Hypothesis Testing and Statistically-sound Pattern Mining

Tutorial — SDM'21

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)

Tutorial webpage: `http://rionda.to/statdmtut`

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

Outline

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**!

Statistically-Sound Pattern Mining

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

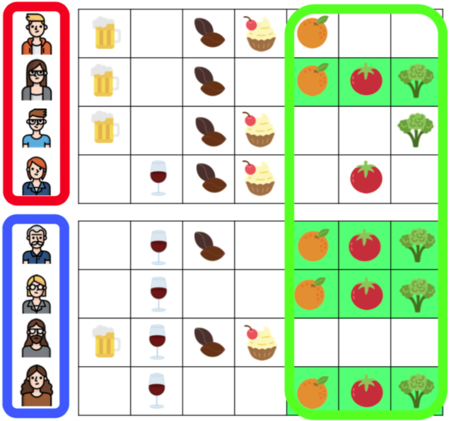
We are given:

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

Example: market basket analysis

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

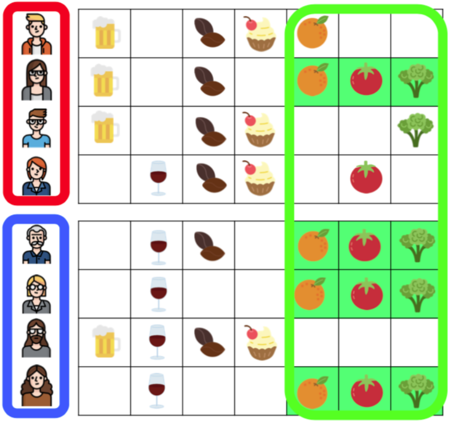
Pattern $\mathcal{S}$: subset of items (orange, tomato, broccoli)



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

Statistical Hypothesis Testing: p-value

Let $x_S = f_S(\mathcal{D})$ 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 $x_S = f_S(\mathcal{D})$ 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**: $p = \Pr[X_S$ more extreme than $x_S : H_0$ is true$]$

Statistical Hypothesis Testing: p-value

Let $x_S = f_S(\mathcal{D})$ 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 $x_S = f_S(\mathcal{D})$ 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

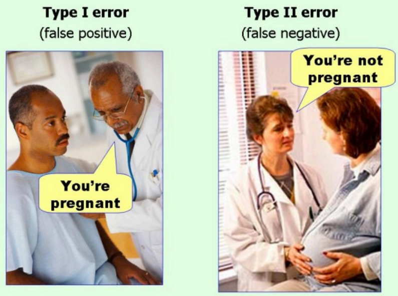
DECISION \ REALITY	H_0 false	H_0 true
reject H_0	Correct!	Type I error
accept H_0	Type II error	Correct!

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!

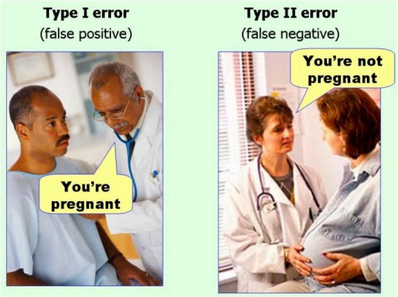


Statistical Hypothesis Testing: Errors

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

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



Theorem

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

Statistical Hypothesis Testing: Power

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...

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 $\Pr[H_0 \text{ is rejected} : H_0 \text{ is false}] = \beta$

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 $\Pr[\text{type II error}] = 1 - \beta$

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 $\Pr[\text{type II error}] = 1 - \beta$

(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

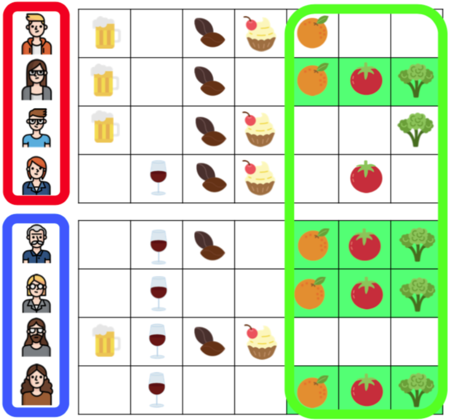
Goal: understand if the appearance of S in transactions ($\mathcal{S} \subseteq t_i$) and the transactions labels ($\ell(t_i)$) are *independent*.

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 "$\ell(t_i) = c_1$"

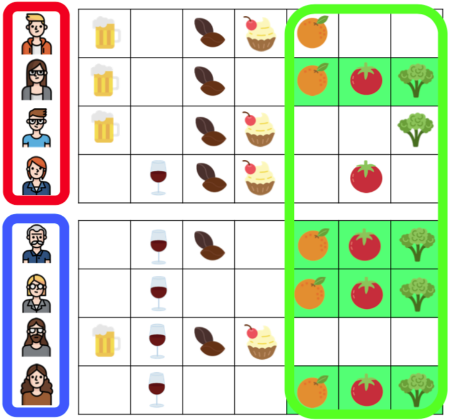
Example: market basket analysis

$\mathcal{S}$ = {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*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	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

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*

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	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
- $\sigma_0(\mathcal{S})$ = 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

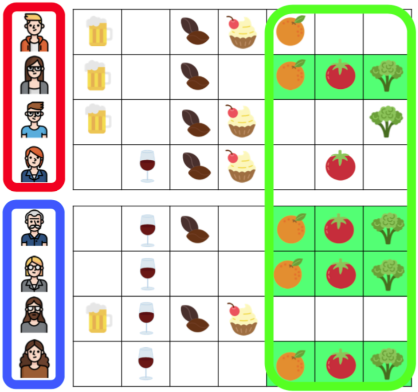
Example: Testing for Independence (3)

Useful representation of the data: *contingency table*

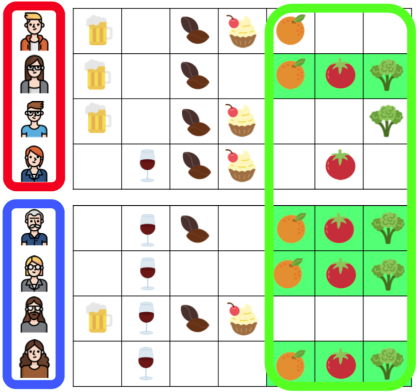
	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

Test statistic $= \sigma_1(S)$

Example: market basket analysis

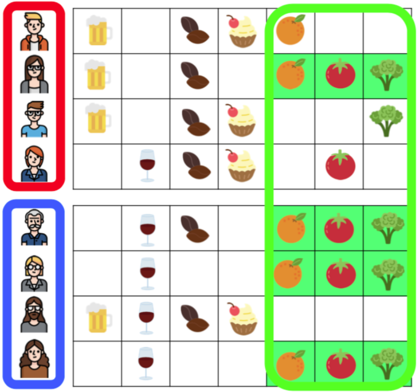


Example: market basket analysis



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

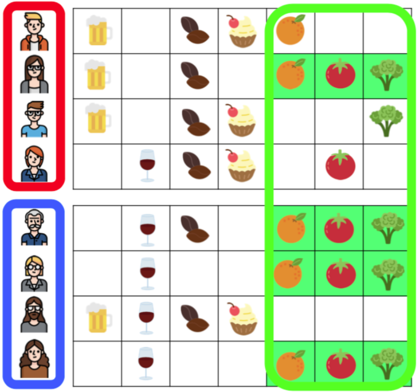
Example: market basket analysis



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

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

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

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

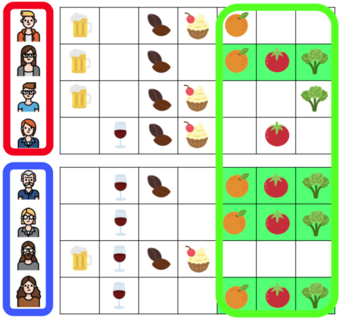
	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	$\sigma_1(\mathcal{S})$	$n_1 - \sigma_1(\mathcal{S})$	n_1
$\ell(t_i) = c_0$	$\sigma_0(\mathcal{S})$	$n_0 - \sigma_0(\mathcal{S})$	n_0
Col. m.	$\sigma(\mathcal{S})$	$n - \sigma(\mathcal{S})$	n

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 $\sigma_{\mathcal{S}}$

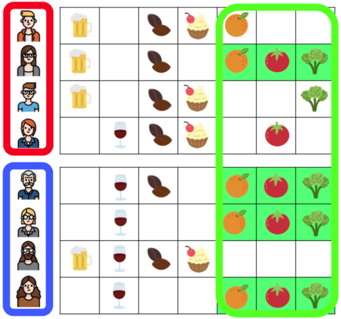
$\Rightarrow$ the p-value is **easily computable!**

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

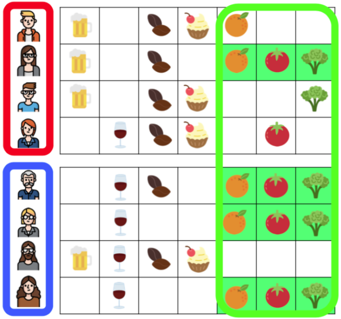
Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	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, 4

Example: market basket analysis

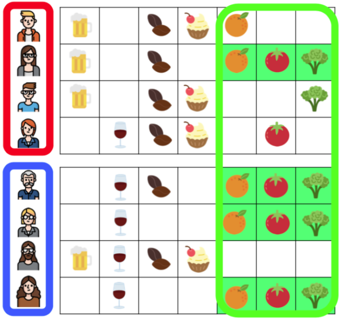


	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

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

$\Rightarrow$ Probability of table $= \Pr(X_{\mathcal{S}} = 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.228$

Example: market basket analysis



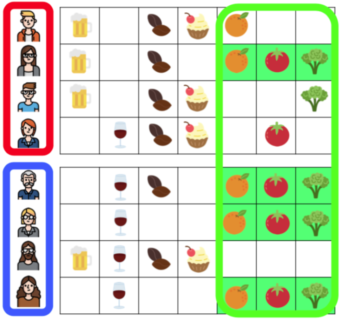
	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

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

$\Rightarrow$ Probability of table $= \Pr(X_{\mathcal{S}} = 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.228$

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

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, 4

$\Rightarrow$ Probability of table $= \Pr(X_{\mathcal{S}} = 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.228$

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

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

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	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

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

χ^2 test

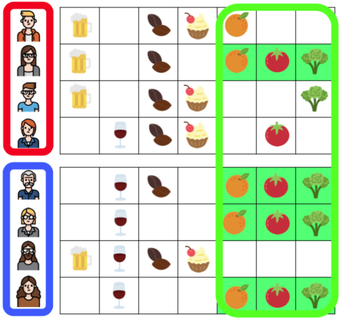
Theorem

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

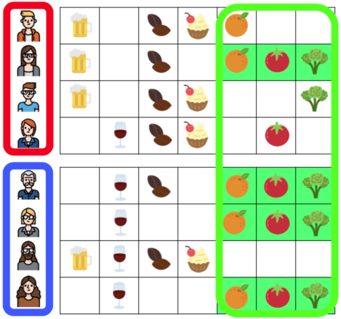
Note: the χ^2 test is the *asymptotic* version of Fisher's exact test.

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

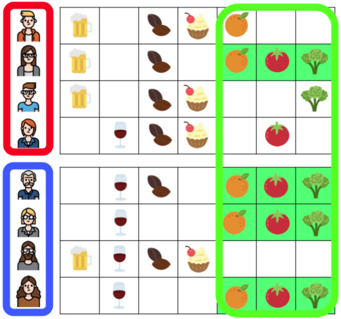
Example: market basket analysis



	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	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

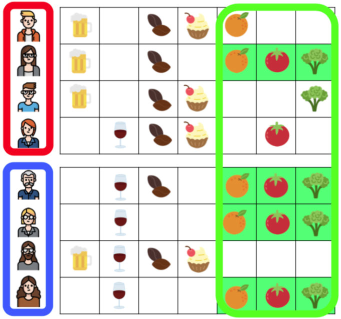


	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

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

Test statistic: 2

Example: market basket analysis



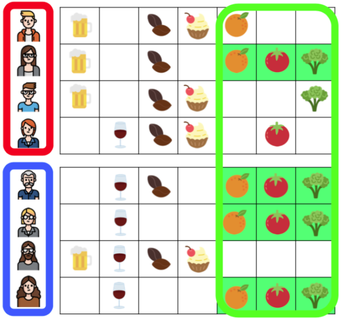
	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

$X_{\mathcal{S}} \sim \chi^2$ 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)

	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \not\subseteq t_i$	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

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

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)

Bernard's exact test(3)

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

Bernard's exact test(3)

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

Fixed π, the probability of the contingency table is easy to compute.

Bernard's exact test(3)

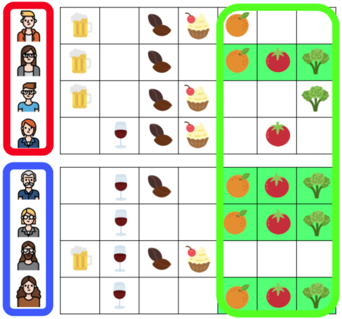
Test statistic: **probability of the contingency table**

Fixed π, the probability of the contingency table is easy to compute.

However, computing the p-value is computationally expensive!

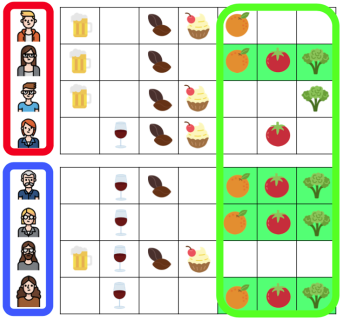
- π is unknown: consider a grid of values for π
- need to enumerate all tables *more extreme* than the observed table *for a given* π

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

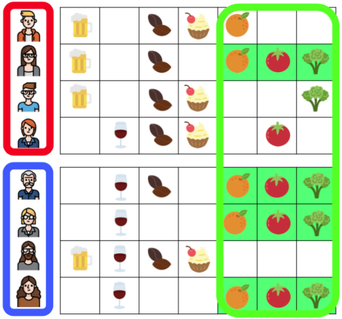
Example: market basket analysis



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

probability of table given π: $\Pr(4, 3|\pi) = \binom{4}{1}\binom{4}{3}(\pi)^4(1-\pi)^4$

Example: market basket analysis



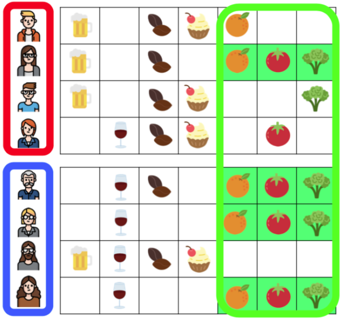
	$\mathcal{S} \subseteq t_i$	$\mathcal{S} \nsubseteq t_i$	Row m.
$\ell(t_i) = c_1$	3	1	4
$\ell(t_i) = c_0$	1	3	4
Col. m.	4	4	8

probability of table given π: $\Pr(4, 3|\pi) = \binom{4}{1}\binom{4}{3}(\pi)^4(1-\pi)^4$

more extreme tables (given π):

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

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

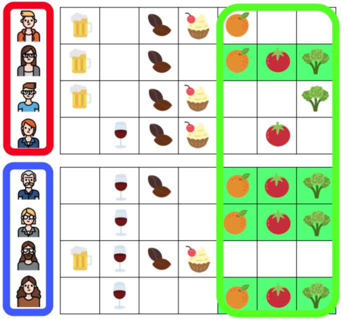
probability of table given π: $\Pr(4, 3|\pi) = \binom{4}{1}\binom{4}{3}(\pi)^4(1-\pi)^4$

more extreme tables (given π):

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

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

probability of table given π: $\Pr(4,3|\pi) = \binom{4}{1}\binom{4}{3}(\pi)^4(1-\pi)^4$

more extreme tables (given π):

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

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 π = 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 $\mathcal{D}$

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 $m = |\mathcal{H}|$.

E.g., 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 $m = |\mathcal{H}|$.

E.g., itemsets from a universe $\mathcal{I}$ of items: $m = 2^{|\mathcal{I}|} - 1$

Proposition

If we use α to test the significance of *each* hypothesis in $\mathcal{H}$, then

$$\mathbb{E}[\text{number of } \textit{false discoveries}] = m \times \alpha$$

Multiple hypothesis testing

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

E.g., itemsets from a universe $\mathcal{I}$ of items: $m = 2^{|\mathcal{I}|} - 1$

Proposition

If we use α to test the significance of *each* hypothesis in $\mathcal{H}$, then

$$\mathbb{E}[\text{number of } \textit{false discoveries}] = m \times \alpha$$

Typical α to test a *single* hypothesis: $\alpha = 0.05$ or 0.01

$\Rightarrow$ *many false discoveries* in expectation

$\Rightarrow$ at least *one with high probability*!

We want **guarantees on the probability of any false discovery**

Multiple Hypothesis testing procedures

We want **guarantees on the probability of any false discovery**

Family-Wise Error Rate (FWER):

$$\Pr[> 0 \text{ false discoveries}]$$

We want *FWER* $\leqslant \alpha$, for some $\alpha \in (0, 1)$.

How to achieve this goal?

Multiple Hypothesis testing procedures

We want **guarantees on the probability of any false discovery**

Family-Wise Error Rate (FWER):

$$\Pr[> 0 \text{ false discoveries}]$$

We want *FWER* $\leqslant \alpha$, for some $\alpha \in (0, 1)$.

How to achieve this goal?

- Bonferroni correction
- Bonferroni-Holm procedure
- ...

Bonferroni correction

$\mathcal{H}$: set of hypotheses (*patterns*) to test, $m = |\mathcal{H}|$.
For $\mathcal{S} \in \mathcal{H}$, let $H_{\mathcal{S},0}$ be the corresponding *null hypothesis*.

Bonferroni correction

$\mathcal{H}$: set of hypotheses (*patterns*) to test, $m = |\mathcal{H}|$.
For $\mathcal{S} \in \mathcal{H}$, let $H_{\mathcal{S},0}$ be the corresponding *null hypothesis*.
Rejection rule: Given a *statistical level* $\alpha \in (0, 1)$:

reject $H_{S,0}$ (i.e., flag $\mathcal{S}$ as significant) iff $p \leqslant \frac{\alpha}{m}$

Bonferroni correction

$\mathcal{H}$: set of hypotheses (*patterns*) to test, $m = |\mathcal{H}|$.
For $\mathcal{S} \in \mathcal{H}$, let $H_{\mathcal{S},0}$ be the corresponding *null hypothesis*.
Rejection rule: Given a *statistical level* $\alpha \in (0,1)$:

reject $H_{S,0}$ (i.e., flag $\mathcal{S}$ as significant) iff $p \leqslant \frac{\alpha}{m}$

Why does this approach controls the FWER?

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

Bonferroni correction

$\mathcal{H}$: set of hypotheses (*patterns*) to test, $m = |\mathcal{H}|$.
For $\mathcal{S} \in \mathcal{H}$, let $H_{\mathcal{S},0}$ be the corresponding *null hypothesis*.
Rejection rule: Given a *statistical level* $\alpha \in (0,1)$:

reject $H_{S,0}$ (i.e., flag $\mathcal{S}$ as significant) iff $p \leqslant \frac{\alpha}{m}$

Why does this approach controls the FWER?

- 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 $] \leqslant |\mathcal{H}| \frac{\alpha}{m} \leqslant \alpha$

Choosing hypotheses *before* testing?

Alphabet of items $\mathcal{I}$ with $|\mathcal{I}| = 6000$
Dataset $\mathcal{D}$ with 10 transactions with label c_1, 10 with label c_0
Hypotheses $\mathcal{H} = \mathcal{I}$

- "large m, small data: nothing will be flagged as significant!" 

Choosing hypotheses *before* testing?

Alphabet of items $\mathcal{I}$ with $|\mathcal{I}| = 6000$
Dataset $\mathcal{D}$ with 10 transactions with label c_1, 10 with label c_0
Hypotheses $\mathcal{H} = \mathcal{I}$

- "large m, small data: nothing will be flagged as significant!" 
- "let's select some hypotheses first, and then do the testing...":
  find pattern $\mathcal{S}^* = \arg\max_{\mathcal{S}\in\mathcal{H}}(\sigma_1(\mathcal{S}) - \sigma_0(\mathcal{S}))$.
- "I am going to test only $\mathcal{S}^*$!"
  E.g., $\sigma_1(\mathcal{S}^*) = 10, \sigma_0(\mathcal{S}^*) = 0$. Fisher's test p-value = 0.0001

Choosing hypotheses *before* testing?

Alphabet of items $\mathcal{I}$ with $|\mathcal{I}| = 6000$
Dataset $\mathcal{D}$ with 10 transactions with label c_1, 10 with label c_0
Hypotheses $\mathcal{H} = \mathcal{I}$

- "large m, small data: nothing will be flagged as significant!"
- "let's select some hypotheses first, and then do the testing...":
  find pattern $\mathcal{S}^* = \arg\max_{\mathcal{S} \in \mathcal{H}} (\sigma_1(\mathcal{S}) - \sigma_0(\mathcal{S}))$.
- "I am going to test only $\mathcal{S}^*$!"
  E.g., $\sigma_1(\mathcal{S}^*) = 10, \sigma_0(\mathcal{S}^*) = 0$. Fisher's test p-value = 0.0001
- "$\mathcal{S}_*$ 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 generated as follows:

- Each item/pattern $\mathcal{S}$ will appear exactly 10 times
- For $i = 1, \ldots, 10$, place $\mathcal{S}$ in the i-th transaction labeled c_0 with probability $1/2$, and the i-th transaction labeled c_1 otherwise

No pattern $\mathcal{S}$ **is associated with class labels!**

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

BUT IT IS NOT!

Assume that $\mathcal{D}$ is generated as follows:

- Each item/pattern $\mathcal{S}$ will appear exactly 10 times
- For $i = 1, \ldots, 10$, place $\mathcal{S}$ in the i-th transaction labeled c_0 with probability $1/2$, and the i-th transaction labeled c_1 otherwise

No pattern $\mathcal{S}$ **is associated with class labels!**

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

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

BUT IT IS NOT!

Assume that $\mathcal{D}$ is generated as follows:

- Each item/pattern $\mathcal{S}$ will appear exactly 10 times
- For $i = 1, \ldots, 10$, place $\mathcal{S}$ in the i-th transaction labeled c_0 with probability $1/2$, and the i-th transaction labeled c_1 otherwise

No pattern $\mathcal{S}$ **is associated with class labels!**

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

In *expectation*, ≈ 5 patterns with $\sigma_1(\mathcal{S}) = 10$ and $\sigma_0(\mathcal{S}) = 0$.

they are *all* false discoveries!

Where is the problem?

We selected the hypothesis to test on the basis of its support $\sigma_1(\mathcal{S})$

Where is the problem?

We selected the hypothesis to test on the basis of its support $\sigma_1(\mathcal{S})$

$\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$ *is clearly related to the* p*-value*

Where is the problem?

We selected the hypothesis to test on the basis of its support $\sigma_1(\mathcal{S})$

$\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$ *is clearly related to the p-value*

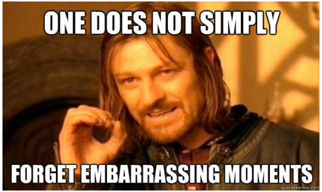
We have essentially **looked at the p-values of all hypotheses** and then **acted as if we did not!** 

Where is the problem?

We selected the hypothesis to test on the basis of its support $\sigma_1(\mathcal{S})$

$\sigma_1(\mathcal{S}) = 10 - \sigma_0(\mathcal{S})$ *is clearly related to the p-value*

We have essentially **looked at the p-values of all hypotheses** and then **acted as if we did not!**



Outline

Selecting hypotheses

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

Selecting hypotheses

A smaller $\mathcal{H}$ will lead to a *higher corrected significance threshold* $\alpha/|\mathcal{H}|$, thus *may* lead 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

A smaller $\mathcal{H}$ will lead to a *higher corrected significance threshold* $\alpha/|\mathcal{H}|$, thus *may* lead 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 Bonferroni correction on $\mathcal{H}'$ to bound the FWER (for $\mathcal{H}$)

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 Bonferroni correction on $\mathcal{H}'$ to bound the FWER (for $\mathcal{H}$)

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 $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$.

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 the Bonferroni correction on $\mathcal{H}'$.

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 $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$.

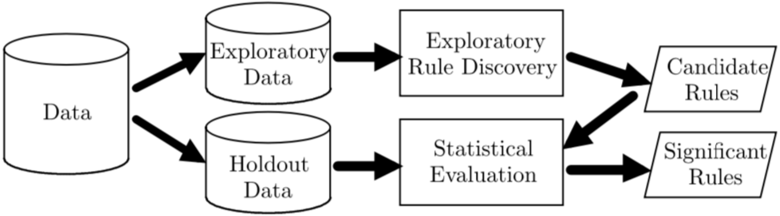
2. Apply some selection procedure to $\mathcal{D}_1$ to select $\mathcal{H}'$
   (it may include performing the tests on $\mathcal{D}_1$).

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

Splitting $\mathcal{D}$ is *similar* to using a training set and a test set.

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?

How selective shall we be?

Let $\mathcal{Z}_\alpha \subseteq \mathcal{H}$ be the 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 increases because the corrected significance threshold increases?

How selective shall we be?

Let $\mathcal{Z}_\alpha \subseteq \mathcal{H}$ be the 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 increases 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]

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

A breakthrough [Tarone 1990]

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

A breakthrough [Tarone 1990]

The statistic of Fisher's exact test 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]

The statistic of Fisher's exact test 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]

The statistic of Fisher's exact test 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]

The statistic of Fisher's exact test 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

A breakthrough [Tarone 1990]

The statistic of Fisher's exact test 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)

The statistic of Fisher's exact test 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 the statistic for pattern $\mathcal{S}$ with support $\sigma(\mathcal{S})$ assuming $\sigma_1(\mathcal{S}) = x$.

A breakthrough [Tarone 1990] (2)

The statistic of Fisher's exact test 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 the statistic for pattern $\mathcal{S}$ with support $\sigma(\mathcal{S})$ assuming $\sigma_1(\mathcal{S}) = x$.

It must be $\max\{0, n_1 - (n - \sigma(\mathcal{S}))\} \leqslant x \leqslant \min\{\sigma(\mathcal{S}), n_1\}$

A breakthrough [Tarone 1990] (2)

The statistic of Fisher's exact test 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 the statistic for pattern $\mathcal{S}$ with support $\sigma(\mathcal{S})$ assuming $\sigma_1(\mathcal{S}) = x$.

It must be $\max\{0, n_1 - (n - \sigma(\mathcal{S}))\} \leqslant x \leqslant \min\{\sigma(\mathcal{S}), n_1\}$
$\Rightarrow$ the range of $p^F(\sigma(\mathcal{S}), x)$ depends only on $\sigma(\mathcal{S})$ (n, n_1 are fixed)

A breakthrough [Tarone 1990] (3)

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

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

A breakthrough [Tarone 1990] (3)

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

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

Tarone's result: when testing each hypothesis with significance level δ, then **the hypotheses that will certainly have p-value greater than δ do not need to be counted when using Bonferroni's correction!** ☺

A breakthrough [Tarone 1990] (4)

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

A breakthrough [Tarone 1990] (4)

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

A breakthrough [Tarone 1990] (4)

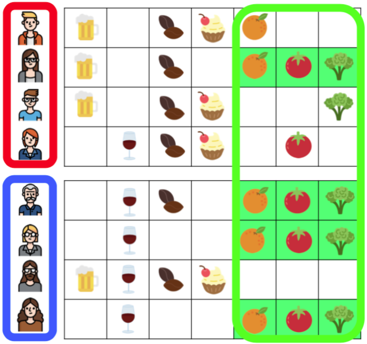
$\mathcal{S}$ cannot be significant with significance level δ if $\psi(\sigma(\mathcal{S})) > \delta \Rightarrow \mathcal{S}$ is **untestable**.

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

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

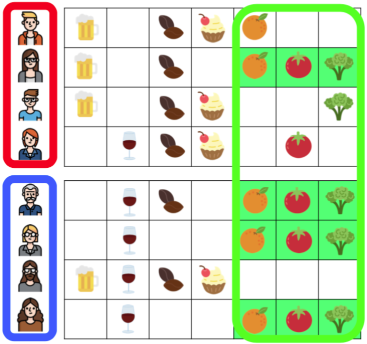
All the others do not really matter, and should not be counted when applying the Bonferroni correction to control for the FWER.

Example: market basket analysis



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

Example: market basket analysis

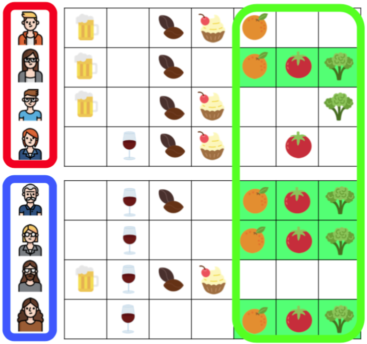


$\mathcal{S} = \{\text{orange, tomato, broccoli}\}$

minimum attainable p-value

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

Example: market basket analysis



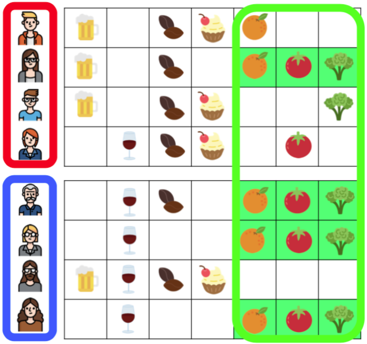
$\mathcal{S} = \{$orange, tomato, broccoli$\}$

minimum attainable p-value

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

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

Example: market basket analysis



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

minimum attainable p-value

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

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

$\Rightarrow$ if the significance level used to test each hypothesis 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 $\delta^* = \max\{\delta : \delta \leqslant \alpha/|\mathcal{T}(\delta)|\}$!

Now, like always, is a good time for questions on:

Multiple hypothesis testing

Bonferroni Correction

Tarone's approach to selecting hypotheses

Minimal attainable p-value

Anything else $=)$

Now, like always, is a good time for questions on:

Multiple hypothesis testing

Bonferroni Correction

Tarone's approach to selecting hypotheses

Minimal attainable p-value

Anything else $=$)

Let's take a 5–10 minutes break.

Outline

Selecting testable patterns

Minimum attainable p-value $\psi(\sigma(\mathcal{S}))$ of a pattern $\mathcal{S}$: select patterns to test from $\mathcal{H}$.

Selecting testable patterns

Minimum attainable p-value $\psi(\sigma(\mathcal{S}))$ of a pattern $\mathcal{S}$: select patterns to test from $\mathcal{H}$.

Naïve approach: compute $\psi(\sigma(\mathcal{S}))$ for all $\mathcal{S} \in \mathcal{H}$, find $\delta^{\star}$

Selecting testable patterns

Minimum attainable p-value $\psi(\sigma(\mathcal{S}))$ of a pattern $\mathcal{S}$: select patterns to test from $\mathcal{H}$.

Naïve approach: compute $\psi(\sigma(\mathcal{S}))$ for all $\mathcal{S} \in \mathcal{H}$, find $\delta^\star$

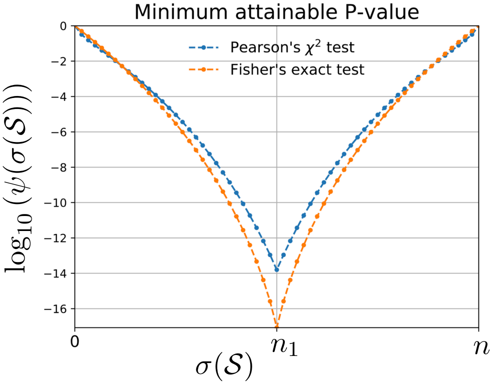
Not possible to enumerate all $\mathcal{S} \in \mathcal{H}$...

Minimum attainable p-value $\psi(\sigma(\mathcal{S}))$ of a pattern $\mathcal{S}$ is a function of its support $\sigma(\mathcal{S})$ in the data.
Low (and very high) support $\sigma(\mathcal{S}) \rightarrow$ large $\psi(\sigma(\mathcal{S}))$

[1] A. Terada, et. al. *Statistical significance of combinatorial regulations.* PNAS, 2013.

Minimum attainable p-value $\psi(\sigma(\mathcal{S}))$ of a pattern $\mathcal{S}$ is a function of its support $\sigma(\mathcal{S})$ in the data.

Low (and very high) support $\sigma(\mathcal{S}) \rightarrow$ large $\psi(\sigma(\mathcal{S}))$



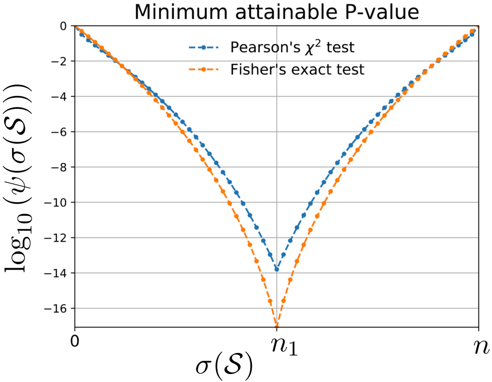
$n = 60,\ n_1 = 30.$

(from F. Llinares-López, D. Roqueiro, ISMB'18 Tutorial.)

[1] A. Terada, et. al. *Statistical significance of combinatorial regulations.* PNAS, 2013.

Minimum attainable p-value $\psi(\sigma(\mathcal{S}))$ of a pattern $\mathcal{S}$ is a function of its support $\sigma(\mathcal{S})$ in the data.

Low (and very high) support $\sigma(\mathcal{S}) \rightarrow$ large $\psi(\sigma(\mathcal{S}))$



$n = 60,\ n_1 = 30.$

(from F. Llinares-López, D. Roqueiro, ISMB'18 Tutorial.)

Intuition of `LAMP`[1]: connection betw. *testable* and *frequent* patterns!

[1] A. Terada, et. al. *Statistical significance of combinatorial regulations.* PNAS, 2013.

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

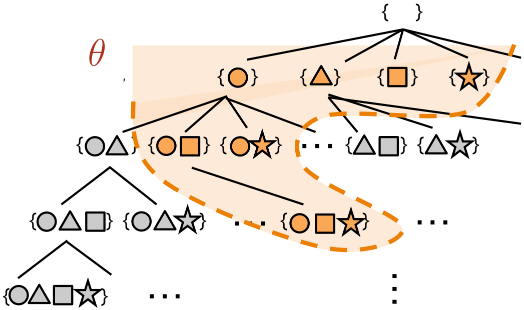
$$FP(\mathcal{D}, \mathcal{H}, \theta) := \{\mathcal{S} \in \mathcal{H} : \sigma(\mathcal{S}) \geqslant \theta\} .$$

Frequent Pattern Mining

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

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

Typical approach: Explore the *search tree* of $\mathcal{H}$, *pruning* subtrees with support $< \theta$ (monotonicity of support)

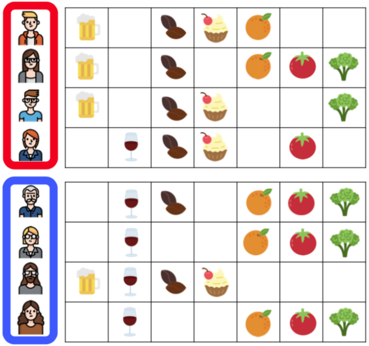


Frequent Pattern Mining

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:

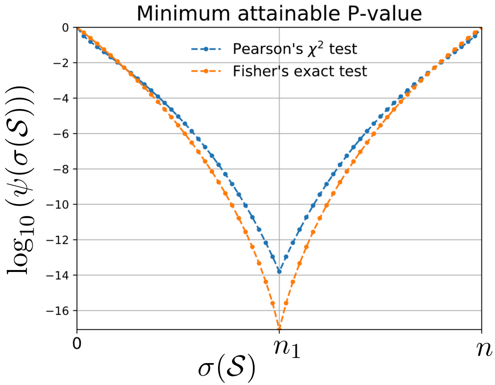
$\mathcal{S}' = \{$🫘, 🍊, 🍅, 🥦$\}$, $\mathcal{S} = \{$🍅$\}$

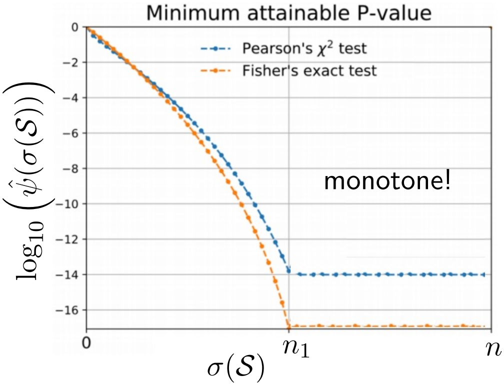
$\sigma(\mathcal{S}') = 2 \leqslant \sigma(\mathcal{S}) = 5.$

Valid for many other patterns (e.g., *subgraphs, sequential patterns, subgroups*, ...)

LAMP: monotone minimum achievable p-value function $\hat{\psi}(\cdot)$:

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





We obtain the equivalence:

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

Thus:

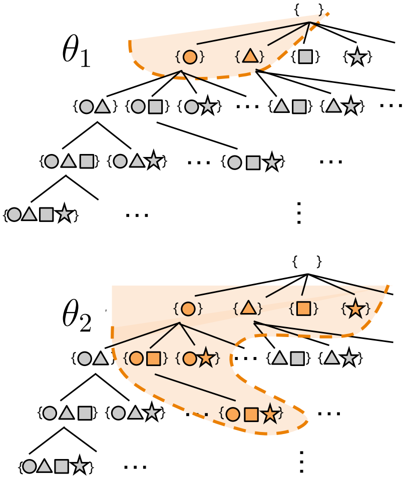
$$|\mathcal{T}(\hat{\psi}(\theta))| = |FP(\mathcal{D}, \mathcal{H}, \theta)|.$$

We can use $|FP(\mathcal{D}, \mathcal{H}, \theta)|$ to find

$$\delta^* = \max\{\delta : \delta|\mathcal{T}(\delta)| \leqslant \alpha\}.$$

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

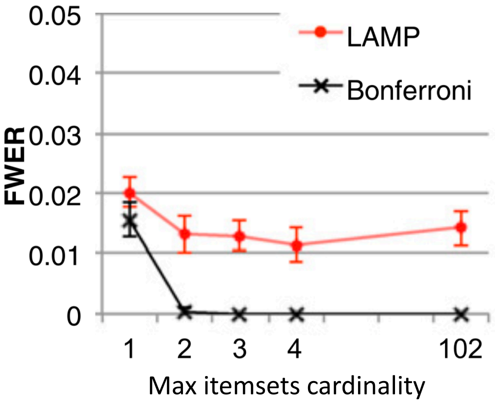
Performs multiple Frequent Pattern Mining instances (decreasing values of θ) to evaluate $|FP(\mathcal{D}, \mathcal{H}, \theta)|$.
Find minimum θ such that it holds
$\alpha/|FP(\mathcal{D}, \mathcal{H}, \theta)| \geqslant \hat{\psi}(\theta)$



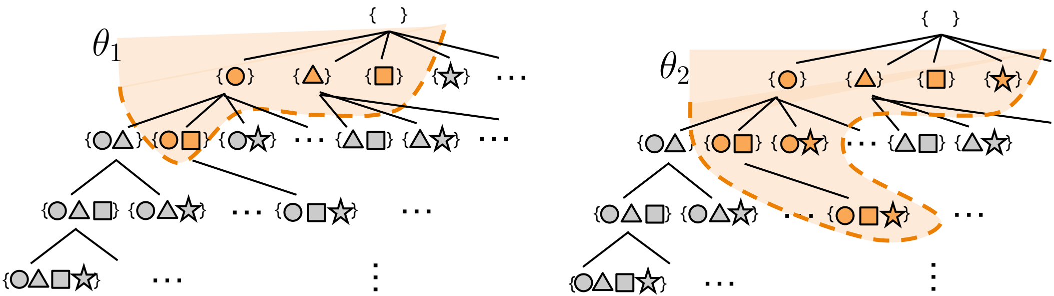
(imgs. from LAMP paper)

LAMP: Experimental Results

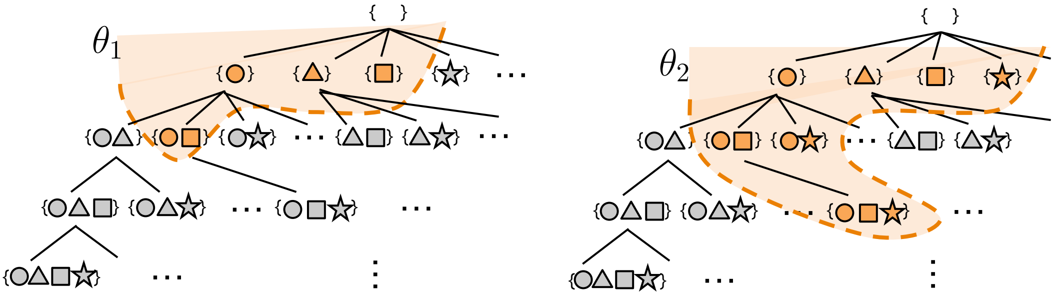
(imgs. from LAMP)



Estimated $FWER$ ($\alpha = 0.05$) of LAMP vs Bonferroni correction.



For θ_2 we count again all patterns already counted for $\theta_1 \geqslant \theta_2$!

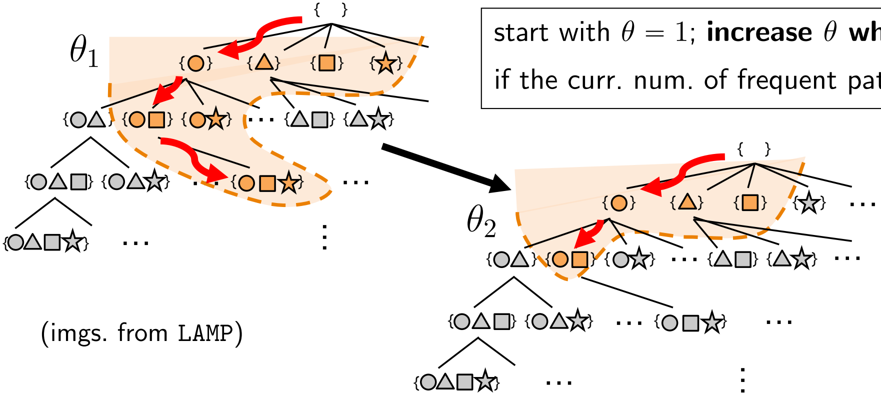


For θ_2 we count again all patterns already counted for $\theta_1 \geqslant \theta_2$!

Is it possible to explore patterns only once?

`SupportIncrease`[2]: `LAMP` with only *one* Depth-First (DF) exploration of $\mathcal{H}$.

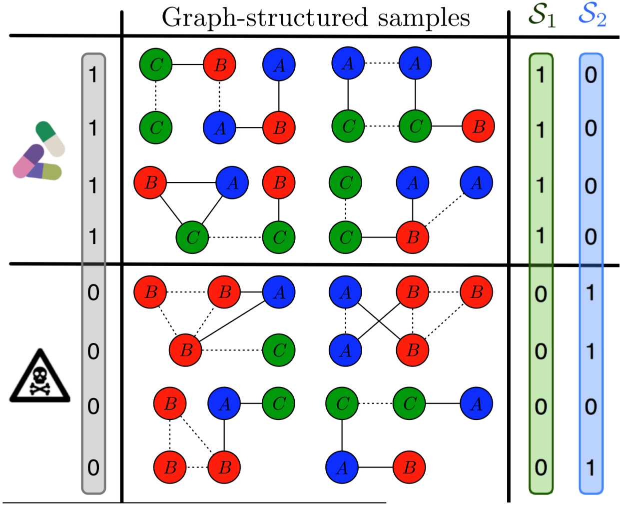
start with $\theta = 1$; **increase** θ **while exploring** if the curr. num. of frequent patterns $\geqslant \alpha/\hat{\psi}(\theta)$

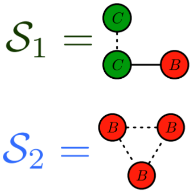


(imgs. from `LAMP`)

[2] Minato, S. I., et al. *A fast method of statistical assessment for combinatorial hypotheses based on frequent itemset enumeration*. ECML-PKDD 2014.

Mining Significant Subgraphs[4]





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

(imgs. from [3])

[3] F. Llinares-López, D. Roqueiro, *Significant Pattern Mining for Biomarker Discovery*, ISMB'18 Tutorial.

[4] M. Sugiyama, F. Llinares-López, N. Kasenburg, K.M. Borgwardt. *Significant subgraph mining with multiple testing correction.* ICDM 2015.

Outline

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

(gray = fixed,
yellow = random)

Recap: Assumptions of Fisher's test: all marginals of all the tested contingency tables are *fixed* by design of the experiment.

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

(gray = fixed, yellow = random)

Recap: Assumptions of Fisher's test: all marginals of all the tested contingency tables are *fixed* by design of the experiment.

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

(gray $=$ fixed, yellow $=$ random)

Recap: Assumptions of Fisher's test: all marginals of all the tested contingency tables are *fixed* by design of the experiment.

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

Not used in practice, mainly for computational reasons...

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

(gray = fixed,

yellow = random)

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

(gray = fixed,
yellow = random)

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 $\mathcal{C}_{\mathcal{S}}$ = observed contingency table for $\mathcal{S}$.

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

(gray = fixed,

yellow = random)

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 $\mathcal{C}_{\mathcal{S}}$ = observed contingency table for $\mathcal{S}$.

$P(\mathcal{C} \mid \pi)$ = prob. of a table $\mathcal{C}$ assuming NH and $\pi_{\mathcal{S}} = \pi$

$T(\mathcal{C}_{\mathcal{S}}, \pi) = \{$more extreme cont. tables of $\mathcal{C}_{\mathcal{S}}\}$

$$\phi(\mathcal{C}_{\mathcal{S}}, \pi) = \sum_{\mathcal{C} \in T(\mathcal{C}_{\mathcal{S}}, \pi)} P(\mathcal{C} \mid \pi)$$

p-**value:** $p_{\mathcal{S}} = \max_{\pi \in [0,1]} \{\phi(\mathcal{C}_{\mathcal{S}}, \pi)\}$

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

(gray = fixed,
yellow = random)

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 $\mathcal{C}_{\mathcal{S}}$ = observed contingency table for $\mathcal{S}$.

$P(\mathcal{C} \mid \pi)$ = prob. of a table $\mathcal{C}$ assuming NH and $\pi_{\mathcal{S}} = \pi$

$T(\mathcal{C}_{\mathcal{S}}, \pi)$ = {more extreme cont. tables of $\mathcal{C}_{\mathcal{S}}$}

$$\phi(\mathcal{C}_{\mathcal{S}}, \pi) = \sum_{\mathcal{C} \in T(\mathcal{C}_{\mathcal{S}}, \pi)} P(\mathcal{C} \mid \pi)$$

p-**value:** $p_{\mathcal{S}} = \max_{\pi \in [0,1]} \{\phi(\mathcal{C}_{\mathcal{S}}, \pi)\}$ → **hard to compute!**

Efficient Unconditional Testing: SPuManTE[5]

1) Computes *confidence intervals* $C_j(\mathcal{S})$ for $\pi_{\mathcal{S},j}$

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

Efficient Unconditional Testing: SPuManTE[6]

1) Computes *confidence intervals* $C_j(\mathcal{S})$ for $\pi_{\mathcal{S},j}$

Compute a probabilistic (high prob.) upper bound to

$$\sup_{\mathcal{S}\in\mathcal{H}, j\in\{0,1\}} \left| \pi_{\mathcal{S},j} - \frac{\sigma_j(\mathcal{S})}{n_j} \right|$$

(note: $\sigma_j(\mathcal{S})/n_j$ is *observed* from $\mathcal{D}$, $\pi_{\mathcal{S},j}$ is *unknown*)

How? Upper bound[5] to Rademacher Complexity of $\mathcal{H}$.

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

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

Efficient Unconditional Testing: SPuManTE

2) p-value p_S according to confidence intervals:

$$p_S = \begin{cases} 0 & \text{, if } C_0(\mathcal{S}) \cap C_1(\mathcal{S}) = \emptyset \\ \max\{\phi(\mathcal{C}_{\mathcal{S}}, \pi), \pi \in C_0(\mathcal{S}) \cap C_1(\mathcal{S})\} & \text{, othw.} \end{cases}$$

Flag $\mathcal{S}$ as significant if $p_S \leqslant \delta$.

Efficient Unconditional Testing: SPuManTE

p-value p_S according to confidence intervals:

$$p_S = \begin{cases} 0 & \text{, if } C_0(\mathcal{S}) \cap C_1(\mathcal{S}) = \emptyset \\ \max\{\phi(\mathcal{C}_{\mathcal{S}}, \pi), \pi \in C(\mathcal{S})\} & \text{, othw.} \end{cases}$$

p-value p_S is still expensive to compute in second case!

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

Efficient Unconditional Testing: SPuManTE

p-value p_S according to confidence intervals:

$$p_S = \begin{cases} 0 & \text{, if } C_0(\mathcal{S}) \cap C_1(\mathcal{S}) = \emptyset \\ \max\{\phi(\mathcal{C}_{\mathcal{S}}, \pi), \pi \in C(\mathcal{S})\} & \text{, othw.} \end{cases}$$

p-value p_S is still expensive to compute in second case!

3) **Upper and Lower bounds** to p_S, and **efficient algorithm** for computation of $\phi(\cdot)$

More in the paper[7] :)

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

Outline

1. Introduction and Theoretical Foundations
2. **Mining Statistically-Sound Patterns**
   2.1 `LAMP`: Tarone's method for Significant Pattern Mining
   2.2 `SPuManTE`: relaxing conditional assumptions
   2.3 **Permutation Testing**
   2.4 `WY` Permutation Testing
3. Recent developments and advanced topics
4. Final Remarks

Permutation Testing

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

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, assumptions

Permutation Testing: Setting

$\mathcal{D}_0$: observed dataset from some generative process $\mathcal{G}$.

E.g., a transactional dataset

Permutation Testing: Setting

$\mathcal{D}_0$: observed dataset from some generative process $\mathcal{G}$.

E.g., a transactional dataset

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

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

Permutation Testing: Setting

$\mathcal{D}_0$: observed dataset from some generative process $\mathcal{G}$.

E.g., a transactional dataset

$T_0 = \mathcal{A}(\mathcal{D}_0) \in \mathbb{R}$: 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 $\mathcal{D}_0$ satisfied by all $\mathcal{D} \in \mathcal{G}$

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

Permutation Testing: Setting

$\mathcal{D}_0$: observed dataset from some generative process $\mathcal{G}$.

E.g., a transactional dataset

$T_0 = \mathcal{A}(\mathcal{D}_0) \in \mathbb{R}$: 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 $\mathcal{D}_0$ satisfied by all $\mathcal{D} \in \mathcal{G}$

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

QUESTION: Is T_0 surprising? Or just 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 from $\mathcal{G}$.

I.e., it is *very likely* to observe a value $\mathcal{A}(\mathcal{D}) \geqslant T_0$ in a dataset $\mathcal{D}$ taken from $\mathcal{G}$.

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 from $\mathcal{G}$.

I.e., it is *very likely* to observe a value $\mathcal{A}(\mathcal{D}) \geqslant T_0$ in a dataset $\mathcal{D}$ taken from $\mathcal{G}$.

Ideally:

$$Q(T_0) = \Pr_{\mathcal{D}\sim\mathcal{G}} \left(\mathcal{A}(\mathcal{D}) \geqslant T_0\right). \quad \text{Reject } H_0 \text{ if } Q(T_0) \leqslant \delta.$$

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 from $\mathcal{G}$.

I.e., it is *very likely* to observe a value $\mathcal{A}(\mathcal{D}) \geqslant T_0$ in a dataset $\mathcal{D}$ taken from $\mathcal{G}$.

Ideally:

$$Q(T_0) = \Pr_{\mathcal{D} \sim \mathcal{G}} (\mathcal{A}(\mathcal{D}) \geqslant T_0) . \quad \text{Reject } H_0 \text{ if } Q(T_0) \leqslant \delta.$$

Very often: no closed form for $Q(T_0)$!

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 from $\mathcal{G}$.

I.e., it is *very likely* to observe a value $\mathcal{A}(\mathcal{D}) \geqslant T_0$ in a dataset $\mathcal{D}$ taken from $\mathcal{G}$.

Ideally:

$$Q(T_0) = \Pr_{\mathcal{D} \sim \mathcal{G}} \left(\mathcal{A}(\mathcal{D}) \geqslant T_0 \right). \quad \text{Reject } H_0 \text{ if } Q(T_0) \leqslant \delta.$$

Very often: no closed form for $Q(T_0)$!
Instead: empirical estimate $\tilde{Q}(T_0)$ of $Q(T_0)$ using samples from $\mathcal{G}$

Permutation Testing

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ *independent uniform* samples taken from $\mathcal{G}$.

Permutation Testing

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ *independent uniform* samples taken from $\mathcal{G}$.

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

Permutation Testing

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ *independent uniform* samples taken from $\mathcal{G}$.

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

3. Compute the *empirical p-value* $\tilde{Q}(T_0)$:

$$\tilde{Q}(T_0) = \frac{|\{i : T_i \geqslant T_0\}| + 1}{m + 1}$$

Permutation Testing

1. *Generate* $\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ *independent uniform* samples taken from $\mathcal{G}$.

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

3. Compute the *empirical p-value* $\tilde{Q}(T_0)$:

$$\tilde{Q}(T_0) = \frac{|\{i : T_i \geqslant T_0\}| + 1}{m + 1}$$

4. If $\tilde{Q}(T_0) \leqslant \delta$, reject H_0.

Generating uniform samples

1. Assumption: there exists a perturbation operation

$$\phi : \mathcal{G} \to \mathcal{G}$$

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

Generating uniform samples

1. Assumption: there exists a perturbation operation

$$\phi : \mathcal{G} \to \mathcal{G}$$

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

2. We need to derive sufficient number of perturbations to obtain an independent and uniform sample from $\mathcal{G}$

Example

$\mathcal{D}_0$: observed dataset (*binary matrix*).
rows: transactions: columns: items

1	0	1	1
0	1	1	0
1	0	1	0
1	0	0	1

$T_0 = \mathcal{A}(\mathcal{D}_0) =$ *number* of frequent itemsets w.r.t. frequency threshold θ

Example

$\mathcal{D}_0$: observed dataset (*binary matrix*).
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

$T_0 = \mathcal{A}(\mathcal{D}_0) =$ *number* of frequent itemsets w.r.t. frequency threshold θ

$\mathbf{P} =$ the rows and columns *totals*

Example

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

$\mathcal{D}_0$: observed dataset (*binary matrix*).
rows: transactions: columns: items

$T_0 = \mathcal{A}(\mathcal{D}_0) =$ *number* of frequent itemsets w.r.t. frequency threshold θ

$\mathbf{P}$ = the rows and columns *totals*

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

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 $u(A) = v(B) = 0$

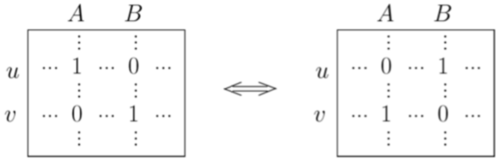


Fig. 1. A swap in a 0–1 matrix.

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

Advantages and disadvantages of permutation testing

Conceptually very natural

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

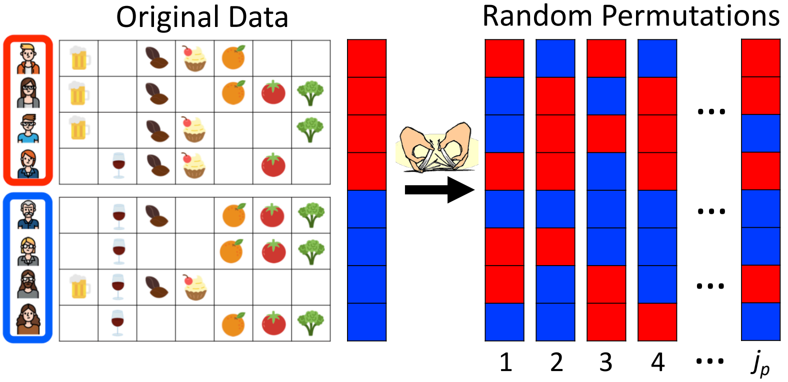
Computationally very expensive:

m times: sample generation + running $\mathcal{A}$

Outline

Westfall-Young[8] (WY) Permutation Testing

Perturbation: random shuffle of the labels (repeated m times).



Compare p-values from original data with random labels.

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

$p^j_{\min} =$ minimum p-value (over $\mathcal{H}$) on j-th random label

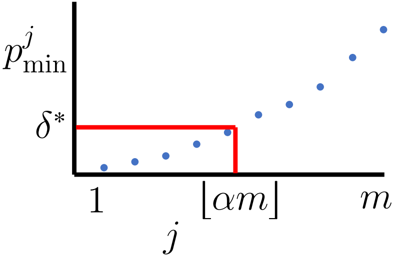
Estimated $FWER$ for sign. thr. δ: $\overline{FWER}(\delta) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[p^j_{\min} \leqslant \delta\right]$

$p^j_{\min} =$ minimum p-value (over $\mathcal{H}$) on j-th random label

Estimated $FWER$ for sign. thr. δ: $\overline{FWER}(\delta) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[p^j_{\min} \leqslant \delta\right]$

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

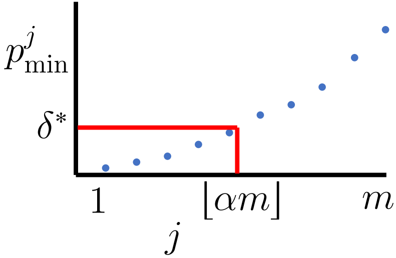
$= \alpha$-quantile of $\{p^j_{\min}\}$



$p^j_{\min}$ = minimum p-value (over $\mathcal{H}$) on j-th random label

Estimated $FWER$ for sign. thr. δ: $\overline{FWER}(\delta) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[p^j_{\min} \leqslant \delta\right]$

Compute $\delta^* = \max\left\{\delta : \overline{FWER}(\delta) \leqslant \alpha\right\}$
$= \alpha$-quantile of $\{p^j_{\min}\}$



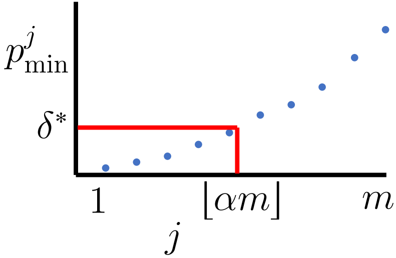
Output $\{\mathcal{S} : p_{\mathcal{S}} \leqslant \delta^*\}$.

$p^j_{\min}$ = minimum p-value (over $\mathcal{H}$) on j-th random label

Estimated $FWER$ for sign. thr. δ: $\overline{FWER}(\delta) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[p^j_{\min} \leqslant \delta\right]$

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

$= \alpha$-quantile of $\{p^j_{\min}\}$



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

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

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

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

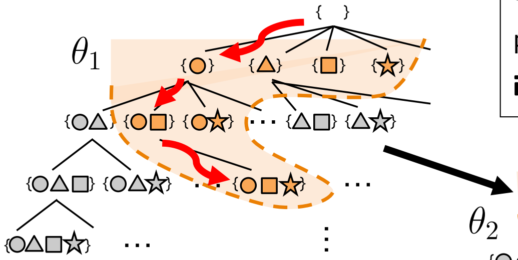
FASTWY[9]: **Intuition:**

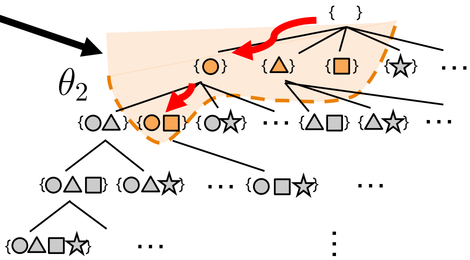
$$\hat{\psi}(\mathcal{S}) \geqslant p^j_{\min} = \mathcal{S} \text{ is } \textit{untestable} \;\Rightarrow\; \text{cannot improve } p^j_{\min}!$$

[9] A. Terada, K. Tsuda, and J. Sese. *Fast westfall-young permutation procedure for combinatorial regulation discovery*. ICBB, 2013.

(improved version[10] 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 $p^j_{\min} \leqslant \hat{\psi}(\theta)$





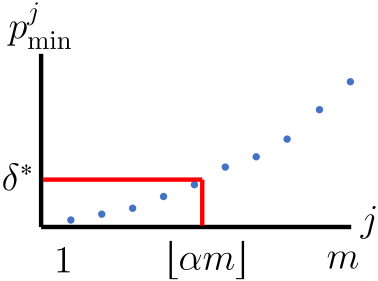
(imgs. from `LAMP`)

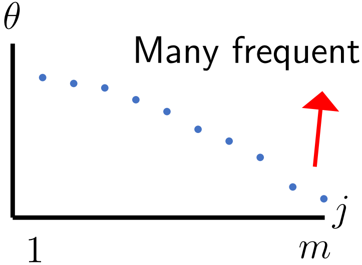
[10] T. Aika, H. Kim, and J. Sese. *High-speed westfall-young permutation procedure for genome-wide association studies*, ACM-BCB 2015.

Issues of `FASTWY`:

1) repeat the procedure m times ($m \simeq 10^3$-10^4 for $\alpha \simeq 0.05$);

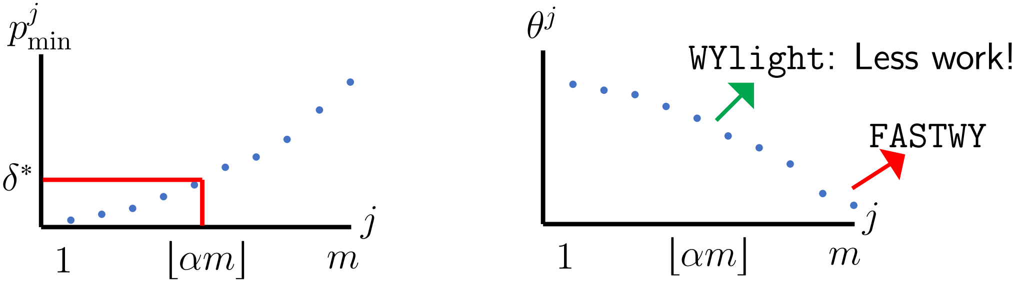
2) for some j, the min. p-value $p^j_{\min}$ is large $\rightarrow$ large space of testable patterns! (small freq. threshold θ)





WYlight

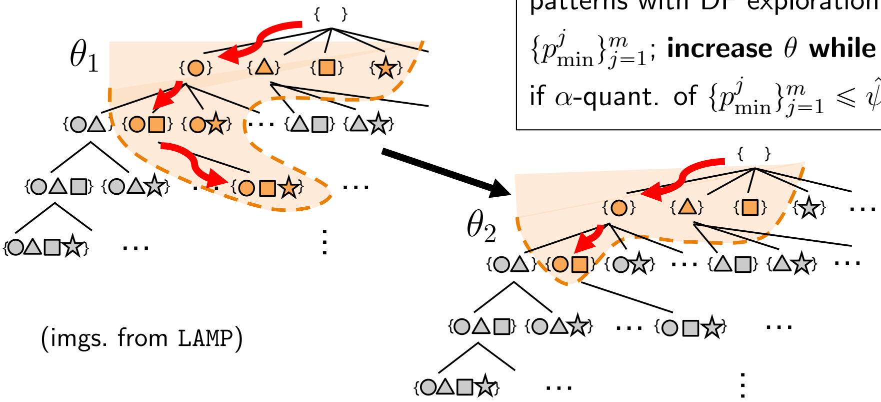
WYlight[11]: **Intuition:** to find δ^* we only need to **compute exactly the lower α-quantile of $\{p^j_{\min}\}_{j=1}^m$.**



[11] 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 m permutations at once.



(imgs. from `LAMP`)

start with $\theta = 1$ and $p^j_{\min} = 1, \forall j$; explore patterns with DF exploration, updating $\{p^j_{\min}\}_{j=1}^m$; **increase θ while exploring** if α-quant. of $\{p^j_{\min}\}_{j=1}^m \leqslant \hat{\psi}(\theta)$

Too many results!

Motivation: for many datasets, impractically large set of results ($SP(0.05)$) are found even when controlling $FWER \leq 0.05$:

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	$> 10^7$

What if we want (quickly!) only the **top-k significant patterns**, with same guarantees on $FWER$?

[12] L. Pellegrina and F. Vandin. *Efficient mining of the most significant patterns with permutation testing*. KDD 2018, DAMI 2020.

What if we want (quickly!) only the **top-k significant patterns**, with same guarantees on $FWER$?

$p^k = k$-th smallest p-value of $\mathcal{S} \in \mathcal{H}$,
$\delta^* = \max\left\{x : \overline{FWER}(x) \leqslant \alpha\right\}$,
$\overline{\delta} = \min\left\{p^k, \delta^*\right\}$.

[12] L. Pellegrina and F. Vandin. *Efficient mining of the most significant patterns with permutation testing*. KDD 2018, DAMI 2020.

What if we want (quickly!) only the **top-k significant patterns**, with same guarantees on $FWER$?

$p^k = k$-th smallest p-value of $\mathcal{S} \in \mathcal{H}$,
$\delta^* = \max\left\{x : \overline{FWER}(x) \leqslant \alpha\right\}$,
$\overline{\delta} = \min\left\{p^k, \delta^*\right\}$.

Set of top-k significant patterns:

$$TKSP(\mathcal{D}, \mathcal{H}, \alpha, k) := \left\{\mathcal{S} : p_{\mathcal{S}} \leqslant \overline{\delta}\right\}.$$

[12] L. Pellegrina and F. Vandin. *Efficient mining of the most significant patterns with permutation testing*. KDD 2018, DAMI 2020.

What if we want (quickly!) only the **top-k significant patterns**, with same guarantees on $FWER$?

$p^k = k$-th smallest p-value of $\mathcal{S} \in \mathcal{H}$,
$\delta^* = \max\left\{x : \overline{FWER}(x) \leqslant \alpha\right\}$,
$\overline{\delta} = \min\left\{p^k, \delta^*\right\}$.

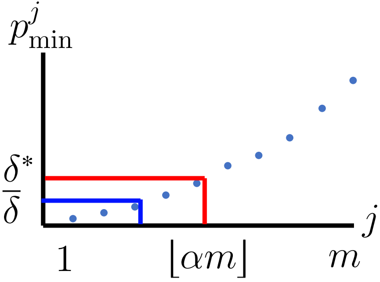
Set of top-k significant patterns:

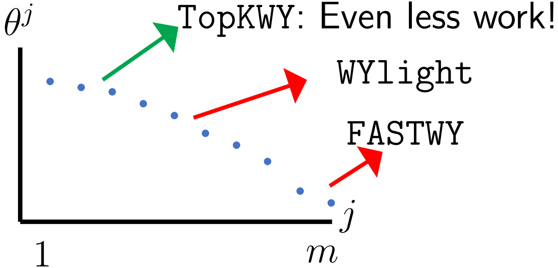
$$TKSP(\mathcal{D}, \mathcal{H}, \alpha, k) := \left\{\mathcal{S} : p_{\mathcal{S}} \leqslant \overline{\delta}\right\}.$$

Computed efficiently with `TopKWY`[12]!

[12] L. Pellegrina and F. Vandin. *Efficient mining of the most significant patterns with permutation testing*. KDD 2018, DAMI 2020.

Intuition: to compute $TKSP(\mathcal{D}, \mathcal{H}, \alpha, k)$ we only need to compute exactly the values of the set $\left\{p^j_{\min}\right\}_{j=1}^{m}$ that are $\leqslant \overline{\delta}$.

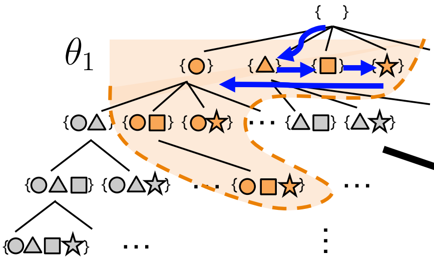


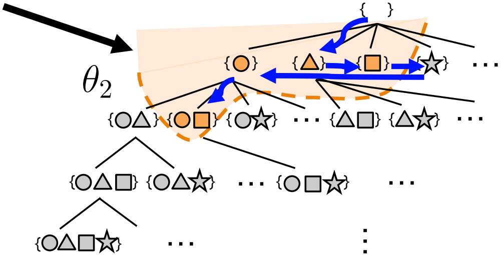


Algorithm: Best First (BF) exploration of $\mathcal{H}$ to compute $\overline{\delta}$.

(Approach similar to TopKMiner (Pietracaprina and Vandin, 2007) for **top-k freq. itemsets**).

start with $\theta = 1$ and $p^j_{\min} = 1, \forall j$; explore patterns with BF exploration, updating $\{p^j_{\min}\}_{j=1}^m$ and p^k; **increase θ while exploring** if $\min\left\{\alpha\text{-quant. of } \{p^j_{\min}\}_{j=1}^m \;,\; p^k\right\} \leqslant \hat{\psi}(\theta)$





(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 $FP(\mathcal{D}, \mathcal{H}, \theta^*) = \mathcal{T}(\overline{\delta})$.

Instead, the DF search *always* explores a super-set of $\mathcal{T}(\overline{\delta})$.

[13] L. Pellegrina, F. Vandin, *Efficient mining of the most significant patterns with permutation testing*. KDD 2018, DAMI 2020.

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 $FP(\mathcal{D}, \mathcal{H}, \theta^*) = \mathcal{T}(\overline{\delta})$.

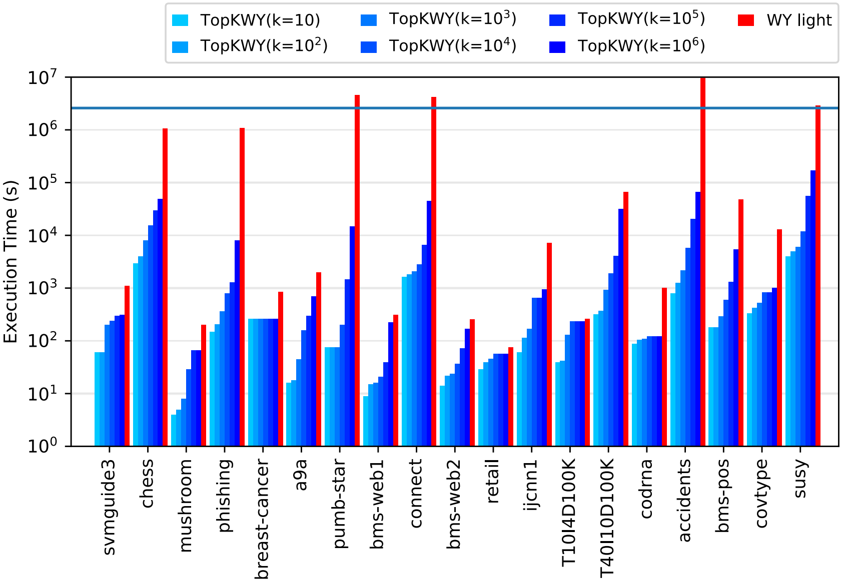
Instead, the DF search *always* explores a super-set of $\mathcal{T}(\overline{\delta})$.

2) Improved bounds to *skip* the processing of the permutations for many patterns.

(More details on the paper[13] ☺)

[13] L. Pellegrina, F. Vandin, *Efficient mining of the most significant patterns with permutation testing*. KDD 2018, DAMI 2020.

TopKWY: Running time



Outline

Recent developments and advanced topics

1. Controlling the FDR
2. Covariate-adaptive methods
3. Relaxing all conditional assumptions

More details and references at

`http://rionda.to/statdmtut`

Outline

Final Remarks

Knowledge Discovery 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

Hypothesis Testing and Statistically-sound Pattern Mining

Tutorial — SDM'21

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)

Tutorial webpage: `http://rionda.to/statdmtut`

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

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.

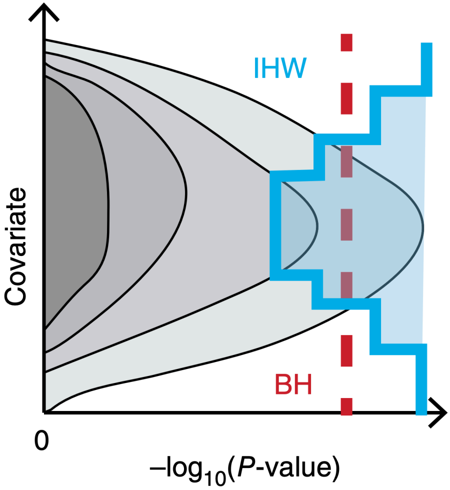
Example: the support $\sigma(\mathcal{S})$ of $\mathcal{S}$ has an impact on its minimum achivable p-value for Fisher's exact test

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)[14]

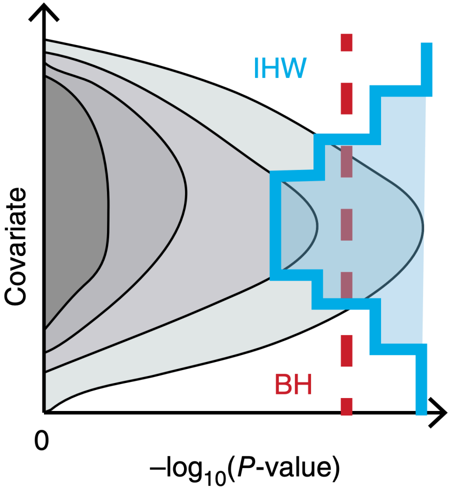
[14] Ignatiadis, Nikolaos, et al. *Data-driven hypothesis weighting increases detection power in genome-scale multiple testing.* Nature methods 13.7 (2016): 577.

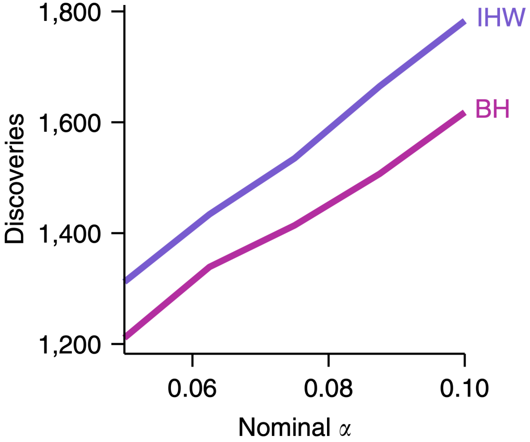
Independent Hypothesis Weighting (IHW)[14]



[14] 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)[14]





[14] Ignatiadis, Nikolaos, et al. *Data-driven hypothesis weighting increases detection power in genome-scale multiple testing.* Nature methods 13.7 (2016): 577.

Outline

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).

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?