is fixed

Which one is more appropriate?

Depends on how the data is collected!

In practice: everybody uses Fisher's text (computational reasons?)

Pattern mining and statistical hypothesis testing

Previous part: we had **one** pattern S we are interested in

Let p_S be the p-value for S.

Pattern mining and statistical hypothesis testing

Previous part: we had **one** pattern S we are interested in

Let p_S be the p-value for S.

Rejection rule:
Given a *statistical level* $\alpha \in (0,1)$: **reject** H_0 iff $p \leqslant \alpha \Rightarrow \mathcal{S}$ is significant!

Pattern mining and statistical hypothesis testing

Previous part: we had **one** pattern S we are interested in

Let p_S be the p-value for S.

Rejection rule:
Given a *statistical level* $\alpha \in (0,1)$: **reject** H_0 iff $p \leqslant \alpha \Rightarrow \mathcal{S}$ is significant!
$\Rightarrow$ probability false discovery $\leqslant \alpha$

Pattern mining and statistical hypothesis testing

Previous part: we had **one** pattern S we are interested in

Let p_S be the p-value for S.

Rejection rule:
Given a *statistical level* $\alpha \in (0, 1)$: **reject** H_0 iff $p \leqslant \alpha \Rightarrow \mathcal{S}$ is significant!
$\Rightarrow$ probability false discovery $\leqslant \alpha$

KDD scenario: we consider **multiple hypotheses** given by our dataset $\mathcal{D}$

Pattern mining and statistical hypothesis testing

Previous part: we had **one** pattern S we are interested in

Let p_S be the p-value for S.

Rejection rule:
Given a *statistical level* $\alpha \in (0, 1)$: **reject** H_0 iff $p \leqslant \alpha \Rightarrow \mathcal{S}$ is significant!
$\Rightarrow$ probability false discovery $\leqslant \alpha$

KDD scenario: we consider **multiple hypotheses** given by our dataset $\mathcal{D}$

What happens if we use the rejection rule above?

Outline

Multiple hypothesis testing

Let $\mathcal{H}$ be the **set of hypotheses** we want to test, and $m = |\mathcal{H}|$.

Multiple hypothesis testing

Let $\mathcal{H}$ be the **set of hypotheses** we want to test, and $m = |\mathcal{H}|$.

Proposition

$\mathbb{E}[\text{num. false discoveries}] = m \times \alpha$.

Multiple hypothesis testing

Let $\mathcal{H}$ be the **set of hypotheses** we want to test, and $m = |\mathcal{H}|$.

Proposition

$\mathbb{E}[\text{num. false discoveries}] = m \times \alpha$.

Typical values of α: 0.01, 0.05.

Multiple hypothesis testing

Let $\mathcal{H}$ be the **set of hypotheses** we want to test, and $m = |\mathcal{H}|$.

Proposition

$\mathbb{E}[\text{num. false discoveries}] = m \times \alpha$.

Typical values of α: 0.01, 0.05.

Value of m?

Multiple hypothesis testing

Let $\mathcal{H}$ be the **set of hypotheses** we want to test, and $m = |\mathcal{H}|$.

Proposition

$\mathbb{E}[\text{num. false discoveries}] = m \times \alpha$.

Typical values of α: 0.01, 0.05.

Value of m? If you are looking at itemsets from a universe $\mathcal{I}$ of items: $m = 2^{\mathcal{I}} - 1$

Multiple hypothesis testing

Let $\mathcal{H}$ be the **set of hypotheses** we want to test, and $m = |\mathcal{H}|$.

Proposition

$\mathbb{E}[\text{num. false discoveries}] = m \times \alpha$.

Typical values of α: 0.01, 0.05.

Value of m? If you are looking at itemsets from a universe $\mathcal{I}$ of items: $m = 2^{\mathcal{I}} - 1$

$\Rightarrow m \times \alpha$ is extremely high!

Multiple hypothesis testing

Let $\mathcal{H}$ be the **set of hypotheses** we want to test, and $m = |\mathcal{H}|$.

Proposition

$\mathbb{E}[\text{num. false discoveries}] = m \times \alpha$.

Typical values of α: 0.01, 0.05.

Value of m? If you are looking at itemsets from a universe $\mathcal{I}$ of items: $m = 2^{\mathcal{I}} - 1$

$\Rightarrow m \times \alpha$ is extremely high!

Need to consider the fact that we are testing multiple hypotheses!

We want **guarantees on the (expected) number of false discoveries**.

Multiple Hypothesis testing procedures

We want **guarantees on the (expected) number of false discoveries**.

V = number of false discoveries.

Multiple Hypothesis testing procedures

We want **guarantees on the (expected) number of false discoveries**.

V = number of false discoveries.

Family-Wise Error Rate (FWER): $\Pr[V \geqslant 1]$.

Multiple Hypothesis testing procedures

We want **guarantees on the (expected) number of false discoveries**.

V = number of false discoveries.

Family-Wise Error Rate (FWER): $\Pr[V \geqslant 1]$.

Two procedures with guarantees on the FWER

- Bonferroni correction
- Bonferroni-Holm procedure

Bonferroni correction

Let $\mathcal{H}$ be the set of hypotheses (*patterns*) we want to test, and $m = |\mathcal{H}|$.

Bonferroni correction

Let $\mathcal{H}$ be the set of hypotheses (*patterns*) we want to test, and $m = |\mathcal{H}|$.

Given a pattern $S \in \mathcal{H}$, let $H_{S,0}$ be the corresponding null hypothesis.

Bonferroni correction

Let $\mathcal{H}$ be the set of hypotheses (*patterns*) we want to test, and $m = |\mathcal{H}|$.

Given a pattern $S \in \mathcal{H}$, let $H_{S,0}$ be the corresponding null hypothesis.

Rejection rule: Given a *statistical level* $\alpha \in (0, 1)$: **reject** $H_{S,0}$ iff $p \leqslant \frac{\alpha}{m} \Rightarrow \mathcal{S}$ is significant!

Bonferroni correction

Let $\mathcal{H}$ be the set of hypotheses (*patterns*) we want to test, and $m = |\mathcal{H}|$.

Given a pattern $S \in \mathcal{H}$, let $H_{S,0}$ be the corresponding null hypothesis.

Rejection rule: Given a *statistical level* $\alpha \in (0, 1)$: **reject** $H_{S,0}$ iff $p \leqslant \frac{\alpha}{m} \Rightarrow \mathcal{S}$ is significant!

Intuition

Bonferroni correction

Let $\mathcal{H}$ be the set of hypotheses (*patterns*) we want to test, and $m = |\mathcal{H}|$.

Given a pattern $S \in \mathcal{H}$, let $H_{S,0}$ be the corresponding null hypothesis.

Rejection rule: Given a *statistical level* $\alpha \in (0, 1)$: **reject** $H_{S,0}$ iff $p \leqslant \frac{\alpha}{m} \Rightarrow \mathcal{S}$ is significant!

Intuition

- for each $\mathcal{S}$, $\Pr[\mathcal{S}$ is a false discovery $] \leqslant \frac{\alpha}{m}$

Bonferroni correction

Let $\mathcal{H}$ be the set of hypotheses (*patterns*) we want to test, and $m = |\mathcal{H}|$.

Given a pattern $S \in \mathcal{H}$, let $H_{S,0}$ be the corresponding null hypothesis.

Rejection rule: Given a *statistical level* $\alpha \in (0, 1)$: **reject** $H_{S,0}$ iff $p \leqslant \frac{\alpha}{m} \Rightarrow \mathcal{S}$ is significant!

Intuition

- for each $\mathcal{S}$, $\Pr[\mathcal{S}$ is a false discovery $] \leqslant \frac{\alpha}{m}$
- union bound on m events: $\Pr[> 0$ false discoveries $]$

Bonferroni correction

Let $\mathcal{H}$ be the set of hypotheses (*patterns*) we want to test, and $m = |\mathcal{H}|$.

Given a pattern $S \in \mathcal{H}$, let $H_{S,0}$ be the corresponding null hypothesis.

Rejection rule: Given a *statistical level* $\alpha \in (0,1)$: **reject** $H_{S,0}$ iff $p \leqslant \frac{\alpha}{m} \Rightarrow \mathcal{S}$ is significant!

Intuition

- for each $\mathcal{S}$, $\Pr[\mathcal{S}$ is a false discovery $] \leqslant \frac{\alpha}{m}$
- union bound on m events: $\Pr[> 0$ false discoveries $]$ $\leqslant \sum_{\mathcal{S}\in\mathcal{H}} \Pr[S$ is false discovery $] \leqslant |\mathcal{H}|\frac{\alpha}{m} \leqslant \alpha$

Bonferroni-Holm procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

Bonferroni-Holm procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values

Bonferroni-Holm procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the minimum value such that $p_k > \frac{\alpha}{m+1-k}$

Bonferroni-Holm procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the minimum value such that $p_k > \frac{\alpha}{m+1-k}$
3. **rejection rule**: reject the hypotheses (patterns) associated with $p_1, p_2, \ldots, p_{k-1}$

Bonferroni-Holm procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the minimum value such that $p_k > \frac{\alpha}{m+1-k}$
3. **rejection rule**: reject the hypotheses (patterns) associated with $p_1, p_2, \ldots, p_{k-1}$

More powerful than Bonferroni correction: p_i compared with $\frac{\alpha}{m+1-i}$ vs $\frac{\alpha}{m}$.

Bonferroni-Holm procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the minimum value such that $p_k > \frac{\alpha}{m+1-k}$
3. **rejection rule**: reject the hypotheses (patterns) associated with $p_1, p_2, \ldots, p_{k-1}$

More powerful than Bonferroni correction: p_i compared with $\frac{\alpha}{m+1-i}$ vs $\frac{\alpha}{m}$.

However: both **require very small p-values** to flag patterns as significant when m is large.

False Discovery Rate

Let V be the number of false discoveries.

False Discovery Rate

Let V be the number of false discoveries.

The requirement on **FWER** can be **too strict**!

False Discovery Rate

Let V be the number of false discoveries.

The requirement on **FWER** can be **too strict**!

Family-Wise Error Rate (FWER): $\Pr[V \geqslant 1]$.

False Discovery Rate

Let V be the number of false discoveries.

The requirement on **FWER** can be **too strict**!

Family-Wise Error Rate (FWER): $\Pr[V \geqslant 1]$.

Let R the number of discoveries (i.e., rejected hypotheses).

False Discovery Rate

Let V be the number of false discoveries.

The requirement on **FWER** can be **too strict**!

Family-Wise Error Rate (FWER): $\Pr[V \geqslant 1]$.

Let R the number of discoveries (i.e., rejected hypotheses).

Relaxed requirement: control the False Discovery Rate

False Discovery Rate

Let V be the number of false discoveries.

The requirement on **FWER** can be **too strict**!

Family-Wise Error Rate (FWER): $\Pr[V \geqslant 1]$.

Let R the number of discoveries (i.e., rejected hypotheses).

Relaxed requirement: control the False Discovery Rate

False Discovery Rate (FDR): $\mathbb{E}[V/R]$ (assuming $V/R = 0$ when $R = 0$).

Benjamini-Hochberg procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Benjamini-Hochberg procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

Benjamini-Hochberg procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values

Benjamini-Hochberg procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the maximum value such that $p_k \leqslant \frac{\alpha k}{m}$

Benjamini-Hochberg procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the maximum value such that $p_k \leqslant \frac{\alpha k}{m}$
3. **rejection rule**: reject the hypotheses (patterns) associated with $p_1, p_2, \ldots, p_k$

Benjamini-Hochberg procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the maximum value such that $p_k \leqslant \frac{\alpha k}{m}$
3. **rejection rule**: reject the hypotheses (patterns) associated with $p_1, p_2, \ldots, p_k$

Note: more powerful than Bonferroni and Bonferroni-Holm

Benjamini-Hochberg procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the maximum value such that $p_k \leqslant \frac{\alpha k}{m}$
3. **rejection rule**: reject the hypotheses (patterns) associated with $p_1, p_2, \ldots, p_k$

Note: more powerful than Bonferroni and Bonferroni-Holm

Assumption: hypotheses are independent.

Benjamini-Yekutieli procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Benjamini-Yekutieli procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

Benjamini-Yekutieli procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values

Benjamini-Yekutieli procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the maximum value such that $p_k \leqslant \frac{\alpha k}{m \sum_{i=1}^{m}(1/i)}$

Benjamini-Yekutieli procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the maximum value such that $p_k \leqslant \frac{\alpha k}{m \sum_{i=1}^{m}(1/i)}$
3. **rejection rule**: reject the hypotheses (patterns) associated with $p_1, p_2, \ldots, p_k$

Benjamini-Yekutieli procedure

Let $\mathcal{H}$ the set of hypotheses (*patterns*) to be tested, and $m = |\mathcal{H}|$.

Sequential procedure:

1. order the hypotheses (patterns) by increasing p-values: let $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_m$ be the sorted p-values
2. let k be the maximum value such that $p_k \leqslant \frac{\alpha k}{m \sum_{i=1}^{m}(1/i)}$
3. **rejection rule**: reject the hypotheses (patterns) associated with $p_1, p_2, \ldots, p_k$

Note: does not require independence of hypotheses.

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Number of hypotheses $m = \binom{15}{6} = 6435$

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Number of hypotheses $m = \binom{15}{6} = 6435$

- "m is large, will never find significant results"! 

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Number of hypotheses $m = \binom{15}{6} = 6435$

- "m is large, will never find significant results"! 
- "let me select some hypotheses first, and then do the testing..."

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Number of hypotheses $m = \binom{15}{6} = 6435$

- "m is large, will never find significant results"! 
- "let me select some hypotheses first, and then do the testing..."
- find pattern $\mathcal{S}$ with highest value $\sigma_1(\mathcal{S}) - \sigma_0(\mathcal{S})$: $\sigma_1(\mathcal{S}) = 10, \sigma_0(\mathcal{S}) = 0$

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Number of hypotheses $m = \binom{15}{6} = 6435$

- "m is large, will never find significant results"!
- "let me select some hypotheses first, and then do the testing..."
- find pattern $\mathcal{S}$ with highest value $\sigma_1(\mathcal{S}) - \sigma_0(\mathcal{S})$:
  $\sigma_1(\mathcal{S}) = 10, \sigma_0(\mathcal{S}) = 0$
- "I am going to test only $\mathcal{S}$!"

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Number of hypotheses $m = \binom{15}{6} = 6435$

- "m is large, will never find significant results"! 
- "let me select some hypotheses first, and then do the testing..."
- find pattern $\mathcal{S}$ with highest value $\sigma_1(\mathcal{S}) - \sigma_0(\mathcal{S})$: $\sigma_1(\mathcal{S}) = 10, \sigma_0(\mathcal{S}) = 0$
- "I am going to test only $\mathcal{S}$!"
- Fisher's exact test p-value $= 0.0001$

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Number of hypotheses $m = \binom{15}{6} = 6435$

- "m is large, will never find significant results"! 
- "let me select some hypotheses first, and then do the testing..."
- find pattern $\mathcal{S}$ with highest value $\sigma_1(\mathcal{S}) - \sigma_0(\mathcal{S})$:
  $\sigma_1(\mathcal{S}) = 10, \sigma_0(\mathcal{S}) = 0$
- "I am going to test only $\mathcal{S}$!"
- Fisher's exact test p-value $= 0.0001$
- "$\mathcal{S}$ is very significant!!!"

Choosing hypotheses *before* testing?

Dataset $\mathcal{D}$:

- 10 transactions with label c_1, 10 transactions with label c_0
- items $\mathcal{I}$ with $|\mathcal{I}| = 13$

We are interested only in patterns of size 6.

Number of hypotheses $m = \binom{15}{6} = 6435$

- "m is large, will never find significant results"!
- "let me select some hypotheses first, and then do the testing..."
- find pattern $\mathcal{S}$ with highest value $\sigma_1(\mathcal{S}) - \sigma_0(\mathcal{S})$: $\sigma_1(\mathcal{S}) = 10, \sigma_0(\mathcal{S}) = 0$
- "I am going to test only $\mathcal{S}$!"
- Fisher's exact test p-value $= 0.0001$
- "$\mathcal{S}$ is very significant!!!"

"$\mathcal{S}$ is very significant!!!" ☺

"$\mathcal{S}$ is very significant!!!" ☺

BUT IT IS NOT!

"$\mathcal{S}$ is very significant!!!" ☺

BUT IT IS NOT!

Assume that $\mathcal{D}$ is generate as follows: for each pattern $\mathcal{S}$

"$\mathcal{S}$ is very significant!!!" ☺

BUT IT IS NOT!

Assume that $\mathcal{D}$ is generate as follows: for each pattern $\mathcal{S}$

- consider one of its 10 occurrences

“$\mathcal{S}$ is very significant!!!” ☺

BUT IT IS NOT!

Assume that $\mathcal{D}$ is generate as follows: for each pattern $\mathcal{S}$

- consider one of its 10 occurrences
- place it in a transaction with label c_0 with probability $1/2$, and in a transaction with label c_1 with probability $1/2$ otherwise

"$\mathcal{S}$ is very significant!!!" ☺

BUT IT IS NOT!

Assume that $\mathcal{D}$ is generate as follows: for each pattern $\mathcal{S}$

- consider one of its 10 occurrences
- place it in a transaction with label c_0 with probability $1/2$, and in a transaction with label c_1 with probability $1/2$ otherwise
- $\mathcal{S}$ **is not associated with class labels!**

"$\mathcal{S}$ is very significant!!!" ☺

BUT IT IS NOT!

Assume that $\mathcal{D}$ is generate as follows: for each pattern $\mathcal{S}$

- consider one of its 10 occurrences
- place it in a transaction with label c_0 with probability $1/2$, and in a transaction with label c_1 with probability $1/2$ otherwise
- **$\mathcal{S}$ is not associated with class labels!**

For a *given* $\mathcal{S}$, the probability $\sigma_1(\mathcal{S}) = 10$ and $\sigma_0(\mathcal{S}) = 0$ is $(1/2)^{10} = 1/1024$

"$\mathcal{S}$ is very significant!!!" ☺

BUT IT IS NOT!

Assume that $\mathcal{D}$ is generate as follows: for each pattern $\mathcal{S}$

- consider one of its 10 occurrences
- place it in a transaction with label c_0 with probability $1/2$, and in a transaction with label c_1 with probability $1/2$ otherwise
- **$\mathcal{S}$ is not associated with class labels!**

For a *given* $\mathcal{S}$, the probability $\sigma_1(\mathcal{S}) = 10$ and $\sigma_0(\mathcal{S}) = 0$ is $(1/2)^{10} = 1/1024$

In expectation, there will be 6 patterns with $\sigma_1(\mathcal{S}) = 10$ and $\sigma_0(\mathcal{S}) = 0$ and they are all false discoveries!

Where is the problem?

We selected hypotheses based on $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$,

Where is the problem?

We selected hypotheses based on $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$,
and $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$ *is clearly related to the* p*-value*

Where is the problem?

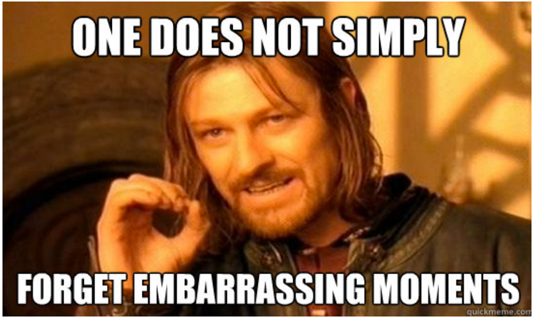
We selected hypotheses based on $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$, and $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$ *is clearly related to the p-value*

So we have essentially **looked at p-values of all hypotheses** and **pretended we did not!** 

Where is the problem?

We selected hypotheses based on $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$,
and $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$ *is clearly related to the p-value*

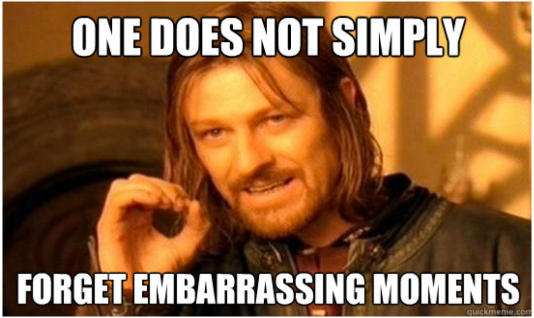
So we have essentially **looked at p-values of all hypotheses** and **pretended we did not!**



Where is the problem?

We selected hypotheses based on $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$,
and $\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$ *is clearly related to the p-value*

So we have essentially **looked at p-values of all hypotheses** and **pretended we did not!**



When in doubt: assume you have looked at all hypotheses!

Outline

Selecting hypotheses

All approaches seen so far for controlling the FWER and the FDR depend on the *set $\mathcal{H}$ of hypotheses*, e.g., on its size.

A smaller $\mathcal{H}$ *may* lead to a *higher corrected significance threshold*, thus to *higher power*.

Selecting hypotheses

All approaches seen so far for controlling the FWER and the FDR depend on the *set $\mathcal{H}$ of hypotheses*, e.g., on its size.

A smaller $\mathcal{H}$ *may* lead to a *higher corrected significance threshold*, thus to *higher power*.

QUESTION: can we *shrink $\mathcal{H}$ a posteriori*?

I.e., Can we use $\mathcal{D}$ to select $\mathcal{H}' \subsetneq \mathcal{H}$

such that $\mathcal{H} \setminus \mathcal{H}'$ only contains *non-significant* hypotheses?

Selecting hypotheses

All approaches seen so far for controlling the FWER and the FDR depend on the *set* $\mathcal{H}$ *of hypotheses*, e.g., on its size.

A smaller $\mathcal{H}$ *may* lead to a *higher corrected significance threshold*, thus to *higher power*.

QUESTION: can we *shrink* $\mathcal{H}$ *a posteriori*?

I.e., Can we use $\mathcal{D}$ to select $\mathcal{H}' \subsetneq \mathcal{H}$

such that $\mathcal{H} \setminus \mathcal{H}'$ only contains *non-significant* hypotheses?

ANSWER: No... and yes! ☺

How not to select hypotheses

The one thing you *must remember* from this tutorial!

Do not do this:

How not to select hypotheses

The one thing you *must remember* from this tutorial!

Do not do this:

1) Perform each individual test for each hypothesis using $\mathcal{D}$.

2) *Use the test results* to select which hypotheses to include in $\mathcal{H}'$.

3) Use your favorite MHC to bound the FWER/FDR on $\mathcal{H}'$.

How not to select hypotheses

The one thing you *must remember* from this tutorial!

Do not do this:

1) Perform each individual test for each hypothesis using $\mathcal{D}$.

2) *Use the test results* to select which hypotheses to include in $\mathcal{H}'$.

3) Use your favorite MHC to bound the FWER/FDR on $\mathcal{H}'$.

Selecting $\mathcal{H}'$ must be done *without performing the tests on* $\mathcal{D}$.

The holdout approach

1. Partition $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$: $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$ and $\mathcal{D}_1 \cap \mathcal{D}_2 = \varnothing$.

2. Apply some selection procedure to $\mathcal{D}_1$ to select $\mathcal{H}'$ (it may include performing the tests on $\mathcal{D}_1$).

3) Perform the individual test for each hypothesis in $\mathcal{H}'$ on $\mathcal{D}_2$, using any MHC method.

The holdout approach

1. Partition $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$: $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{D}$ and $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$.

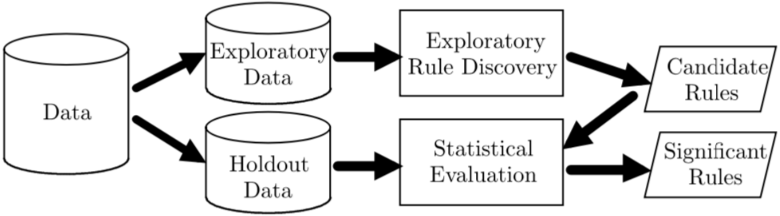
2. Apply some selection procedure to $\mathcal{D}_1$ to select $\mathcal{H}'$ (it may include performing the tests on $\mathcal{D}_1$).

3) Perform the individual test for each hypothesis in $\mathcal{H}'$ on $\mathcal{D}_2$, using any MHC method.

Splitting $\mathcal{D}$ is similar to splitting a labeled set into training and test sets.

An example: holdout for significant itemsets

G. Webb, Discovering Significant Patterns, Mach. Learn. 2007



When holdout works and why

Holdout can be used *only* when $\mathcal{D}$ can be partitioned into $\mathcal{D}_1$ and $\mathcal{D}_2$ s.t. $\mathcal{D}_1$ and $\mathcal{D}_2$ are *samples from the null distribution*.

When holdout works and why

Holdout can be used *only* when $\mathcal{D}$ can be partitioned into $\mathcal{D}_1$ and $\mathcal{D}_2$ s.t. $\mathcal{D}_1$ and $\mathcal{D}_2$ are *samples from the null distribution*.

Such partitioning may *not exist or be known*.

When holdout works and why

Holdout can be used *only* when $\mathcal{D}$ can be partitioned into $\mathcal{D}_1$ and $\mathcal{D}_2$ s.t. $\mathcal{D}_1$ and $\mathcal{D}_2$ are *samples from the null distribution*.

Such partitioning may *not exist or be known*. E.g., for *graphs*:

Split the set of nodes in two and claim that each of the resulting induced subgraphs is a sample from the original distribution:

what do you do with edges crossing the two sets?

Formally: holdout works when the elements of $\mathcal{D}$ are *identically distributed exchangeable random variables*.

How selective shall we be?

$\mathcal{Z}_\alpha \subseteq \mathcal{H}$: set of α-significant hypotheses.

When selecting $\mathcal{H}'$, we may *get rid of some α-significant ones*:

$$\mathcal{Z}_\alpha \cap (\mathcal{H} \backslash \mathcal{H}') \neq \emptyset.$$

Does the power still increases just because the corrected significance threshold increases?

How selective shall we be?

$\mathcal{Z}_\alpha \subseteq \mathcal{H}$: set of α-significant hypotheses.

When selecting $\mathcal{H}'$, we may *get rid of some α-significant ones*:

$$\mathcal{Z}_\alpha \cap (\mathcal{H} \backslash \mathcal{H}') \neq \emptyset.$$

Does the power still increases just because the corrected significance threshold increases? **Unclear!**

One can build examples where power $\uparrow$, $\downarrow$, or $=$.

Take-away message

Being *more or less selective* in choosing $\mathcal{H}'$ has a *complicated effect on power* that cannot be clearly evaluated a priori.

This downside of holdout is due to the fact that

holdout *may* remove α-significant hypotheses from $\mathcal{H}$.

OTOH, holdout is a *simple natural procedure*, and

it *generally* leads to higher power because

most discarded hypotheses are not α-significant.

Take-away message

Being *more or less selective* in choosing $\mathcal{H}'$ has a *complicated effect on power* that cannot be clearly evaluated a priori.

This downside of holdout is due to the fact that

holdout *may* remove α-significant hypotheses from $\mathcal{H}$.

OTOH, holdout is a *simple natural procedure*, and

it *generally* leads to higher power because

most discarded hypotheses are not α*-significant*.

Coming up: how to discard *only* non-α-significant hypotheses.

Outline

A breakthrough [Tarone 1990]

Fisher's exact test statistic is **discrete**

A breakthrough [Tarone 1990]

Fisher's exact test statistic is **discrete**

$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

A breakthrough [Tarone 1990]

Fisher's exact test statistic is **discrete**
$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

Example Consider a dataset with $n_0 = 5$, $n_1 = 10$, $\sigma(S) = 5$ ($\Rightarrow n = 15, n - \sigma(S) = 10$).

A breakthrough [Tarone 1990]

Fisher's exact test statistic is **discrete**
$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

Example Consider a dataset with $n_0 = 5$, $n_1 = 10$, $\sigma(S) = 5$ ($\Rightarrow n = 15, n - \sigma(S) = 10$).

Smallest p-value for S?

A breakthrough [Tarone 1990]

Fisher's exact test statistic is **discrete**
$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

Example Consider a dataset with $n_0 = 5$, $n_1 = 10$, $\sigma(S) = 5$ ($\Rightarrow n = 15, n - \sigma(S) = 10$).

Smallest p-value for S? When $\sigma_1(S) = 5$

A breakthrough [Tarone 1990]

Fisher's exact test statistic is **discrete**
$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

Example Consider a dataset with $n_0 = 5$, $n_1 = 10$, $\sigma(S) = 5$ ($\Rightarrow n = 15, n - \sigma(S) = 10$).

Smallest p-value for S? When $\sigma_1(S) = 5$

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	5	0	5
$\ell(t_i) = c_0$	0	10	10
Col. m.	5	10	15

A breakthrough [Tarone 1990]

Fisher's exact test statistic is **discrete**
$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

Example Consider a dataset with $n_0 = 5$, $n_1 = 10$, $\sigma(S) = 5$ ($\Rightarrow n = 15, n - \sigma(S) = 10$).

Smallest p-value for S? When $\sigma_1(S) = 5$

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	5	0	5
$\ell(t_i) = c_0$	0	10	10
Col. m.	5	10	15

minimum attainable p-value $= 3 \times 10^{-4}$

A breakthrough [Tarone 1990] (2)

Fisher's exact test statistic is **discrete**

$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

A breakthrough [Tarone 1990] (2)

Fisher's exact test statistic is **discrete**

$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

A breakthrough [Tarone 1990] (2)

Fisher's exact test statistic is **discrete**
$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Let $p^F(\sigma(\mathcal{S}), x)$ be Fisher's exact test for pattern $\mathcal{S}$ with support $\sigma(\mathcal{S})$ and $\sigma_1(\mathcal{S}) = x$.

A breakthrough [Tarone 1990] (2)

Fisher's exact test statistic is **discrete**
$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Let $p^F(\sigma(\mathcal{S}), x)$ be Fisher's exact test for pattern $\mathcal{S}$ with support $\sigma(\mathcal{S})$ and $\sigma_1(\mathcal{S}) = x$.

Note that $\max\{0, n_1 - (n - \sigma(\mathcal{S}))\} \leqslant x \leqslant \min\{\sigma_1(\mathcal{S}), n_1\}$

A breakthrough [Tarone 1990] (2)

Fisher's exact test statistic is **discrete**
$\Rightarrow$ there is a **minimum attainable** p**-value** for a pattern $\mathcal{S}$.

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Let $p^F(\sigma(\mathcal{S}), x)$ be Fisher's exact test for pattern $\mathcal{S}$ with support $\sigma(\mathcal{S})$ and $\sigma_1(\mathcal{S}) = x$.

Note that $\max\{0, n_1 - (n - \sigma(\mathcal{S}))\} \leqslant x \leqslant \min\{\sigma_1(\mathcal{S}), n_1\} \Rightarrow$ the range of $p^F(\sigma(\mathcal{S}), x)$ depends only on $\sigma(\mathcal{S})$ (since n_1 is fixed)

A breakthrough [Tarone 1990] (3)

Then the minimum achievable p-value for $\mathcal{S}$ is:

$$\psi(\sigma(\mathcal{S})) = \min_{\max\{0, n_1 - (n - \sigma(\mathcal{S}))\} \leqslant x \leqslant \min\{\sigma_1(\mathcal{S}), n_1\}} \{p^F(\sigma(\mathcal{S}), x)\}$$

A breakthrough [Tarone 1990] (3)

Then the minimum achievable p-value for $\mathcal{S}$ is:

$$\psi(\sigma(\mathcal{S})) = \min_{\max\{0, n_1 - (n - \sigma(\mathcal{S}))\} \leqslant x \leqslant \min\{\sigma_1(\mathcal{S}), n_1\}} \{p^F(\sigma(\mathcal{S}), x)\}$$

Tarone's result: if your are testing hypotheses with significance level δ, then **hypotheses that cannot be significant do not count as hypotheses for Bonferroni's correction!** ☺

A breakthrough [Tarone 1990] (4)

$\mathcal{S}$ cannot be significant with significance level δ if
$\psi(\sigma(\mathcal{S})) > \alpha'$

A breakthrough [Tarone 1990] (4)

$\mathcal{S}$ cannot be significant with significance level δ if $\psi(\sigma(\mathcal{S})) > \alpha' \Rightarrow \mathcal{S}$ is **untestable**.

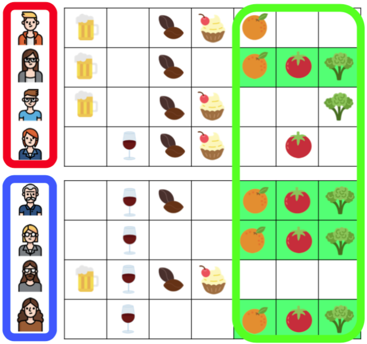
A breakthrough [Tarone 1990] (4)

$\mathcal{S}$ cannot be significant with significance level δ if $\psi(\sigma(\mathcal{S})) > \alpha' \Rightarrow \mathcal{S}$ is **untestable**.

Set of **testable hypotheses** (for significance level δ):

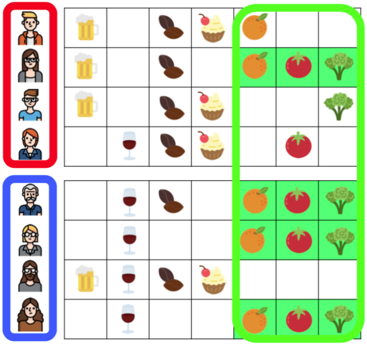
$$\mathcal{T}(\delta) = \{\mathcal{S} \mid \psi(\sigma(\mathcal{S})) \leqslant \delta\}$$

Example: market basket analysis



$$\mathcal{S} = \{\text{orange, tomato, broccoli}\}$$

Example: market basket analysis

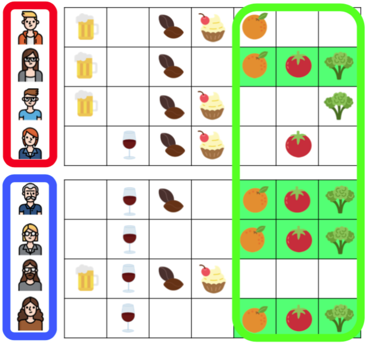


$\mathcal{S} = \{$orange, tomato, broccoli$\}$

minimum achievable p-value

$$\psi(\sigma(\mathcal{S})) = \min_{0 \leqslant x \leqslant \min\{\sigma_1(\mathcal{S}), n_1\}} \{p^F(\sigma(\mathcal{S}), x)\}$$

Example: market basket analysis



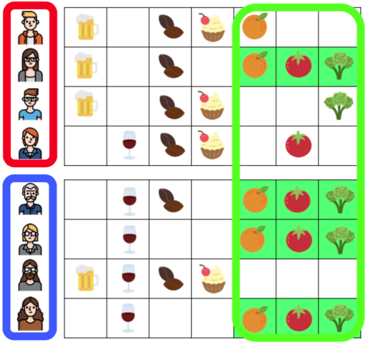
$\mathcal{S} = \{$orange, tomato, broccoli$\}$

minimum achievable p-value

$$\psi(\sigma(\mathcal{S})) = \min_{0 \leqslant x \leqslant \min\{\sigma_1(\mathcal{S}), n_1\}} \{p^F(\sigma(\mathcal{S}), x)\}$$

obtained for $x = 4$: $\psi(4) = 0.014$.

Example: market basket analysis



$\mathcal{S} = \{$orange, tomato, broccoli$\}$

minimum achievable p-value

$$\psi(\sigma(\mathcal{S})) = \min_{0 \leqslant x \leqslant \min\{\sigma_1(\mathcal{S}), n_1\}} \{p^F(\sigma(\mathcal{S}), x)\}$$

obtained for $x = 4$: $\psi(4) = 0.014$.

$\Rightarrow$ if significance level is $\delta = 0.01$, you do not need to count $\mathcal{S}$ among the hypotheses!

Tarone's Improved Bonferroni correction

Set of **testable hypotheses**:

$$\mathcal{T}(\delta) = \{\mathcal{S} \mid \psi(\sigma(\mathcal{S})) \leqslant \delta\}$$

Tarone's Improved Bonferroni correction

Set of **testable hypotheses**:

$$\mathcal{T}(\delta) = \{\mathcal{S} \mid \psi(\sigma(\mathcal{S})) \leqslant \delta\}$$

Rejection rule:
Given a *statistical level* $\alpha \in (0, 1)$, let $\delta \leqslant \alpha / |\mathcal{T}(\delta)|$: **reject** H_0 iff $p \leqslant \delta \Rightarrow \mathcal{S}$ is significant!

Tarone's Improved Bonferroni correction

Set of **testable hypotheses**:

$$\mathcal{T}(\delta) = \{\mathcal{S} \mid \psi(\sigma(\mathcal{S})) \leqslant \delta\}$$

Rejection rule:
Given a *statistical level* $\alpha \in (0,1)$, let $\delta \leqslant \alpha/|\mathcal{T}(\delta)|$: **reject** H_0 iff $p \leqslant \delta \Rightarrow \mathcal{S}$ is significant!

Theorem

The FWER is $\leqslant \alpha$.

Tarone's Improved Bonferroni correction

Set of **testable hypotheses**:

$$\mathcal{T}(\delta) = \{\mathcal{S} \mid \psi(\sigma(\mathcal{S})) \leqslant \delta\}$$

Rejection rule:
Given a *statistical level* $\alpha \in (0, 1)$, let $\delta \leqslant \alpha/|\mathcal{T}(\delta)|$: **reject** H_0 iff $p \leqslant \delta \Rightarrow \mathcal{S}$ is significant!

Theorem

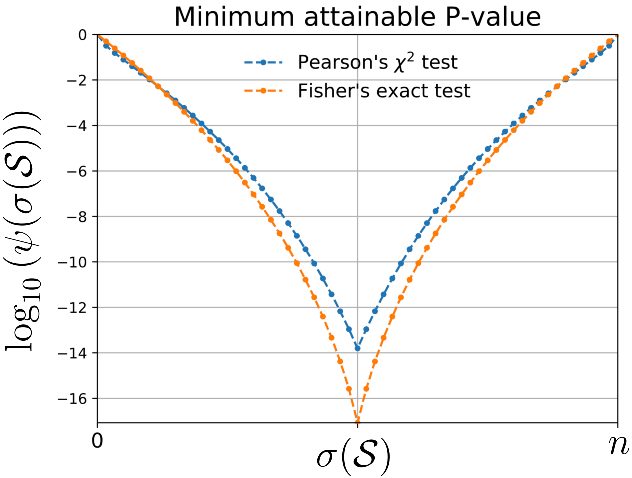
The FWER is $\leqslant \alpha$.

Idea: find $\delta^* = \max\{\delta : \delta \leqslant \alpha/|\mathcal{T}(\delta)|\}$!

Still with us? :)

Outline

Intuitively: patterns with low (and very high) support $\sigma(\mathcal{S})$ in the data provide less "evidence" of being significant $\rightarrow$ higher $\psi(\sigma(\mathcal{S}))$!



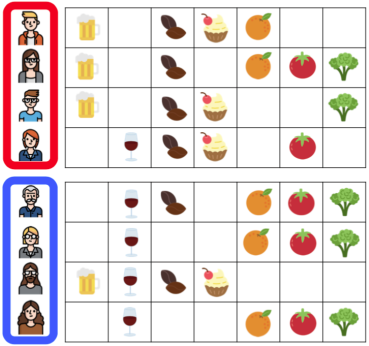
$\boxed{n = 60,\ n_1 = 30.}$

(from F. Llinares-López, D. Roqueiro, *Significant Pattern Mining for Biomarker Discovery*, ISMB18 Tutorial.)

Monotonicity of patterns' support:

Theorem

Let $\mathcal{S}$ be an itemset. Then it holds $\sigma(\mathcal{S}') \leqslant \sigma(\mathcal{S})$ for all $\mathcal{S}' \supseteq \mathcal{S}$.



Example:

$\mathcal{S}' = \{\text{tomato, broccoli}\}$, $\mathcal{S} = \{\text{tomato}\}$

$\sigma(\mathcal{S}') = 4 \leqslant \sigma(\mathcal{S}) = 5.$

Monotonicity of patterns' min. achievable p-value:
LAMP[1]: define the function $\hat{\psi}(\cdot)$ as

$$\hat{\psi}(x) = \begin{cases} \psi(x) & \text{, if } x \leqslant n_1 \\ \psi(n_1) & \text{, othw.} \end{cases}$$

Theorem

For Fisher's test it holds $\hat{\psi}(x) \leqslant \hat{\psi}(y)$ for all $x \geqslant y$.

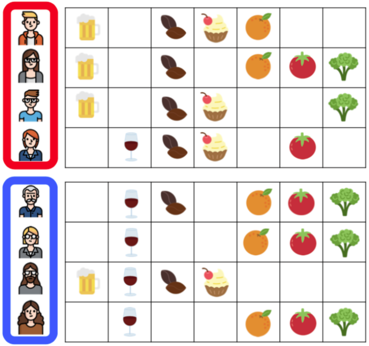
(in simpler terms: $\hat{\psi}(x)$ is monotone)

[1] Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. *Statistical significance of combinatorial regulations.* Proceedings of the National Academy of Sciences (2013).

Intuition: connection between monotonicity of patterns' min. achievable p-value and patterns' support:

Theorem

Let $\mathcal{S}$ be an itemset. Then $\hat{\psi}(\sigma(\mathcal{S})) \leqslant \hat{\psi}(\sigma(\mathcal{S}'))$ for all $\mathcal{S}' \supseteq \mathcal{S}$.



Example:

$\mathcal{S}' = \{\text{wine , coffee}\}$, $\mathcal{S} = \{\text{wine}\}$

$\sigma(\mathcal{S}') = 3 \leqslant \sigma(\mathcal{S}) = 5$

$\hat{\psi}(\sigma(\mathcal{S}')) = \hat{\psi}(3) = 0.14 \geqslant \hat{\psi}(\sigma(\mathcal{S})) = \hat{\psi}(5) = 0.03$

This holds for *itemsets* and many other type of patterns with monotonicity of support (i.e., *subgraphs*, *sequential patterns*, *subgroups*, ...)

Intuition: let's benefit from extensive research in **Frequent Pattern Mining algorithms!**

Frequent Pattern Mining

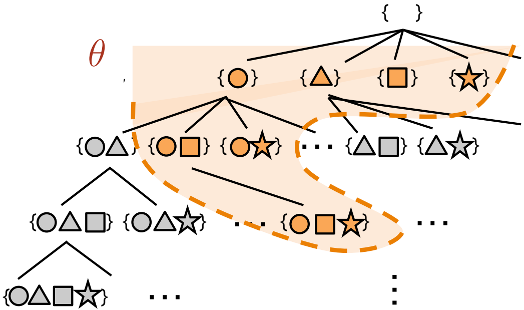
Frequent Pattern Mining: given $\mathcal{D}$, compute the *set of frequent patterns* $FP(\mathcal{D}, \mathcal{H}, \theta) \subseteq \mathcal{H}$ w.r.t. support θ, that is

$$FP(\mathcal{D}, \mathcal{H}, \theta) := \{\mathcal{S} \in \mathcal{H} : \sigma(\mathcal{S}) \geqslant \theta\} .$$

Frequent Pattern Mining: given $\mathcal{D}$, compute the *set of frequent patterns* $FP(\mathcal{D}, \mathcal{H}, \theta) \subseteq \mathcal{H}$ w.r.t. support θ, that is

$$FP(\mathcal{D}, \mathcal{H}, \theta) := \{\mathcal{S} \in \mathcal{H} : \sigma(\mathcal{S}) \geqslant \theta\} \,.$$

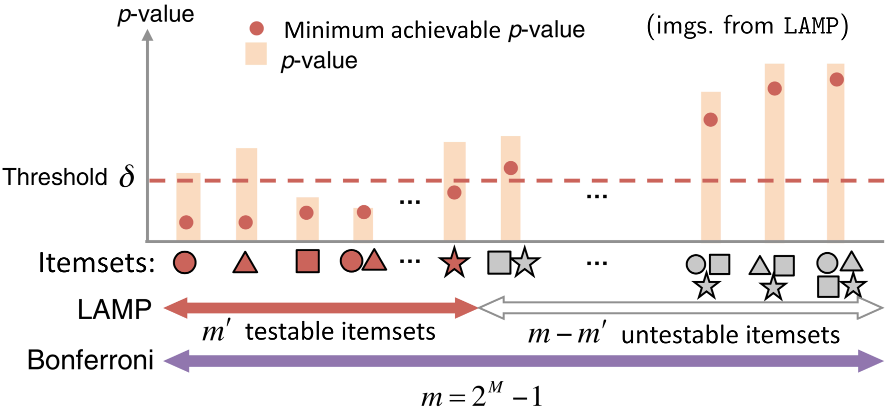
One solution: **Explore the search tree of $\mathcal{H}$, pruning low-support subtrees:**

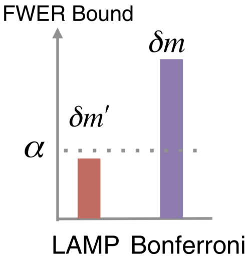


LAMP

LAMP[2]: first method to compute $\delta^* = \max\{\delta : \delta|\mathcal{T}(\delta)| \leqslant \alpha\}$ enumerating Frequent Itemsets.

(imgs. from LAMP)

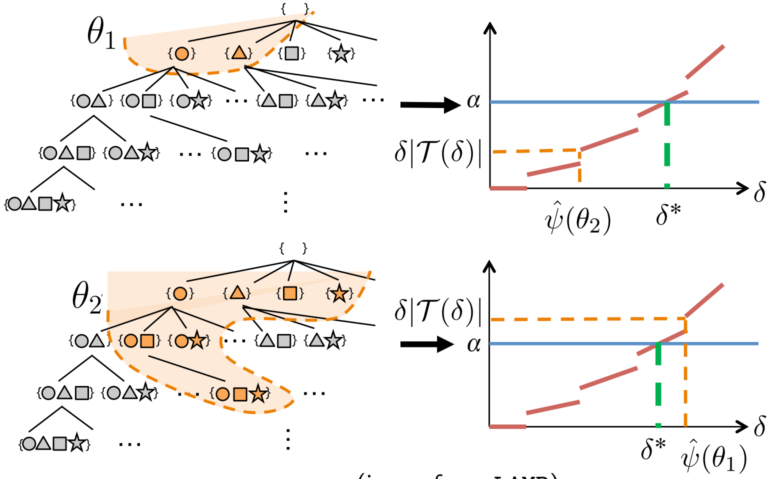




[2] Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. *Statistical significance of combinatorial regulations.* Proceedings of the National Academy of Sciences (2013).

LAMP algorithm

LAMP: compute $\delta^* = \max\{\delta : \delta|\mathcal{T}(\delta)| \leqslant \alpha\}$ enumerating Frequent Itemsets.



Performs multiple Frequent Pattern Mining instances to evaluate $|\mathcal{T}(\delta)|$.
i.e., start with $\theta = n$ and decrease it until δ^* is found.

(imgs. from LAMP)

LAMP algorithm

Let $FP(\mathcal{D}, \mathcal{H}, \theta) := \{\mathcal{S} \in \mathcal{H} : \sigma(\mathcal{S}) \geqslant \theta\}$.

Algorithm 1: LAMP

Input: dataset $\mathcal{D}$, upper bound to $FWER$ α.
Output: $\delta^* = \max\{\delta : \delta \leqslant \alpha / |\mathcal{T}(\delta)|\}$.

$\theta \leftarrow n$;
while $\alpha / |FP(\mathcal{D}, \mathcal{H}, \theta)| \geqslant \hat{\psi}(\theta)$ **do** $\theta \leftarrow \theta - 1$;
return $\alpha / |FP(\mathcal{D}, \mathcal{H}, \theta + 1)|$;

LAMP algorithm

Let $FP(\mathcal{D}, \mathcal{H}, \theta) := \{\mathcal{S} \in \mathcal{H} : \sigma(\mathcal{S}) \geqslant \theta\}$.

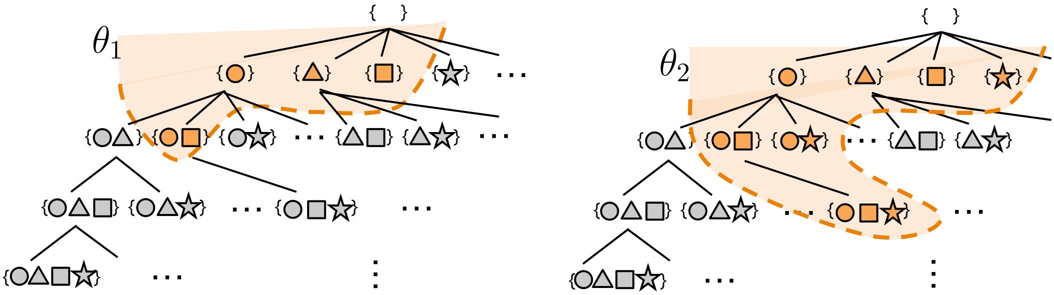
Algorithm 2: LAMP

Input: dataset $\mathcal{D}$, upper bound to $FWER$ α.
Output: $\delta^* = \max\{\delta : \delta \leqslant \alpha / |\mathcal{T}(\delta)|\}$.

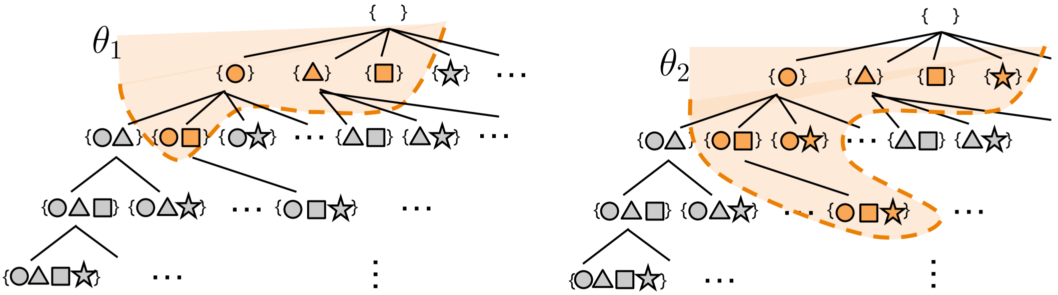
$\theta \leftarrow n$;
while $\alpha / |FP(\mathcal{D}, \mathcal{H}, \theta)| \geqslant \hat{\psi}(\theta)$ **do** $\theta \leftarrow \theta - 1$;
return $\alpha / |FP(\mathcal{D}, \mathcal{H}, \theta + 1)|$;

Problem: the same patterns are explored many times!

i.e.: *all* $\mathcal{S} \in FP(\mathcal{D}, \mathcal{H}, \theta)$ are explored again when $FP(\mathcal{D}, \mathcal{H}, \theta - 1)$ is explored



For $\theta = \theta_2$ we count again all patterns already counted for $\theta = \theta_1 \geqslant \theta_2$!



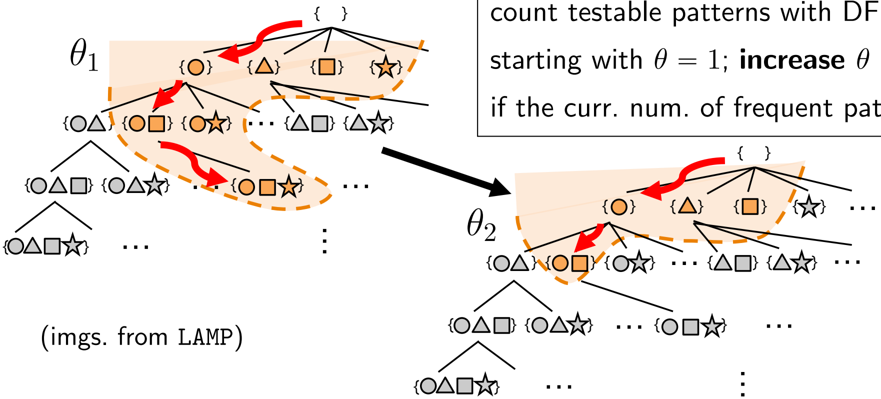
For $\theta = \theta_2$ we count again all patterns already counted for $\theta = \theta_1 \geqslant \theta_2$!

Can we count patterns only once?

SupportIncrease

SupportIncrease[3]: LAMP with only *one* Depth-First (DF) exploration of $\mathcal{H}$.

count testable patterns with DF exploration, starting with $\theta = 1$; **increase θ while exploring** if the curr. num. of frequent patterns $\geqslant \alpha/\hat{\psi}(\theta)$

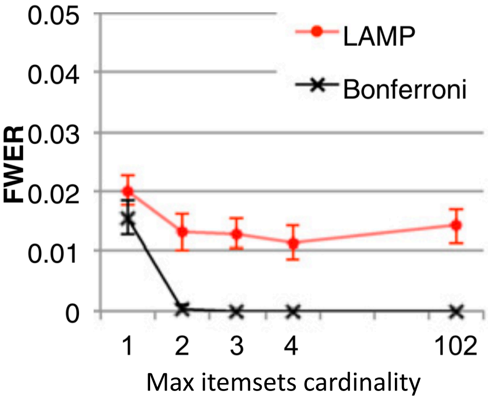


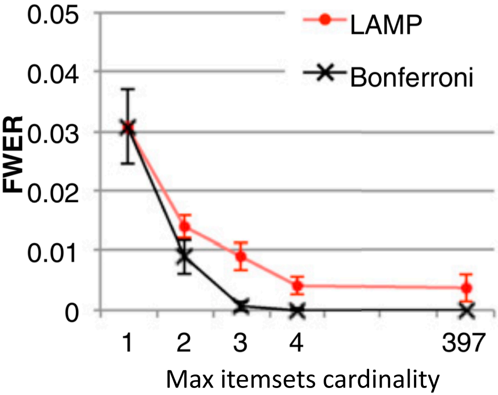
(imgs. from LAMP)

[3] Minato, S. I., Uno, T., Tsuda, K., Terada, A., Sese, J. *A fast method of statistical assessment for combinatorial hypotheses based on frequent itemset enumeration*. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2014)

LAMP: Experimental Results

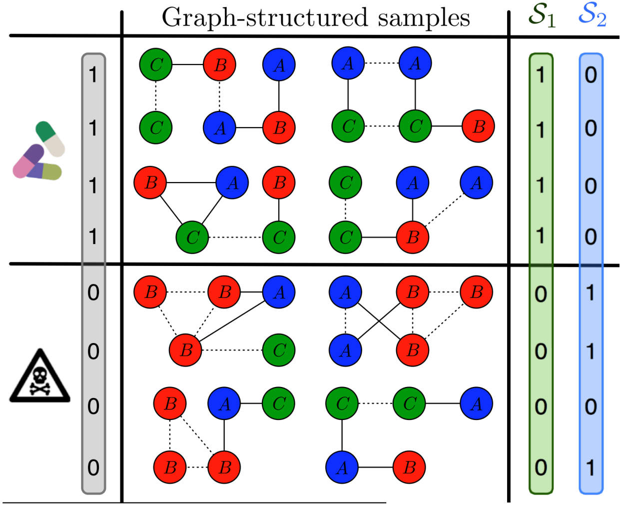
(imgs. from LAMP)

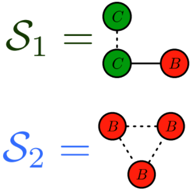




Estimated $FWER$ of LAMP vs Bonferroni correction.

Mining Significant Subgraphs[5]



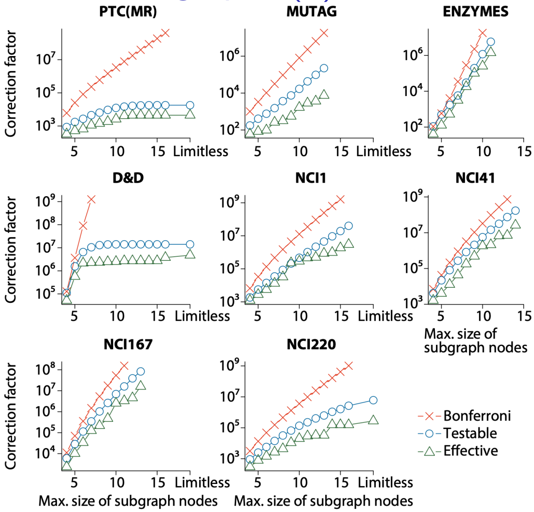


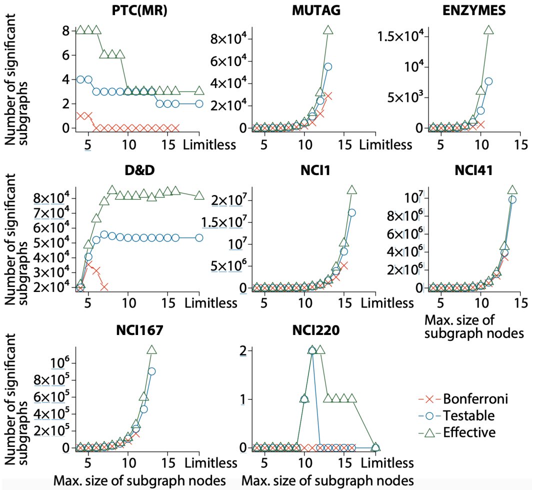
Goal: find induced subgraphs that are significantly enriched in a class of labelled graphs

(imgs. from [4])

[4] F. Llinares-López, D. Roqueiro, *Significant Pattern Mining for Biomarker Discovery*, ISMB18 Tutorial.
[5] M. Sugiyama, F. Llinares-López, N. Kasenburg, K.M. Borgwardt. *Significant subgraph mining with multiple testing correction.* In Proceedings of the International Conference on Data Mining, (2015).

LAMP for subgraphs (2)





From M. Sugiyama,F. Llinares-López, N. Kasenburg, K. M. Borgwardt. *Significant subgraph mining with multiple testing correction.* In Proc. of ICDM (2015).

Outline

Relaxing conditional assumptions

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Recap: Assumptions of Fisher's test: **all marginals of all the tested contingency tables are fixed** by design of the experiment. **Validity of this assumption depends on how the data is collected!**

Relaxing conditional assumptions

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Recap: Assumptions of Fisher's test: **all marginals of all the tested contingency tables are fixed** by design of the experiment. **Validity of this assumption depends on how the data is collected!**

In many cases, *only* n_0, n_1, and n are fixed, while $\sigma(\mathcal{S})$ depends on the data → **Unconditional Test!**

Relaxing conditional assumptions

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Recap: Assumptions of Fisher's test: all marginals of all the tested contingency tables are fixed by design of the experiment.
Validity of this assumption depends on how the data is collected!
In many cases, *only* n_0, n_1, and n are fixed, while $\sigma(\mathcal{S})$ depends on the data → **Unconditional Test!**

Not used in practice, mainly for computational reasons...
Until today ☺

Recap: Barnard's Exact Test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Nuisance variables: $\pi_{\mathcal{S},j} = P(\text{“}\mathcal{S} \subseteq t_i\text{”} \mid \text{“}\ell(t_i) = c_j\text{”})$,

NH: $\pi_{\mathcal{S},0} = \pi_{\mathcal{S},1} = \pi_{\mathcal{S}} = P(\text{“}\mathcal{S} \subseteq t_i\text{”})$.

Recap: Barnard's Exact Test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Nuisance variables: $\pi_{\mathcal{S},j} = P(\text{“}\mathcal{S} \subseteq t_i\text{”} \mid \text{“}\ell(t_i) = c_j\text{”})$,

NH: $\pi_{\mathcal{S},0} = \pi_{\mathcal{S},1} = \pi_{\mathcal{S}} = P(\text{“}\mathcal{S} \subseteq t_i\text{”})$. Let $a = \sigma(\mathcal{S})$, $b = \sigma_1(\mathcal{S})$:

Recap: Barnard's Exact Test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Nuisance variables: $\pi_{\mathcal{S},j} = P(\text{``}\mathcal{S} \subseteq t_i\text{''} \mid \text{``}\ell(t_i) = c_j\text{''})$,

NH: $\pi_{\mathcal{S},0} = \pi_{\mathcal{S},1} = \pi_{\mathcal{S}} = P(\text{``}\mathcal{S} \subseteq t_i\text{''})$. Let $a = \sigma(\mathcal{S})$, $b = \sigma_1(\mathcal{S})$:

$$P(a, b \mid \pi) = \binom{n_1}{b}\binom{n - n_1}{a - b}(\pi)^a (1 - \pi)^{n-a}$$

$$T(a, b, \pi) = \{(x, y) : P(x, y \mid \pi) \leqslant P(a, b \mid \pi)\}$$

$$\phi(a, b, \pi) = \sum_{(x,y)\in T(a,b,\pi)} P(x, y \mid \pi)$$

p-**value:** $p(a, b) = \max_{\pi} \{\phi(a, b, \pi)\}$

Recap: Barnard's Exact Test

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Nuisance variables: $\pi_{\mathcal{S},j} = P(\text{“}\mathcal{S} \subseteq t_i\text{”} \mid \text{“}\ell(t_i) = c_j\text{”})$,

NH: $\pi_{\mathcal{S},0} = \pi_{\mathcal{S},1} = \pi_{\mathcal{S}} = P(\text{“}\mathcal{S} \subseteq t_i\text{”})$. Let $a = \sigma(\mathcal{S})$, $b = \sigma_1(\mathcal{S})$:

$$P(a, b \mid \pi) = \binom{n_1}{b}\binom{n - n_1}{a - b}(\pi)^a (1 - \pi)^{n-a}$$

$$T(a, b, \pi) = \{(x, y) : P(x, y \mid \pi) \leqslant P(a, b \mid \pi)\}$$

$$\phi(a, b, \pi) = \sum_{(x,y) \in T(a,b,\pi)} P(x, y \mid \pi)$$

p-**value:** $p(a, b) = \max_{\pi} \{\phi(a, b, \pi)\}$ $\rightarrow$ **hard to compute!**

Efficient Unconditional Testing: SPuManTE! [6]

(Poster #146 on Tuesday!)

[6] L. Pellegrina, M. Riondato, and F. Vandin. *"SPuManTE: Significant Pattern Mining with Unconditional Testing"*. KDD 2019.

SPuManTE (1)

1) Computes **confidence intervals** $C_j(\mathcal{S})$ for $\pi_{\mathcal{S},j} = P(\text{“}\mathcal{S} \subseteq t_i\text{”} \mid \text{“}\ell(t_i) = c_j\text{”})$;

SPuManTE (1)

1) Computes **confidence intervals** $C_j(\mathcal{S})$ for $\pi_{\mathcal{S},j} = P(\text{“}\mathcal{S} \subseteq t_i\text{”} \mid \text{“}\ell(t_i) = c_j\text{”})$;
How? Compute an upper bound, for all $j \in \{0, 1\}$, on

$$\sup_{\mathcal{S}\in\mathcal{H}} \left| \pi_{\mathcal{S},j} - \frac{\sigma_j(\mathcal{S})}{n_j} \right|$$

(note: $\sigma_j(\mathcal{S})/n_j$ is **observed** from $\mathcal{D}$, $\pi_{\mathcal{S},j}$ is **unknown**)
with probability $\geqslant 1 - \delta$ ($\delta \leqslant \alpha$ for $FWER$ control),

SPuManTE (1)

1) Computes **confidence intervals** $C_j(\mathcal{S})$ for $\pi_{\mathcal{S},j} = P(\text{``}\mathcal{S} \subseteq t_i\text{''} \mid \text{``}\ell(t_i) = c_j\text{''})$;
How? Compute an upper bound, for all $j \in \{0, 1\}$, on

$$\sup_{\mathcal{S} \in \mathcal{H}} \left| \pi_{\mathcal{S},j} - \frac{\sigma_j(\mathcal{S})}{n_j} \right|$$

(note: $\sigma_j(\mathcal{S})/n_j$ is **observed** from $\mathcal{D}$, $\pi_{\mathcal{S},j}$ is **unknown**) with probability $\geqslant 1 - \delta$ ($\delta \leqslant \alpha$ for $FWER$ control), by upper bounding[7] the **Rademacher Complexity** of $\mathcal{H}$. **No assumptions on the input distribution: only information from $\mathcal{D}$!**

[7] M. Riondato and E. Upfal. *Mining frequent itemsets through progressive sampling with Rademacher averages.* KDD 2015.

`SPuManTE` (2)

2) Defines `UT`, an Unconditional Test that conditions (☺) on the event $E_{\mathcal{S}} =$ "$C_0(\mathcal{S}) \cap C_1(\mathcal{S}) = C(\mathcal{S}) = \emptyset$".

SPuManTE (2)

2) Defines UT, an Unconditional Test that conditions (☺) on the event $E_{\mathcal{S}} =$ "$C_0(\mathcal{S}) \cap C_1(\mathcal{S}) = C(\mathcal{S}) = \emptyset$".

p-value p_S according to UT:

$$p_S = \begin{cases} 0 & \text{, if } C(\mathcal{S}) = \emptyset \\ \max\{\phi(\sigma(S), \sigma_1(S), \pi), \pi \in C(\mathcal{S})\} & \text{, othw.} \end{cases}$$

`SPuManTE` (2)

2) Defines `UT`, an Unconditional Test that conditions (☺) on the event $E_{\mathcal{S}} = $ "$C_0(\mathcal{S}) \cap C_1(\mathcal{S}) = C(\mathcal{S}) = \varnothing$".

p-value p_S according to `UT`:

$$p_S = \begin{cases} 0 & \text{, if } C(\mathcal{S}) = \varnothing \\ \max\{\phi(\sigma(S), \sigma_1(S), \pi), \pi \in C(\mathcal{S})\} & \text{, othw.} \end{cases}$$

$\rightarrow$ A pattern is flagged as significant if

$$C(\mathcal{S}) = \varnothing.$$

The confidence of the validity of $C(\mathcal{S})$ provides $FWER$ control.

SPuManTE (3)

p-value p_S according to UT:

$$p_S = \begin{cases} 0 & \text{, if } C(\mathcal{S}) = \emptyset \\ \max\{\phi(\sigma(S), \sigma_1(S), \pi), \pi \in C(\mathcal{S})\} & \text{, othw.} \end{cases}$$

Case $C(\mathcal{S}) \neq \emptyset$: still hard to compute! 

SPuManTE (3)

p-value p_S according to UT:

$$p_S = \begin{cases} 0 & \text{, if } C(\mathcal{S}) = \emptyset \\ \max\{\phi(\sigma(S), \sigma_1(S), \pi), \pi \in C(\mathcal{S})\} & \text{, othw.} \end{cases}$$

Case $C(\mathcal{S}) \neq \emptyset$: still hard to compute! 

3) **Upper and Lower bounds** to p_S, and **efficient algorithms to compute them** $\rightarrow$ requirements to combine UT with LAMP.

SPuManTE (4)

Let
$$\overline{\pi}_{\mathcal{S}} = \frac{\sigma(\mathcal{S})}{n}.$$

Lower bound $\check{p}_{\mathcal{S}}$ to p-value $p_{\mathcal{S}}$:

$$\check{p}_{\mathcal{S}} = \begin{cases} 0 & \text{, if } C(\mathcal{S}) = \emptyset \\ \phi(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}), \overline{\pi}_{\mathcal{S}}) & \text{, othw.} \end{cases}$$

SPuManTE (4)

Let
$$\overline{\pi}_{\mathcal{S}} = \frac{\sigma(\mathcal{S})}{n}.$$

Lower bound $\check{p}_{\mathcal{S}}$ to p-value $p_{\mathcal{S}}$:

$$\check{p}_{\mathcal{S}} = \begin{cases} 0 & \text{, if } C(\mathcal{S}) = \emptyset \\ \phi(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}), \overline{\pi}_{\mathcal{S}}) & \text{, othw.} \end{cases}$$

Compute $\phi(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}), \overline{\pi}_{\mathcal{S}})$ **efficiently? Yes!** ☺
(For more details: paper or come to talk to #146 poster! ☺)

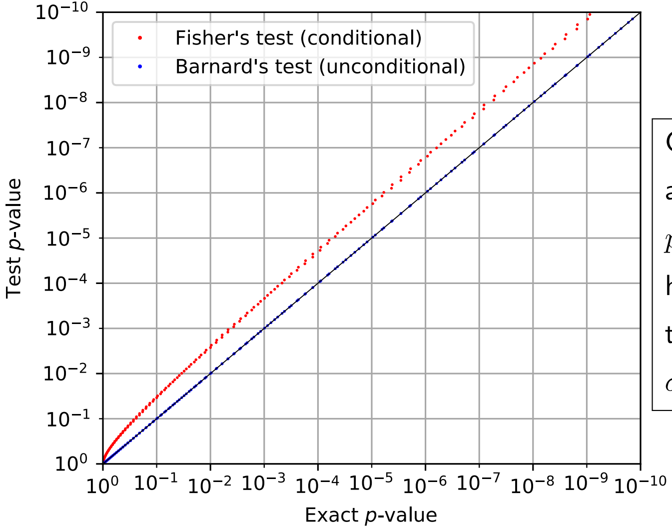
Upper bound $\widehat{p}_{\mathcal{S}}$ to p-value $p_{\mathcal{S}}$:

$$\widehat{p}_{\mathcal{S}} = P(\sigma(\mathcal{S}), \sigma_1(\mathcal{S}) \mid \overline{\pi})(n_0 + 1)(n_1 + 1).$$

Theorem

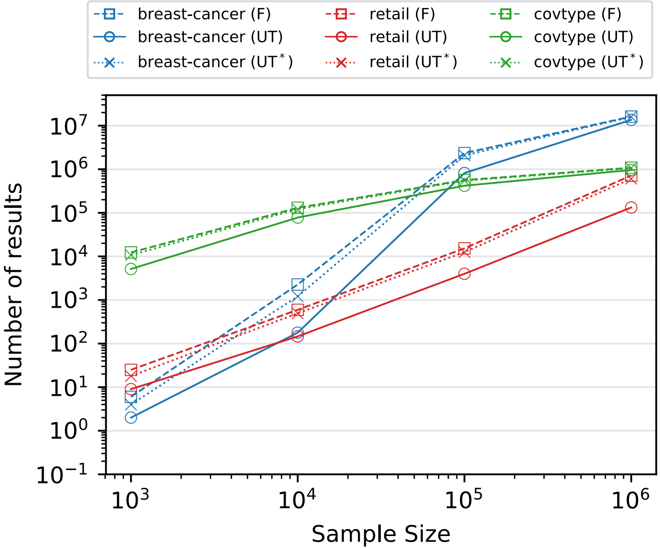
$p_{\mathcal{S}} \leqslant \widehat{p}_{\mathcal{S}}$.

SPuManTE: Experimental Results



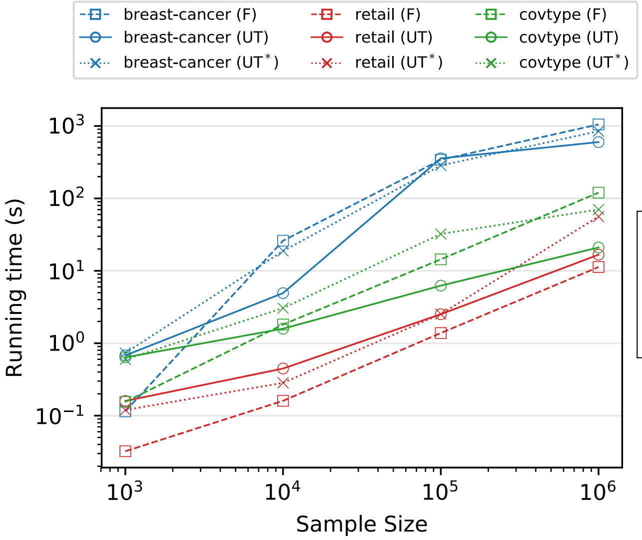
Comparison of p-values of Fisher's and Barnard's tests w.r.t. the exact p-value (under the unconditional null hypothesis) for all contingency tables with $n = 10^4$, $n_1 = 0.25 \cdot n$, $\sigma(\mathcal{S}) = 0.1 \cdot n$.

SPuManTE: Experimental Results



Comparison of number of significant patterns using Fisher's test (F), UT (upper bound $\widehat{p}_{\mathcal{S}}$ to p-values), UT* (lower bound $\check{p}_{\mathcal{S}}$ to p-values). **Additional results**: may not be well supported by the data!

SPuManTE: Experimental Results



Running times of LAMP with Fisher's test (F), SPuManTE using UT and UT*.
SPuManTE: **very efficient!**

Permutation Testing

Main idea: *estimate* the null distribution by *randomly perturbing* the observed data.

Pro: takes advantage of the dependence structure of the hypothesis

Cons: computationally expensive and formally imprecise

Settings

$\mathcal{D}_0$: observed dataset as a *binary matrix*.

E.g., a transactional dataset
(rows: transactions: columns: items)

$$\begin{matrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{matrix}$$

$T_0 = \mathcal{A}(\mathcal{D}_0) \in \mathbb{R}$: output of analysis algorithm $\mathcal{A}$ on $\mathcal{D}_0$.

E.g., the *number* of frequent itemsets w.r.t. min. freq. thresh. θ.

Settings

$\mathcal{D}_0$: observed dataset as a *binary matrix*.

E.g., a transactional dataset
(rows: transactions: columns: items)

3	1	3	2	
1	0	1	1	3
0	1	1	0	2
1	0	1	0	2
1	0	0	1	2

$T_0 = \mathcal{A}(\mathcal{D}_0) \in \mathbb{R}$: output of analysis algorithm $\mathcal{A}$ on $\mathcal{D}_0$.

E.g., the *number* of frequent itemsets w.r.t. min. freq. thresh. θ.

$\mathbf{P}$: a set of properties of $\mathcal{D}_0$ considered important, characteristics.

E.g., the rows and columns *totals*

Settings

$\mathcal{D}_0$: observed dataset as a *binary matrix*.

E.g., a transactional dataset (rows: transactions: columns: items)

3	1	3	2	
1	0	1	1	3
0	1	1	0	2
1	0	1	0	2
1	0	0	1	2

$T_0 = \mathcal{A}(\mathcal{D}_0) \in \mathbb{R}$: output of analysis algorithm $\mathcal{A}$ on $\mathcal{D}_0$.

E.g., the *number* of frequent itemsets w.r.t. min. freq. thresh. θ.

$\mathbf{P}$: a set of properties of $\mathcal{D}_0$ considered important, characteristics.

E.g., the rows and columns *totals*

QUESTION: Is T_0 a "*consequence*" of $\mathbf{P}$?

Null hypothesis

Null hypothesis H_0: T_0 is fully explained by $\mathbf{P}$.

Null hypothesis

Null hypothesis H_0: T_0 is fully explained by $\mathbf{P}$.

I.e., a value of T_0 is "*typical*" for datasets *satisfying* $\mathbf{P}$.

I.e., it is *very likely* to observe a value $\mathcal{A}(\mathcal{D})$ *close to* T_0 in a dataset $\mathcal{D}$ satisfying $\mathbf{P}$.

Null hypothesis

Null hypothesis H_0: T_0 is fully explained by $\mathbf{P}$.

I.e., a value of T_0 is "*typical*" for datasets *satisfying* $\mathbf{P}$.

I.e., it is *very likely* to observe a value $\mathcal{A}(\mathcal{D})$ *close to* T_0 in a dataset $\mathcal{D}$ satisfying $\mathbf{P}$.

I.e., let $\mathbb{D}_\mathbf{P}$: set of datasets satisfying $\mathbf{P}$, then

$$Q(T_0) = \min\left\{\Pr_{\mathcal{U}}\left(\mathcal{A}(\mathcal{D}) \geqslant T_0\right), \Pr_{\mathcal{U}}\left(\mathcal{A}(\mathcal{D}) < T_0\right)\right\} \gg 0,$$

$\mathcal{U}$: *uniform* distribution over $\mathbb{D}_\mathbf{P}$.

Null distribution

To test H_0, we need a *quantitative* approach:

For $\alpha \in (0, 1)$, if $Q(T_0) < \alpha$ then reject H_0.

Null distribution

To test H_0, we need a *quantitative* approach:

For $\alpha \in (0,1)$, if $Q(T_0) < \alpha$ then reject H_0.

Null distribution $\Theta = \Theta(\mathcal{A}, \mathbf{P})$ over values of $T = \mathcal{A}(\mathcal{D})$, $\mathcal{D} \in \mathbb{D}_{\mathbf{P}}$.

Θ has *c.d.f.*

$$\theta(v) = \Pr_{\mathcal{U}}\left(T = \mathcal{A}(\mathcal{D}) \geqslant v\right) = \frac{\left|\left\{\mathcal{D} \in \mathbb{D}_{\mathbf{P}} \ : \ T = \mathcal{A}(\mathcal{D}) \geqslant v\right\}\right|}{\left|\mathbb{D}_{\mathbf{P}}\right|}$$

Null distribution

To test H_0, we need a *quantitative* approach:

For $\alpha \in (0,1)$, if $Q(T_0) < \alpha$ then reject H_0.

Null distribution $\Theta = \Theta(\mathcal{A}, \mathbf{P})$ over values of $T = \mathcal{A}(\mathcal{D})$, $\mathcal{D} \in \mathbb{D}_{\mathbf{P}}$.

Θ has *c.d.f.*

$$\theta(v) = \Pr_{\mathcal{U}}\left(T = \mathcal{A}(\mathcal{D}) \geqslant v\right) = \frac{\left|\{\mathcal{D} \in \mathbb{D}_{\mathbf{P}} \;:\; T = \mathcal{A}(\mathcal{D}) \geqslant v\}\right|}{|\mathbb{D}_{\mathbf{P}}|}$$

We can use $\theta(T_0)$ to test H_0:

if $\min\{\theta(T_0), 1 - \theta(T_)\} < \alpha$, reject H_0.

Null distribution

To test H_0, we need a *quantitative* approach:

For $\alpha \in (0,1)$, if $Q(T_0) < \alpha$ then reject H_0.

Null distribution $\Theta = \Theta(\mathcal{A}, \mathbf{P})$ over values of $T = \mathcal{A}(\mathcal{D})$, $\mathcal{D} \in \mathbb{D}_{\mathbf{P}}$.

Θ has *c.d.f.*

$$\theta(v) = \Pr_{\mathcal{U}}(T = \mathcal{A}(\mathcal{D}) \geqslant v) = \frac{|\{\mathcal{D} \in \mathbb{D}_{\mathbf{P}} \; : \; T = \mathcal{A}(\mathcal{D}) \geqslant v\}|}{|\mathbb{D}_{\mathbf{P}}|}$$

We can use $\theta(T_0)$ to test H_0:

if $\min\{\theta(T_0), 1 - \theta(T_)\} < \alpha$, reject H_0.

Issue: deriving θ is *infeasible* for most $(\mathcal{A}, \mathbf{P})$.

Empiricism to the rescue

ISSUE: deriving θ is infeasible for most $(\mathcal{A}, \mathbf{P})$.

SOLUTION: approximate θ using an *empirical c.d.f.* $\tilde{\theta}$.

Empiricism to the rescue

Issue: deriving θ is infeasible for most $(\mathcal{A}, \mathbf{P})$.

Solution: approximate θ using an *empirical c.d.f.* $\tilde{\theta}$.

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_k\} \subseteq \mathbb{D}_{\mathbf{P}}$ *independent uniform* samples.
2. Run $\mathcal{A}$ on each $\mathcal{D}_i \in \mathbf{D}$ to obtain $\mathbf{T} = \{T_1, \ldots, T_k\}$.
3. Compute an *empirical p-value* from the $\tilde{\theta}$ arising from $\mathbf{T}$:

$$\tilde{p} = \frac{1}{k+1}\left(\min\{|\{i \in [k] \mid T_i < T_0\}|, |\{i \in [k] \mid T_i > T_0\}|\} + 1\right) \in [0, 0.5]$$

4. If $\tilde{p} < \alpha$, reject H_0.

Why does it work?

It is a *consistent* approach:

As the number $k = |\mathbf{D}|$ of samples grows,

the empirical c.d.f. $\tilde{\theta}$ *converges* to θ,

thus, $\tilde{p}$ converges to the exact p-values.

WARNING: Convergence happens *in the limit*,

but there are *finite-sample deviation bounds* for $\tilde{\theta}$ from θ.

The crux of the matter

The steps again:

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_k\} \subseteq \mathbb{D}_{\mathbf{P}}$ *independent uniform* samples.
2. Run $\mathcal{A}$ on each $\mathcal{D}_i \in \mathbf{D}$ to obtain $\mathbf{T} = \{T_1, \ldots, T_k\}$.
3. Compute an *empirical* p*-value* from the $\tilde{\theta}$ arising from $\mathbf{T}$:

$$\tilde{p} = \frac{1}{k+1}\left(\min\{|\{i \in [k] \mid T_i < T_0\}|, |\{i \in [k] \mid T_i > T_0\}|\} + 1\right)$$

4. If $\tilde{p} < \alpha$, reject H_0.

The crux of the matter

The steps again:

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_k\} \subseteq \mathbb{D}_{\mathbf{P}}$ *independent uniform* samples.

2. Run $\mathcal{A}$ on each $\mathcal{D}_i \in \mathbf{D}$ to obtain $\mathbf{T} = \{T_1, \ldots, T_k\}$.

3. Compute an *empirical* p*-value* from the $\tilde{\theta}$ arising from $\mathbf{T}$:

$$\tilde{p} = \frac{1}{k+1}\left(\min\{|\{i \in [k] \mid T_i < T_0\}|, |\{i \in [k] \mid T_i > T_0\}|\} + 1\right)$$

4. If $\tilde{p} < \alpha$, reject H_0. **Easy**

The crux of the matter

The steps again:

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_k\} \subseteq \mathbb{D}_{\mathbf{P}}$ *independent uniform* samples.

2. Run $\mathcal{A}$ on each $\mathcal{D}_i \in \mathbf{D}$ to obtain $\mathbf{T} = \{T_1, \ldots, T_k\}$.

3. Compute an *empirical p-value* from the $\tilde{\theta}$ arising from $\mathbf{T}$: **Easy**

$$\tilde{p} = \frac{1}{k+1} \left(\min\{|\{i \in [k] \mid T_i < T_0\}|, |\{i \in [k] \mid T_i > T_0\}|\} + 1 \right)$$

4. If $\tilde{p} < \alpha$, reject H_0. **Easy**

The crux of the matter

The steps again:

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_k\} \subseteq \mathbb{D}_{\mathbf{P}}$ *independent uniform* samples.
2. Run $\mathcal{A}$ on each $\mathcal{D}_i \in \mathbf{D}$ to obtain $\mathbf{T} = \{T_1, \ldots, T_k\}$. **Easy**
3. Compute an *empirical* p*-value* from the $\tilde{\theta}$ arising from $\mathbf{T}$: **Easy**

$$\tilde{p} = \frac{1}{k+1}\left(\min\{|\{i \in [k] \mid T_i < T_0\}|, |\{i \in [k] \mid T_i > T_0\}|\} + 1\right)$$

4. If $\tilde{p} < \alpha$, reject H_0. **Easy**

The crux of the matter

The steps again:

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_k\} \subseteq \mathbb{D}_{\mathbf{P}}$ *independent uniform* samples. **How?**

2. Run $\mathcal{A}$ on each $\mathcal{D}_i \in \mathbf{D}$ to obtain $\mathbf{T} = \{T_1, \ldots, T_k\}$. **Easy**

3. Compute an *empirical* p*-value* from the $\tilde{\theta}$ arising from $\mathbf{T}$: **Easy**

$$\tilde{p} = \frac{1}{k+1}\left(\min\{|\{i \in [k] \mid T_i < T_0\}|, |\{i \in [k] \mid T_i > T_0\}|\} + 1\right)$$

4. If $\tilde{p} < \alpha$, reject H_0. **Easy**

Perturbing the data

Assumption: there exists a *perturbation operation*

$$\phi : \mathbb{D}_{\mathbf{P}} \times \underbrace{\mathcal{Y}}_{\text{parameters}} \to \mathbb{D}_{\mathbf{P}}$$

s.t. for any $\mathcal{D}', \mathcal{D}'' \in \mathbb{D}_{\mathbf{P}}$, $\mathcal{D}'$ can be obtained by repeatedly applying ϕ to $\mathcal{D}''$.

I.e., there exists a finite sequence $Y_1, \ldots, Y_\ell$, $Y_i \in \mathcal{Y}$ s.t.

$$\mathcal{D}'' = \phi(\phi(\phi(\cdots(\phi(\mathcal{D}'', Y_1), Y_2), \cdots), Y_\ell))$$

If $\mathcal{D}'' = \phi(\mathcal{D}', y)$, then there exists $y^{-1} \in Y$ s.t. $\mathcal{D}' = \phi(\mathcal{D}'', y^{-1})$.

Example: perturbation for rows and columns sums

1. Take two rows u and v and two columns A and B of $\mathcal{D}_0$ such that $u(A) = v(B) = 1$ and $u(B) = v(A) = 0$;
2. Change the rows so that $u(B) = v(A) = 1$ and $u(A) = v(B) = 0$

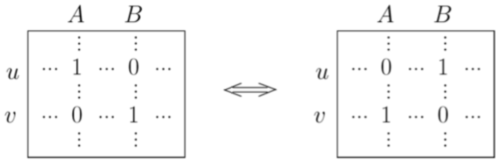


Fig. 1. A swap in a 0–1 matrix.

From Gionis et al., Assessing Data Mining Results via Swap Randomization, ACM TKDD, 2007.

$\mathcal{Y}$ is the set of quadruples of two rows and two columns indices.

Generating the samples

$G = (\mathbb{D}_{\mathbf{P}}, E)$: directed graph s.t. $(\mathcal{D}, \mathcal{D}') \in E$ if $\mathcal{D}'$ can be obtained from $\mathcal{D}$ with *one* perturbation:

$$(\mathcal{D}, \mathcal{D}') \in E \Leftrightarrow \exists y \in \mathcal{Y} \text{ s.t. } \mathcal{D}' = \phi(\mathcal{D}, y)$$

Add *self-loops* and run *Metropolis-Hastings* on the resulting graph G' to obtain *independent and uniform* samples.

Running Metropolis-Hastings

M-H performs a *random walk* on G' with *uniform stationary distribution*.

For each (visited) $\mathcal{D}$, M-H needs its *neighbors*

$$\mathsf{N}(\mathcal{D}) = \{\mathcal{D}' \in \mathbb{D}_{\mathbf{P}} \ : \ \exists y \in \mathcal{Y} \text{ s.t. } \mathcal{D}' = \phi(\mathcal{D}, y)\}$$

Computing $\mathsf{N}(\mathcal{D})$ requires to find all quadruplets $(u, v, A, B) \in \mathcal{Y}$ leading to valid perturbations from $\mathcal{D}$.

Gionis et al. show how to get $\mathsf{N}(\mathcal{D})$ in *expected* constant time when no row/column has too many 1s.

Mixing Time

The samples $\mathcal{D}_1, \ldots, \mathcal{D}_k$ must be *independent* and *uniform*

M-H must make at least M moves after taking each sample

M: *mixing time* of G' with M-H transition probabilities.

Mixing Time

The samples $\mathcal{D}_1, \ldots, \mathcal{D}_k$ must be *independent* and *uniform*

M-H must make at least M moves after taking each sample

M: *mixing time* of G' with M-H transition probabilities.

Deriving M is usually infeasible

so M is fixed to be "large enough" after experimentation.

Advantages and disadvantages of permutation testing

Conceptually very natural

Requires a perturbation operation ϕ for $\mathbf{P}$

Computationally very expensive:

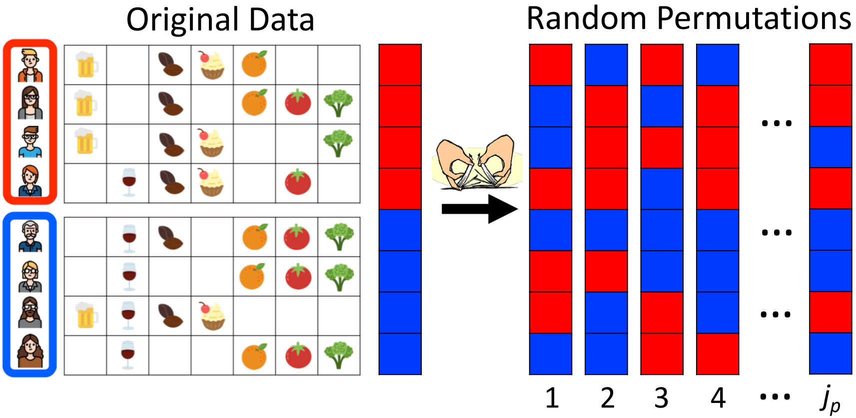
sample generation $+$ running $\mathcal{A}$ on each sample

"Empirical everything": p-value, independence, uniformity, ...

Outline

Westfall-Young ($\mathtt{WY}^8$) Permutation Testing

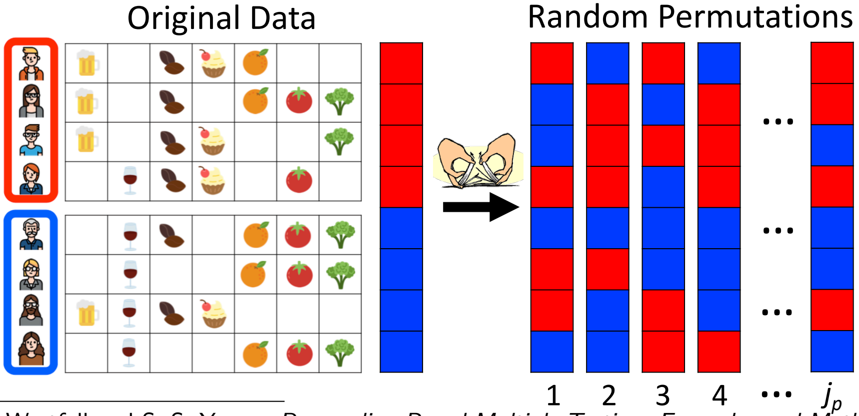
Randomly shuffle the labels; compute patterns' p-values w.r.t. the random labels.



[8] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley-Interscience, 1993.

Westfall-Young (WY[9]) Permutation Testing

Any association found on the random permutations is a **false positive**: directly estimate the p-values from the **null hypothesis joint distribution** → **account for dependencies of hypotheses**



[9] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley-Interscience, 1993.

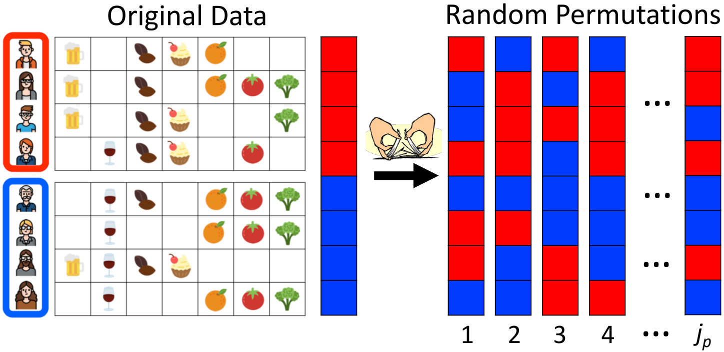
WY Permutation Testing: formally

$$\ell_j(t_i) = j\text{-th permuted label of } t_i \quad , \quad \sigma_1^j(\mathcal{S}) = \sum_{i=1}^{n} \phi_{\mathcal{S}}(t_i) \mathbb{1}\left[\ell_j(t_i) = c_1\right]$$

WY Permutation Testing: formally

$$\ell_j(t_i) = j\text{-th permuted label of } t_i \quad , \quad \sigma_1^j(\mathcal{S}) = \sum_{i=1}^{n} \phi_{\mathcal{S}}(t_i)\mathbb{1}\left[\ell_j(t_i) = c_1\right]$$

Example:



$\mathcal{S} = \{\text{broccoli}\}$
$\sigma_1^1(\mathcal{S}) = 1,$
$\sigma_1^2(\mathcal{S}) = 3,$
$\sigma_1^3(\mathcal{S}) = 2,$
$\ldots$

WY Permutation Testing: formally

$$\ell_j(t_i) = j\text{-th permuted label of } t_i \quad , \quad \sigma_1^j(\mathcal{S}) = \sum_{i=1}^{n} \phi_{\mathcal{S}}(t_i) \mathbb{1}\left[\ell_j(t_i) = c_1\right]$$

$$p_{\min}^j = \min_{\mathcal{S}\in\mathcal{H}} \left\{ p(\sigma(\mathcal{S}), \sigma_1^j(\mathcal{S})) \right\} \quad , \quad \overline{FWER}(x) = \frac{1}{j_p} \sum_{i=1}^{j_p} \mathbb{1}\left[p_{\min}^j \leqslant x\right]$$

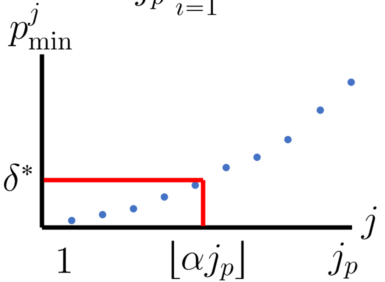
WY Permutation Testing: formally

$$\ell_j(t_i) = j\text{-th permuted label of } t_i \quad , \quad \sigma_1^j(\mathcal{S}) = \sum_{i=1}^{n} \phi_{\mathcal{S}}(t_i) \mathbb{1}\left[\ell_j(t_i) = c_1\right]$$

$$p_{\min}^j = \min_{\mathcal{S} \in \mathcal{H}} \left\{ p(\sigma(\mathcal{S}), \sigma_1^j(\mathcal{S})) \right\} \quad , \quad \overline{FWER}(x) = \frac{1}{j_p} \sum_{i=1}^{j_p} \mathbb{1}\left[p_{\min}^j \leqslant x\right]$$

Compute $\delta^* = \max\left\{x : \overline{FWER}(x) \leqslant \alpha\right\}$

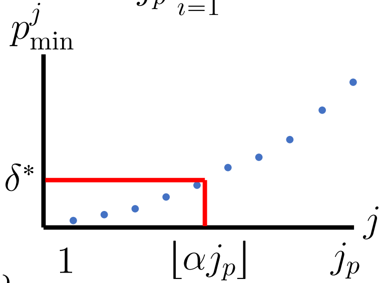
$(j_p \sim 10^3\text{-}10^4$ for $\alpha \sim 0.05)$



WY Permutation Testing: formally

$$\ell_j(t_i) = j\text{-th permuted label of } t_i \quad , \quad \sigma_1^j(\mathcal{S}) = \sum_{i=1}^{n} \phi_{\mathcal{S}}(t_i) \mathbb{1}\left[\ell_j(t_i) = c_1\right]$$

$$p_{\min}^j = \min_{\mathcal{S}\in\mathcal{H}} \left\{ p(\sigma(\mathcal{S}), \sigma_1^j(\mathcal{S})) \right\} \quad , \quad \overline{FWER}(x) = \frac{1}{j_p} \sum_{i=1}^{j_p} \mathbb{1}\left[p_{\min}^j \leqslant x \right]$$



Compute $\delta^* = \max\left\{x : \overline{FWER}(x) \leqslant \alpha\right\}$

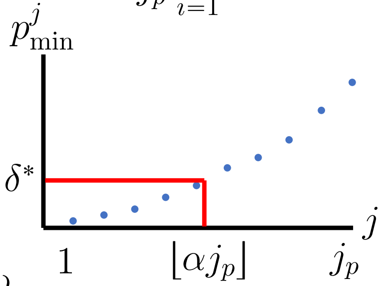
$(j_p \sim 10^3\text{-}10^4$ for $\alpha \sim 0.05)$

Output $\{\mathcal{S} : p_{\mathcal{S}} \leqslant \delta^*\}$.

WY Permutation Testing: formally

$$\ell_j(t_i) = j\text{-th permuted label of } t_i \quad , \quad \sigma_1^j(\mathcal{S}) = \sum_{i=1}^{n} \phi_{\mathcal{S}}(t_i) \mathbb{1}\left[\ell_j(t_i) = c_1\right]$$

$$p_{\min}^j = \min_{\mathcal{S} \in \mathcal{H}} \left\{ p(\sigma(\mathcal{S}), \sigma_1^j(\mathcal{S})) \right\} \quad , \quad \overline{FWER}(x) = \frac{1}{j_p} \sum_{i=1}^{j_p} \mathbb{1}\left[p_{\min}^j \leqslant x\right]$$



Compute $\delta^* = \max\left\{x : \overline{FWER}(x) \leqslant \alpha\right\}$

$(j_p \sim 10^3\text{-}10^4$ for $\alpha \sim 0.05)$

Output $\{\mathcal{S} : p_{\mathcal{S}} \leqslant \delta^*\}$.

Problem: **exhaustive enumeration** of $\mathcal{H}$ to compute $p_{\min}^j$.

Computing $p^j_{\min}$: `FASTWY`

How to compute $p^j_{\min}$ efficiently?

Computing $p^j_{\min}$: `FASTWY`

How to compute $p^j_{\min}$ efficiently?

Tarone saves us again ☺

`FASTWY`[10]: **Intuition:**

$$\hat{\psi}(\mathcal{S}) \geqslant p^j_{\min} \;\Rightarrow\; p\left(\sigma(\mathcal{S}), \sigma^j_1(\mathcal{S})\right) \geqslant p^j_{\min}$$

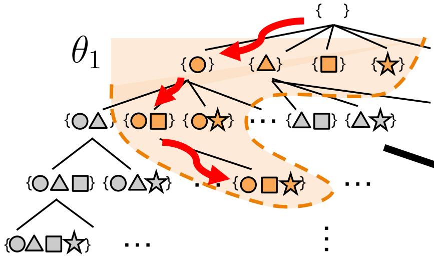
Pattern $\mathcal{S}$ is *untestable* $\Rightarrow$ cannot improve $p^j_{\min}$!

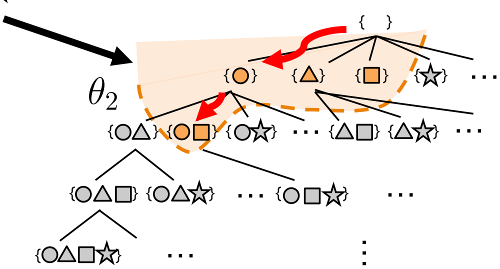
[10] A. Terada, K. Tsuda, and J. Sese. *Fast westfall-young permutation procedure for combinatorial regulation discovery*. In IEEE International Conference on Bioinformatics and Biomedicine, 2013.

Computing $p^j_{\min}$: FASTWY

(improved version[11] of) FASTWY: computes efficiently $p^j_{\min}$ with a **branch-and-bound search** over $\mathcal{H}$, pruning subtrees with $\hat{\psi}(\cdot)$:

start with $\theta = 1$ and $p^j_{\min} = 1$; explore patterns with DF exploration, updating $p^j_{\min}$; **increase θ while exploring** if $p^j_{\min} \leqslant \hat{\psi}(\theta)$





(imgs. from LAMP)

[11] T. Aika, H. Kim, and J. Sese. *High-speed westfall-young permutation procedure for genome-wide association studies*, ACM-BCB 2015.

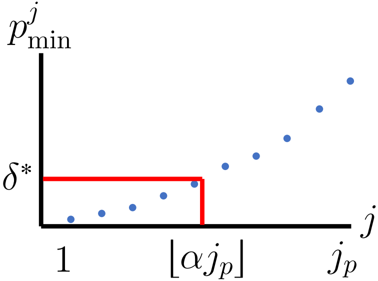
FASTWY

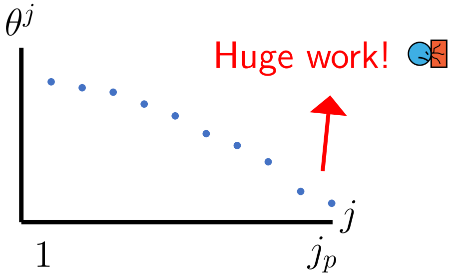
Issues of FASTWY:

1) repeat the procedure j_p times ($j_p \sim 10^3$-10^4);

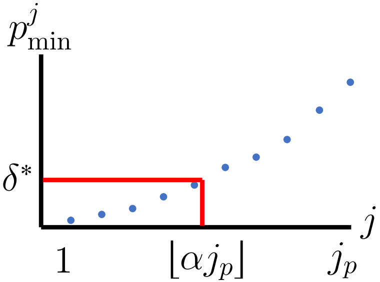
2) for some $j \in [1, j_p]$:

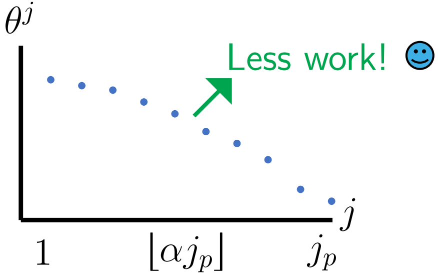
$p^j_{\min}$ **may not be very small** $\rightarrow \theta^j$ **very small** $\rightarrow$ **impractically large number of hypotheses to explore.**





`WYlight`[12]: **Intuition:** to find δ^* we only need to **compute exactly the lower α-quantile of** $\{p_{\min}^j\}_{j=1}^{j_p}$.



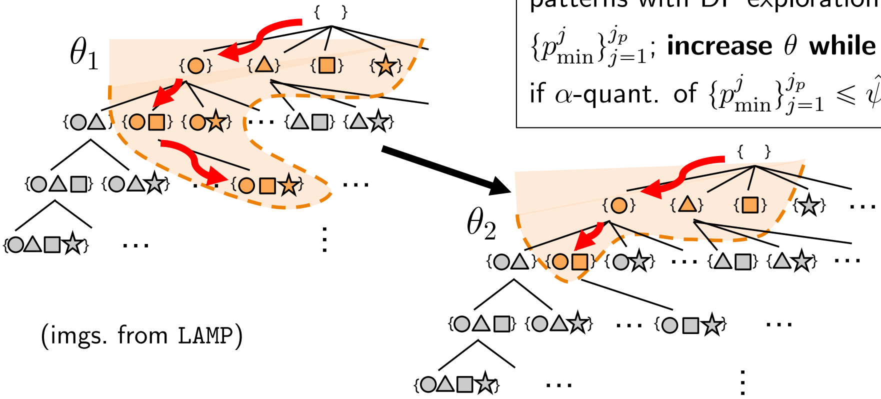


[12] F. Llinares-López, M. Sugiyama, L. Papaxanthos, and K. Borgwardt. *Fast and memory-efficient significant pattern mining via permutation testing*, KDD 2015.

WYlight

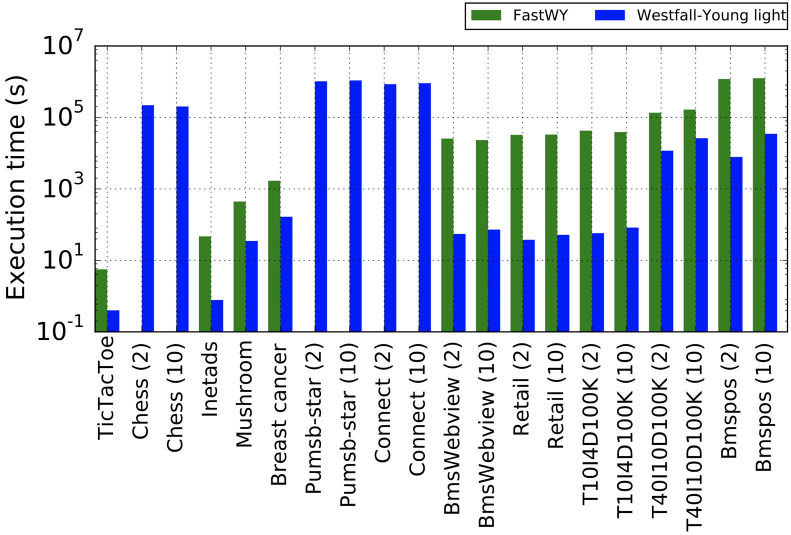
`WYlight` **algorithm**: one DF exploration of $\mathcal{H}$ processing all j_p permutations at once.

start with $\theta = 1$ and $p^j_{\min} = 1, \forall j$; explore patterns with DF exploration, updating $\{p^j_{\min}\}_{j=1}^{j_p}$; **increase θ while exploring** if α-quant. of $\{p^j_{\min}\}_{j=1}^{j_p} \leqslant \hat{\psi}(\theta)$



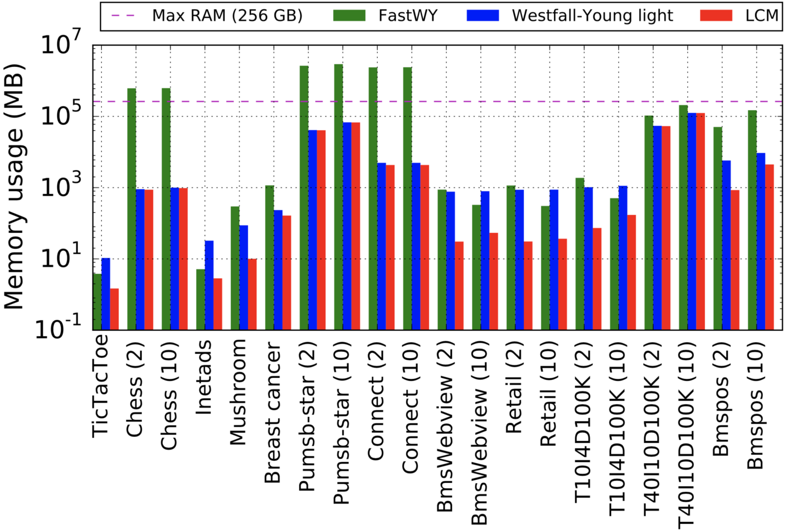
(imgs. from `LAMP`)

WYlight[13] - Running time



[13] F. Llinares-López, M. Sugiyama, L. Papaxanthos, and K. Borgwardt. *Fast and memory-efficient significant pattern mining via permutation testing*, KDD 2015.

WYlight[14] - Memory



[14] F. Llinares-López, M. Sugiyama, L. Papaxanthos, and K. Borgwardt. *Fast and memory-efficient significant pattern mining via permutation testing*, KDD 2015.

Too many results!

Motivation: for many datasets, impractically large set of results ($SP(0.05)$) are found even when controlling $FWER \leqslant 0.05$:

dataset	$\|D\|$	$\|I\|$	avg	n_1/n	$SP(0.05)$
svmguide3(L)	1,243	44	21.9	0.23	36,736
chess(U)	3,196	75	37	0.05	$> 10^7$
mushroom(L)	8,124	118	22	0.48	71,945
phishing(L)	11,055	813	43	0.44	$> 10^7$
breast cancer(L)	12,773	1,129	6.7	0.09	6
a9a(L)	32,561	247	13.9	0.24	348,611
pumb-star(U)	49,046	7117	50.5	0.44	$> 10^7$
bms-web1(U)	58,136	60,978	2.51	0.03	704,685
connect(U)	67,557	129	43	0.49	$> 10^8$
bms-web2(U)	77,158	330,285	4.59	0.04	289,012
retail(U)	88,162	16,470	10.3	0.47	3,071
ijcnn1(L)	91,701	44	13	0.10	607,373
T10I4D100K(U)	100,000	870	10.1	0.08	3,819
T40I10D100K(U)	100,000	942	39.6	0.28	5,986,439
codrna(L)	271,617	16	8	0.33	4,088
accidents(U)	340,183	467	33.8	0.49	$> 10^7$
bms-pos(U)	515,597	1,656	6.5	0.40	26,366,131
covtype(L)	581,012	64	11.9	0.49	542,365
susy(U)	5,000,000	190	43	0.48	$> 10^7$

TopKWY

What if we want (more efficiently!) only the **top-k significant patterns**, retaining the guarantees of WY procedure? $\rightarrow$ TopKWY[15]!

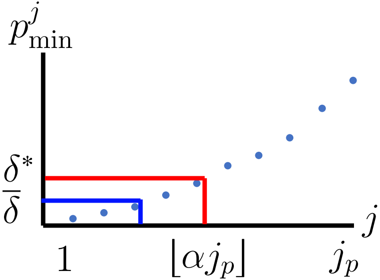
$p^k = k$-th smallest element of $\{p_{\mathcal{S}} : \mathcal{S} \in \mathcal{H}\}$,
$\delta^* = \max\left\{x : \overline{FWER}(x) \leqslant \alpha\right\}$,
$\overline{\delta} = \min\left\{p^k, \delta\right\}$.

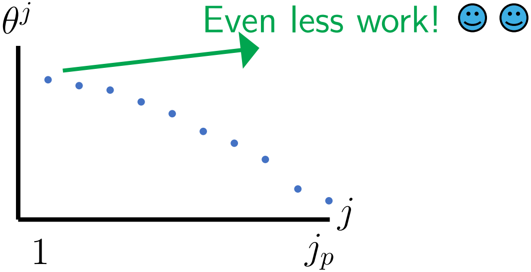
Set of **top-k significant patterns**:

$$TOPKSP(\mathcal{D}, \mathcal{H}, \alpha, k) := \left\{\mathcal{S} : p_{\mathcal{S}} \leqslant \overline{\delta}\right\}.$$

[15] L. Pellegrina and F. Vandin. *Efficient mining of the most significant patterns with permutation testing*. KDD 2018.

Intuition: to compute $TOPKSP(\mathcal{D}, \mathcal{H}, \alpha, k)$ we only need to compute exactly the values of the set $\left\{p_{\min}^{j}\right\}_{j=1}^{j_p}$ that are $\leqslant \overline{\delta}$.

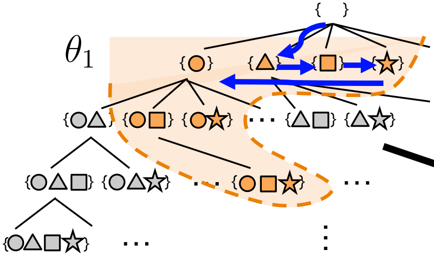


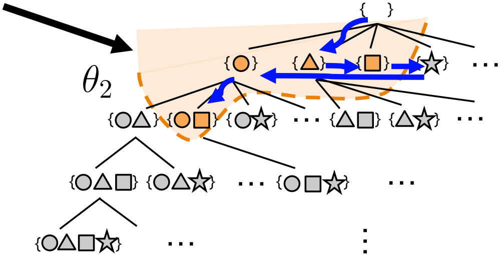


Algorithm: Best First (BF) exploration of $\mathcal{H}$ to compute $\overline{\delta}$.

(Approach similar to `TopKMiner` for **top-k frequent itemsets**).

start with $\theta = 1$ and $p^j_{\min} = 1, \forall j$; explore patterns with BF exploration, updating $\{p^j_{\min}\}_{j=1}^{j_p}$ and p^k; **increase θ while exploring** if $\min\left\{\alpha\text{-quant. of } \{p^j_{\min}\}_{j=1}^{j_p},\ p^k\right\} \leqslant \hat{\psi}(\theta)$





(imgs. from `LAMP`)

TopKWY: Guarantees

1) BF search: guarantees on the set of explored patterns:

Theorem

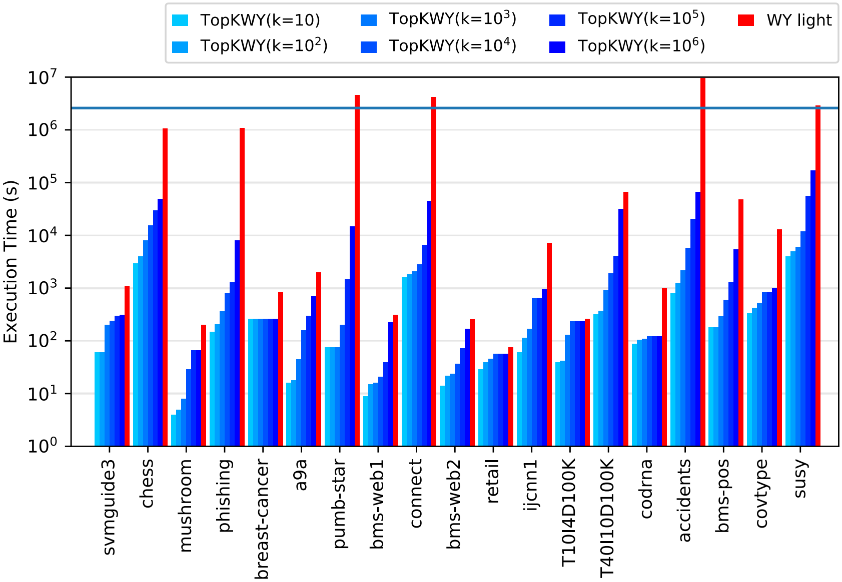
Let $\overline{\delta} = \min\{p^k, \delta\}$, *and* $\theta^* = \max\{x : \hat{\psi}(x) > \overline{\delta}\}$.
TopKWY will process only the set $FP(\mathcal{D}, \mathcal{H}, \theta^*) = \mathcal{T}(\overline{\delta})$.

$\rightarrow$ the DF search *always* explores a super-set of $\mathcal{T}(\overline{\delta})$.

2) Improved bounds to *skip* the processing of the permutations for many patterns.

(More details on the paper ☺)

TopKWY: Running time



Outline

What about controlling the FDR?

Let V the number of false discoveries (rejected *null* hypotheses).

Family-Wise Error Rate (FWER): $\Pr[V \geqslant 1]$.

Let R the number of discoveries (i.e., rejected hypotheses).

False Discovery Rate (FDR): $\mathbb{E}[V/R]$ (assuming $V/R = 0$ when $R = 0$).

What about controlling the FDR?

Let V the number of false discoveries (rejected *null* hypotheses).

Family-Wise Error Rate (FWER): $\Pr[V \geqslant 1]$.

Let R the number of discoveries (i.e., rejected hypotheses).

False Discovery Rate (FDR): $\mathbb{E}[V/R]$ (assuming $V/R = 0$ when $R = 0$).

Significant pattern mining while controlling the FDR?

What about controlling the FDR? (2)

Some methods for scenario where *significance* $\neq$ association with a class label:

What about controlling the FDR? (2)

Some methods for scenario where *significance* $\neq$ association with a class label:

- significance $=$ deviation from expectation when items place **independently** in transactions (with same frequency as in dataset $\mathcal{D}$) [Kirsch, Mitzenmacher, Pietracaprina, Pucci, Upfal, Vandin. Journal of the ACM 2012]

What about controlling the FDR? (2)

Some methods for scenario where *significance* $\neq$ association with a class label:

- significance $=$ deviation from expectation when items place **independently** in transactions (with same frequency as in dataset $\mathcal{D}$) [Kirsch, Mitzenmacher, Pietracaprina, Pucci, Upfal, Vandin. Journal of the ACM 2012]
- *statistical emerging patterns*: given a threshold $a \in (0, 1)$, probability class label is c_1 when pattern $\mathcal{S}$ is present is $\geqslant a$ [Komiyama, Ishihata, Arimura, Nishibayashi, Minato. KDD 2017.]

What about controlling the FDR? (2)

Some methods for scenario where *significance* $\neq$ association with a class label:

- significance = deviation from expectation when items place **independently** in transactions (with same frequency as in dataset $\mathcal{D}$) [Kirsch, Mitzenmacher, Pietracaprina, Pucci, Upfal, Vandin. Journal of the ACM 2012]
- *statistical emerging patterns*: given a threshold $a \in (0, 1)$, probability class label is c_1 when pattern $\mathcal{S}$ is present is $\geqslant a$ [Komiyama, Ishihata, Arimura, Nishibayashi, Minato. KDD 2017.]

Not a solved problem!

Outline

1. Introduction and Theoretical Foundations
2. Mining Statistically-Sound Patterns
3. **Recent developments and advanced topics**
   - 3.1 Controlling the FDR
   - 3.2 **Covariate-adaptive methods**
   - 3.3 Relaxing all conditional assumptions
4. Final Remarks

Using additional information

Sometimes there are additional measures (*covariates*) that provide information on *whether* a pattern *can* be significant.

Using additional information

Sometimes there are additional measures (*covariates*) that provide information on *whether* a pattern *can* be significant.

Example: the support $\sigma(\mathcal{S})$ of $\mathcal{S}$ has an impact on its minimum achivable p-value for Fisher's exact test

Using additional information

Sometimes there are additional measures (*covariates*) that provide information on *whether* a pattern *can* be significant.

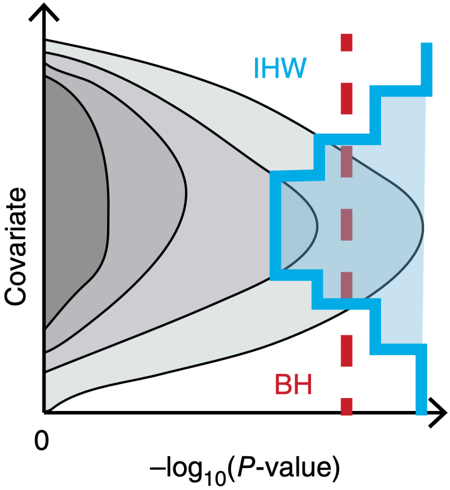
Example: the support $\sigma(\mathcal{S})$ of $\mathcal{S}$ has an impact on its minimum achivable p-value for Fisher's exact test

The covariate can be used to *weight* hypotheses/patterns or, equivalently, use different correction thresholds for False Discovery Rate (FDR) based on the covariate

Independent Hypothesis Weighting (IHW)[16]

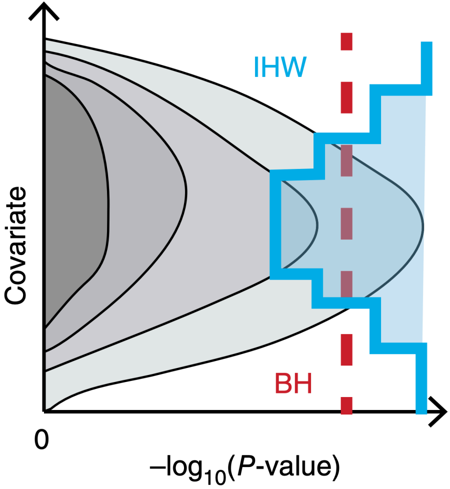
[16] Ignatiadis, Nikolaos, et al. *Data-driven hypothesis weighting increases detection power in genome-scale multiple testing.* Nature methods 13.7 (2016): 577.

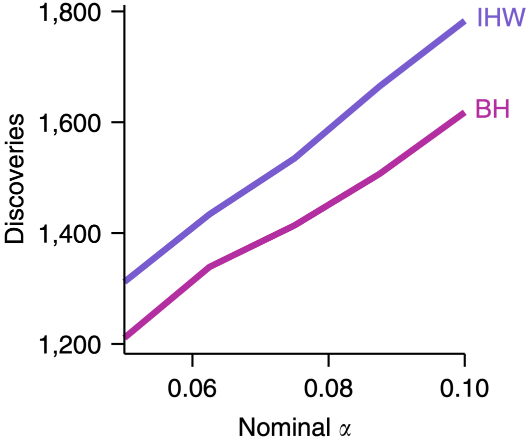
Independent Hypothesis Weighting (IHW)[16]



[16] Ignatiadis, Nikolaos, et al. *Data-driven hypothesis weighting increases detection power in genome-scale multiple testing.* Nature methods 13.7 (2016): 577.

Independent Hypothesis Weighting (IHW)[16]





[16] Ignatiadis, Nikolaos, et al. *Data-driven hypothesis weighting increases detection power in genome-scale multiple testing.* Nature methods 13.7 (2016): 577.

Outline

1. Introduction and Theoretical Foundations
2. Mining Statistically-Sound Patterns
3. **Recent developments and advanced topics**
    3.1 Controlling the FDR
    3.2 Covariate-adaptive methods
    3.3 **Relaxing all conditional assumptions**
4. Final Remarks

No conditioning?

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Fisher's test: conditioning on *both row and column totals*

Barnard's test: conditioning only on *row totals*.

Removing the conditioning on the columns was *really controversial*.

It makes sense in a *pattern mining setting* (and others).

No conditioning?

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Fisher's test: conditioning on *both row and column totals*

Barnard's test: conditioning only on *row totals*.

Removing the conditioning on the columns was *really controversial*.

It makes sense in a *pattern mining setting* (and others).

Q: Shall we stop conditioning on the *row totals*?

In general, removing assumptions is a blessed goal.

Why no conditioning? (2)

Conditioning is *bad*, even when it *approximately* preserve the likelihood.

It destroys the *repeated-sampling* (frequentist) interpretation of p-value, because it *reduces the sample space*:

fewer datasets are considered possible,

often too few to be realistic.

Why no conditioning? (1)

Single-experiment: removing row conditioning is *almost unnatural*.

No one does it → no controversy! ☺

Why no conditioning? (1)

Single-experiment: removing row conditioning is *almost unnatural*.

No one does it → no controversy! ☺

KDD settings: $\mathcal{D}$ is built by *actually sampling* from a distribution whose domain also include the group label:

the row totals are *random variables* and rightly so.

So *let's stop conditioning*, and only keep the sample size n as fixed.

Why no conditioning? (1)

Single-experiment: removing row conditioning is *almost unnatural*.

No one does it → no controversy!

KDD settings: $\mathcal{D}$ is built by *actually sampling* from a distribution whose domain also include the group label:

the row totals are *random variables* and rightly so.

So *let's stop conditioning*, and only keep the sample size n as fixed.

How?

Outline

Final Remarks

Knowl. Disc. should be based on hypothesis testing:

the data is never the whole universe.

Lots of room for research: we scratched the surface

Statistics: tests with higher power, fewer assumptions

CS: *scalability* (wrt many dimensions) is still an issue.

Balance theory and practice (that's what *we* are good at)

Work with real scientists, with real data, with real problems.

Hypothesis Testing and Statistically-sound Pattern Mining

Tutorial - KDD 2019

Leonardo Pellegrina[1] Matteo Riondato[2] Fabio Vandin[1]

[1]Dept. of Information Engineering, University of Padova (IT)

[2]Dept. of Computer Science, Amherst College (USA)