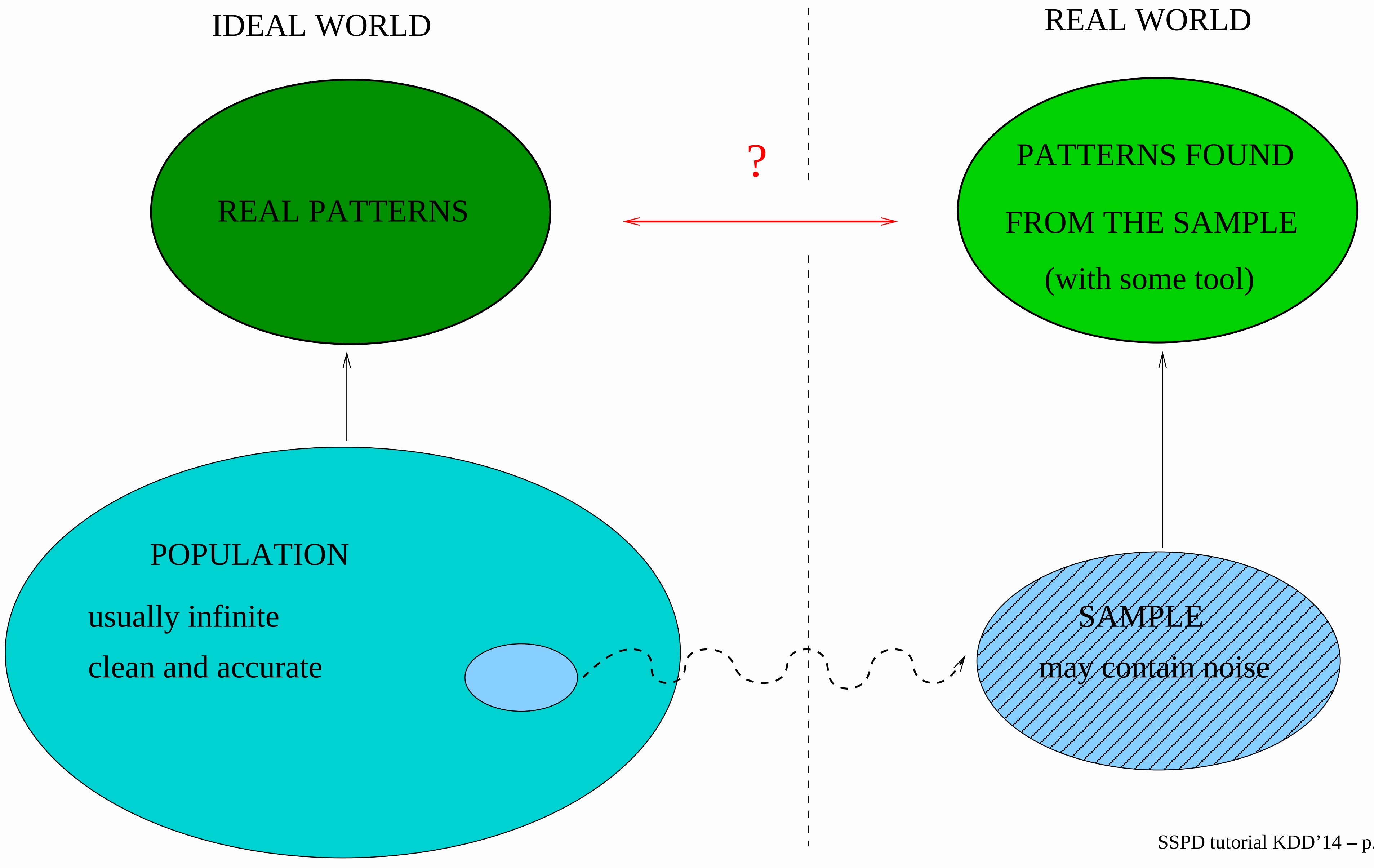
STATISTICALLY SOUND PATTERN DISCOVERY

Wilhelmiina Hämäläinen
University of Eastern Finland
whamalai@cs.uef.fi

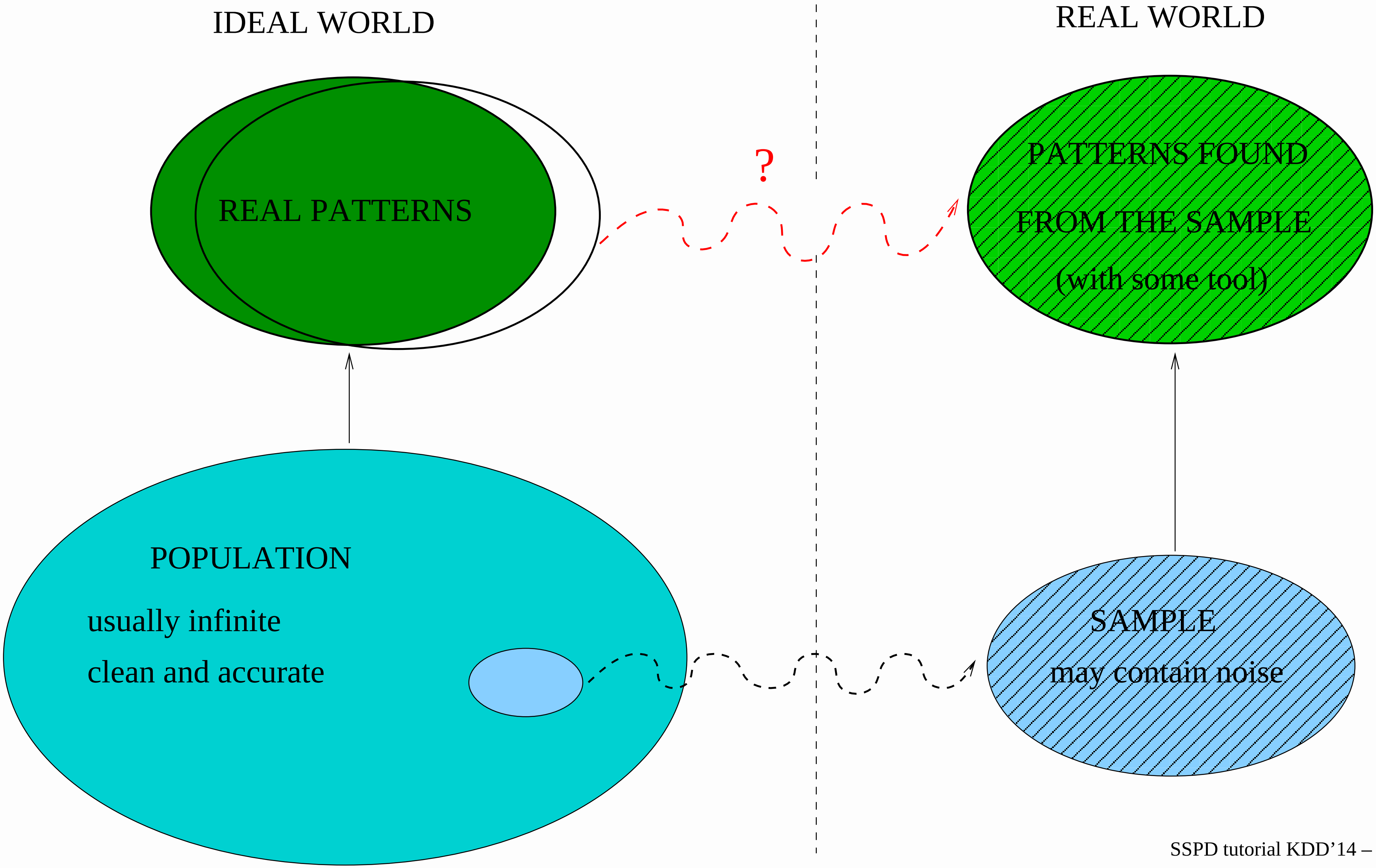
Geoff Webb
Monash University
Australia
geoff.webb@monash.edu

`http://www.cs.joensuu.fi/pages/whamalai/kdd14/sspdtutorial.html`

Statistically sound pattern discovery: Problem



Statistically sound pattern discovery: Problem



Statistically Sound vs. Unsound DM?

Pattern-type-first:
Given a desired classical pattern, invent a search method.

Method-first:
Invent a new pattern type which has an easy search method

e.g., an antimonotonic "interestingness" property

Tricks to sell it:

- overload statistical terms
- don't specify exactly

Statistically Sound vs. Unsound DM?

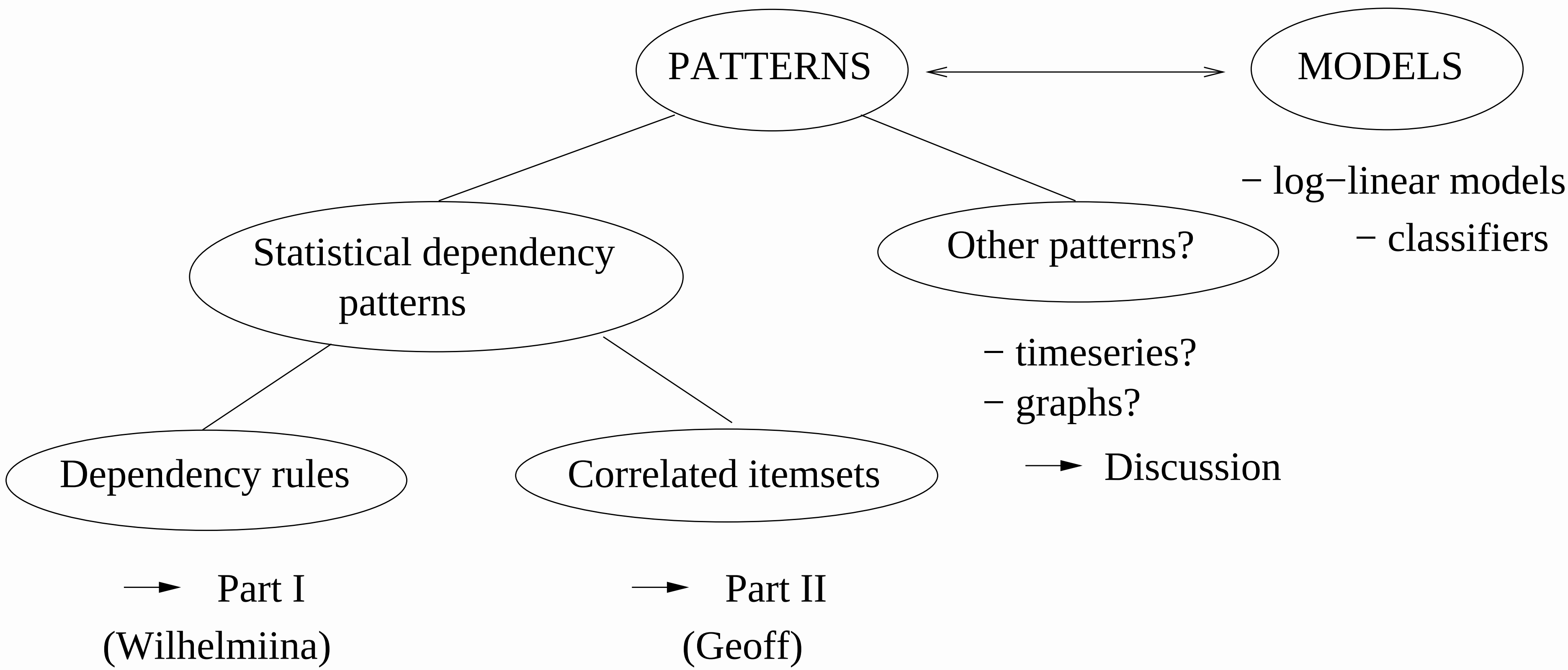
Pattern-type-first:
Given a desired classical pattern, invent a search method.

Method-first:
Invent a new pattern type which has an easy search method

- \+ easy to interprete correctly
- \+ informative
- \+ likely to hold in future
- – computationally demanding

- – difficult to interprete
- – misleading “information”
- – no guarantees on validity
- \+ computationally easy

Statistically sound pattern discovery: Scope



Contents

Statistical dependence: Many interpretations!

Events $(X = x)$ and $(Y = y)$ are statistically **independent**, if $P(X = x, Y = y) = P(X = x)P(Y = y)$.

- When variables (or variable-value combinations) are statistically **dependent**?
- When the dependency is genuine? $\rightarrow$ measures for the **strength** and **significance** of dependence
- How to define mutual dependence between three or more variables?

Statistical dependence: 3 main interpretations

Let A, B, C binary variables. Notate $\neg A \equiv (A = 0)$ and $A \equiv (A = 1)$

1. **Dependency rule** $AB \rightarrow C$: must be $\delta = P(ABC) - P(AB)P(C) > 0$ (positive dependence).
2. **Full probability model**:
   $\delta_1 = P(ABC) - P(AB)P(C)$,
   $\delta_2 = P(A\neg BC) - P(A\neg B)P(C)$,
   $\delta_3 = P(\neg ABC) - P(\neg AB)P(C)$ and
   $\delta_4 = P(\neg A\neg BC) - P(\neg A\neg B)P(C)$.
   - If $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$, no dependence
   - Otherwise decide from δ_i ($i = 1, \ldots, 4$) (with some equation)

Statistical dependence: 3 interpretations

3. **Correlated set** ABC
   - Starting point mutual independence: $P(A=a,B=b,C=c)=P(A=a)P(B=b)P(C=c)$ for all $a,b,c\in\{0,1\}$
   - different variations (and names)! e.g.
     (i) $P(ABC)>P(A)P(B)P(C)$ (positive dependence) or
     (ii) $P(A=a,B=b,C=c)\neq P(A=a)P(B=b)P(C=c)$ for some $a,b,c\in\{0,1\}$

   + extra criteria

In addition, **conditional independence** sometimes useful
$P(B=b,C=c|A=a)=P(B=b|A=a)P(C=c|A=a)$

Statistical dependence: no single correct definition

One of the most important problems in the philosophy of natural sciences is – in addition to the well-known one regarding the essence of the concept of probability itself – **to make precise the premises** *which would make it possible* **to regard any given real events as independent***.*

A.N. Kolmogorov

Part I Contents

1. Statistical dependency rules
2. Variable- and value-based interpretations
3. Statistical significance testing
   3.1 Approaches
   3.2 Sampling models
   3.3 Multiple testing problem
4. Redundancy and significance of improvement
5. Search strategies

1. Statistical dependency rules

Requirements for a genuine statistical dependency rule $X \rightarrow A$:

(i) Statistical dependence

(ii) Statistically significant

- likely not due to chance

(iii) Non-redundant

- not a side-product of another dependency
- added value

Why?

Example: Dependency rules on atherosclerosis

1. Statistical dependencies:
   smoking $\rightarrow$ atherosclerosis
   sports $\rightarrow \neg$ atherosclerosis
   ABCA1-R219K $\perp\!\!\!\perp$ atherosclerosis **?**
2. Statistical significance?
   spruce sprout extract $\rightarrow \neg$ atherosclerosis **?**
   dark chocolate $\rightarrow \neg$ atherosclerosis
3. Redundancy?
   stress, smoking $\rightarrow$ atherosclerosis
   smoking, coffee $\rightarrow$ atherosclerosis **?**
   high cholesterol, sports $\rightarrow$ atherosclerosis **?**
   male, male pattern baldness $\rightarrow$ atherosclerosis **?**

Part I Contents

1. Statistical dependency rules
2. **Variable- and value-based interpretations**
3. Statistical significance testing
   3.1 Approaches
   3.2 Sampling models
   3.3 Multiple testing problem
4. Redundancy and significance of improvement
5. Search strategies

2. Variable-based vs. Value-based interpretation

Meaning of dependency rule $X \rightarrow A$

1. Variable-based: dependency between binary **variables** X and A
   - Positive dependency $X \rightarrow A$ the same as $\neg X \rightarrow \neg A$
   - Equally strong as negative dependency between X and $\neg A$ (or $\neg X$ and A)
2. Value-based: positive dependency between **values** $X = 1$ and $A = 1$
   - different from $\neg X \rightarrow \neg A$ which may be weak!

Strength of statistical dependence

The most common measures:

1. Variable-based: leverage

$$\delta(X, A) = P(XA) - P(X)P(A)$$

2. Value-based: lift

$$\gamma(X, A) = \frac{P(XA)}{P(X)P(A)} = \frac{P(A|X)}{P(A)} = \frac{P(X|A)}{P(X)}$$

$P(A|X)$ = “confidence” of the rule
Remember: $X \equiv (X = 1)$ and $A \equiv (A = 1)$

Contingency table

	A	$\neg A$	All
X	$fr(XA) = n[P(X)P(A) + \delta]$	$fr(X\neg A) = n[P(X)P(\neg A) - \delta]$	$fr(X)$
$\neg X$	$fr(\neg XA) = n[P(\neg X)P(A) - \delta]$	$fr(\neg X\neg A) = n[P\neg(X)P(\neg A) + \delta]$	$fr(\neg X)$
All	$fr(A)$	$fr(\neg A)$	n

All value combinations have the same $|\delta|$!
$\Leftrightarrow \gamma$ depends on the value combination

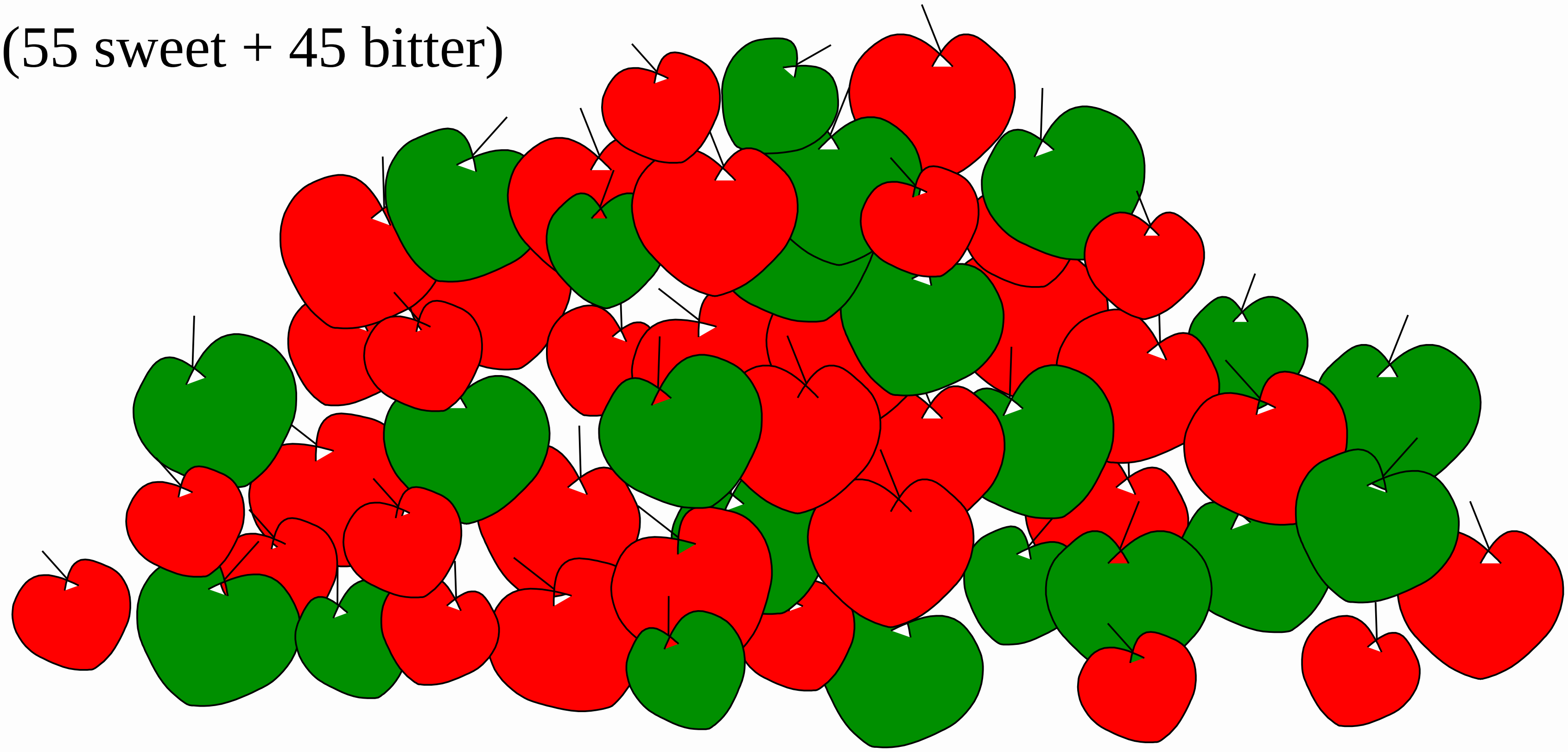
$fr(X)$=absolute frequency of X
$P(X)$=relative frequency of X

Example: The Apple problem

Variables: Taste, smell, colour, size, weight, variety, grower, . . .

100 apples

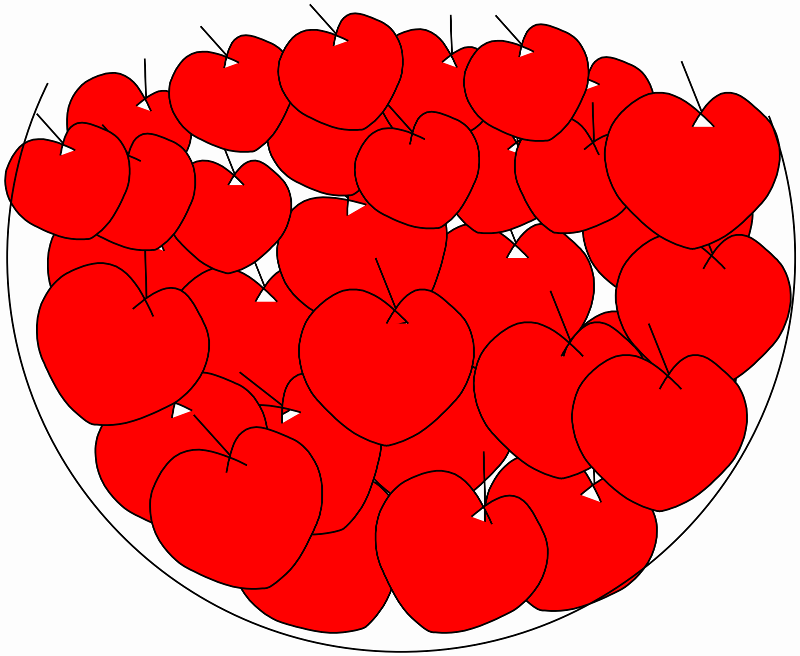
(55 sweet + 45 bitter)



Rule RED → SWEET ($Y \rightarrow A$)

$P(A|Y) = 0.92$, $P(\neg A|\neg Y) = 1.0$
$\delta = 0.22$, $\gamma = 1.67$

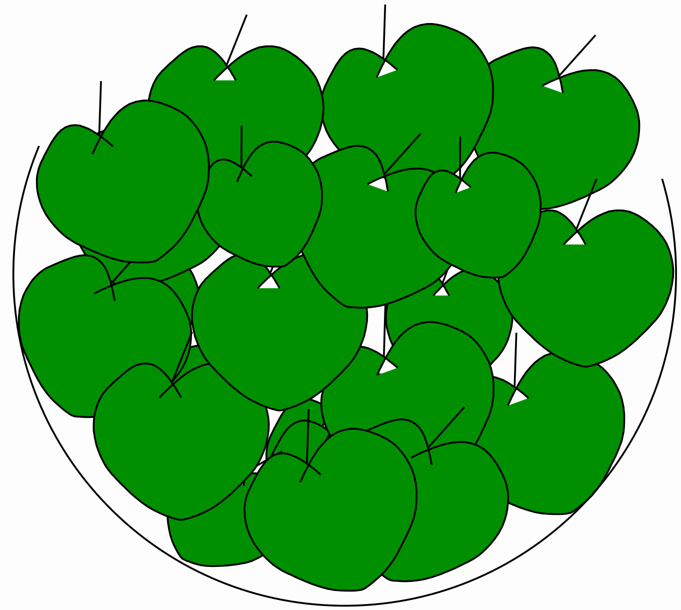
A=sweet, $\neg A$=bitter
Y=red, $\neg Y$=green



Basket 1

60 red apples

(55 sweet)



Basket 2

40 green apples

(all bitter)

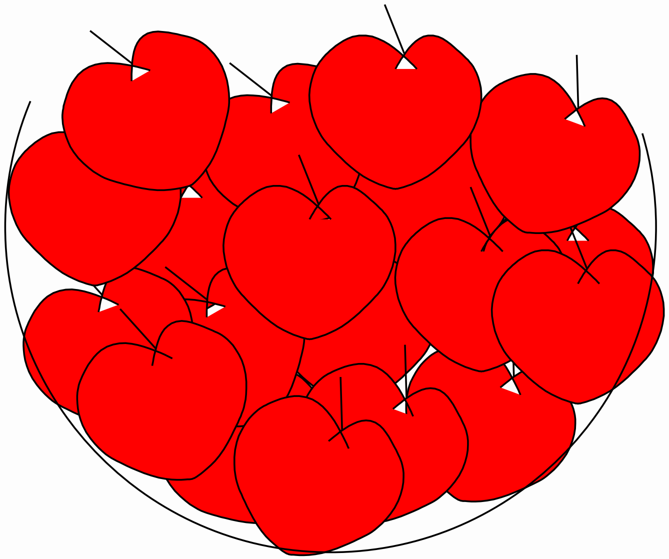
Rule RED and BIG → SWEET ($X \rightarrow A$)

$P(A|X) = 1.0$, $P(\neg A|\neg X) = 0.75$

$\delta = 0.18$, $\gamma = 1.82$

X=(red ∧ big)

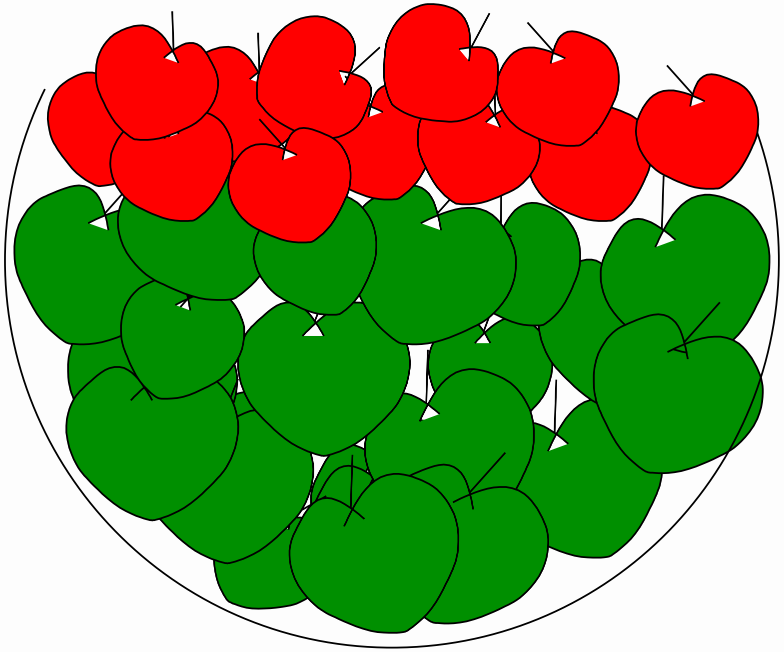
$\neg X$=(green ∨ small)



Basket 1

40 large red apples

(all sweet)



Basket 2

40 green + 20 small red apples

(45 bitter)

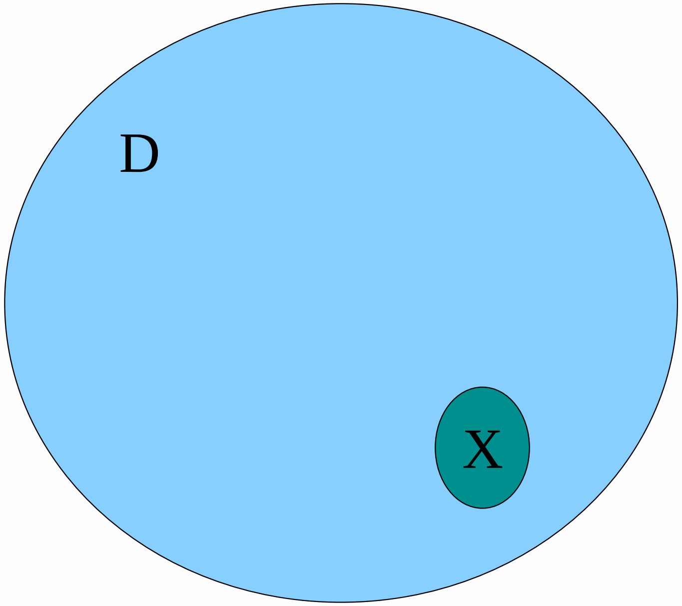
When the value-based interpretation could be useful? Example

D=disease, X=allele combination
$P(X)$ small and $P(D|X) = 1.0$

$\Rightarrow \gamma(X, D) = P(D)^{-1}$ can be **large**

$P(D|\neg X) \approx P(D)$
$P(\neg D|\neg X) \approx P(\neg D)$

$\Rightarrow \delta(X, D) = P(X)P(\neg D)$ **small**.



Now dependency strong in the value-based but weak in the variable-based interpretation!

(Usually, **variable-based** dependencies tend to be more reliable)

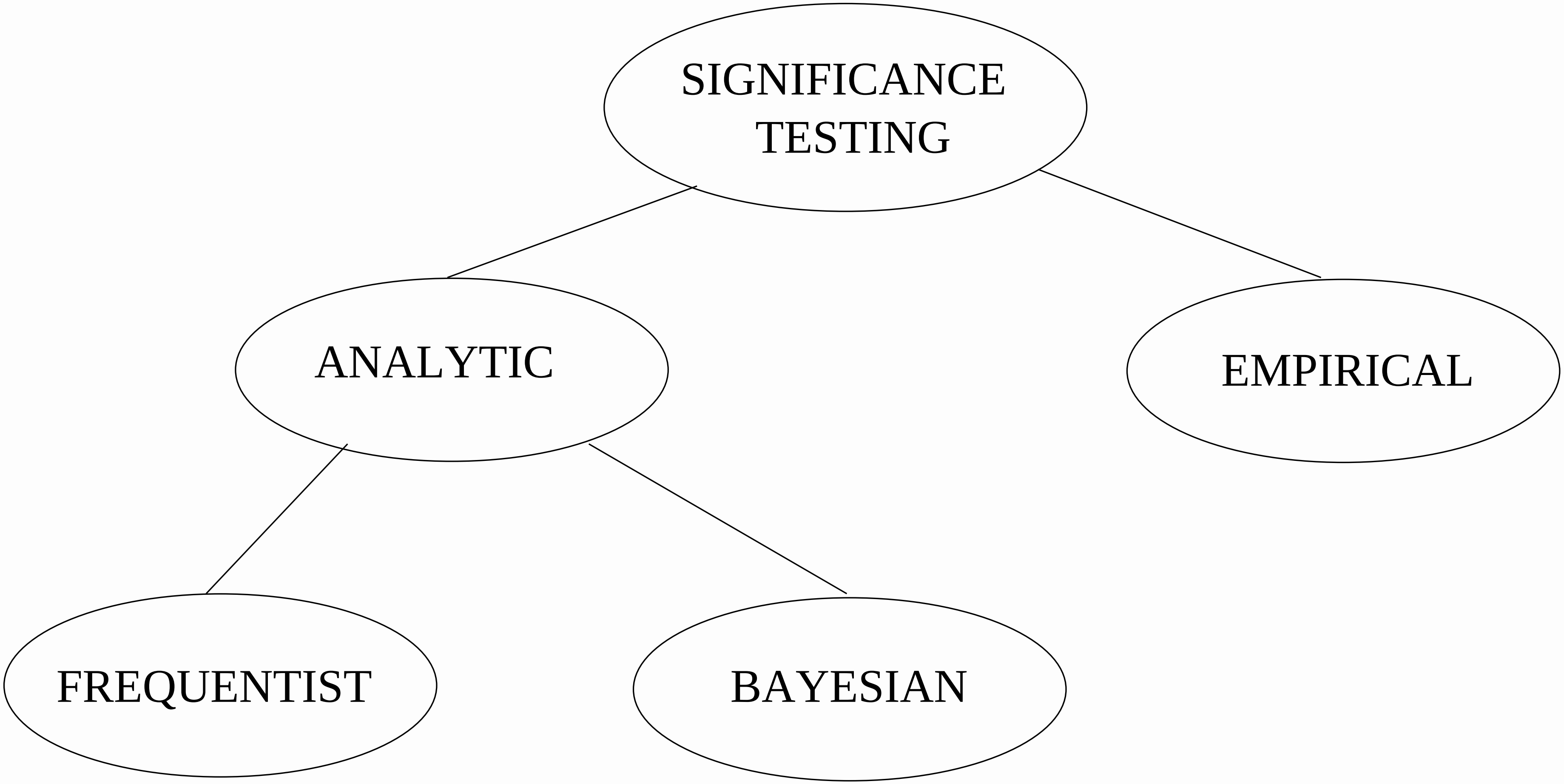
Part I Contents

3. Statistical significance of $X \rightarrow A$

What is the probability of the observed or a stronger dependency, if X and A were independent? If small probability, then $X \rightarrow A$ likely genuine (not due to chance).

- Significant $X \rightarrow A$ is likely to hold in future (in similar data sets)
- How to estimate the probability??
- How small the probability should be?
  - Fisherian vs. Neyman-Pearsonian schools
  - multiple testing problem

3.1 Main approaches



different schools

different sampling models

Analytic approaches

H_0: X and A independent (null hypothesis)
H_1: X and A positively dependent (research hypothesis)

- Frequentist: Calculate
  $p = P(\text{observed or stronger dependency}|H_0)$
- Bayesian:
  (i) Set $P(H_0)$ and $P(H_1)$
  (ii) Calculate $P(\text{observed or stronger dependency}|H_0)$ and $P(\text{observed or stronger dependency}|H_1)$
  (iii) Derive (with Bayes' rule) $P(H_0|\text{observed or stronger dependency})$ and $P(H_1|\text{observed or stronger dependency})$

Analytic approaches: pros and cons

+ p-values relatively fast to calculate
+ can be used as search criteria
– How to define the distribution under H_0? (assumptions)
– If data not representative, the discoveries cannot be generalized to the whole population
  - describe only the sample data or other similar samples
  - random samples not always possible (infinite population)

Note: Differences between Fisherian vs. Neyman-Pearsonian schools

- significance testing vs. hypothesis testing
- role of nominal p-values (thresholds 0.05, 0.01)
- many textbooks represent a hybrid approach

→ see Hubbard & Bayarri

Empirical approach (randomization testing)

Generate random data sets according to H_0 and test how many of them contain the observed or stronger dependency $X \rightarrow A$.

(i) Fix a permutation scheme (how to express H_0 + which properties of the original data should hold)

(ii) Generate a random subset $\{d_1, \ldots, d_b\}$ of all possible permutations

(iii)

$$p = \frac{|\{d_i | \text{contains observed or stronger dependency}\}|}{b}$$

Empirical approach: pros and cons

- \+ no assumptions on any underlying parametric distribution
- \+ can test null hypotheses for which no closed form test exists
- \+ offers an approach to multiple testing problem $\rightarrow$ Later
- \+ data doesn't have to be a random sample
  $\rightarrow$ discoveries hold for the whole population ...
- – ... **defined by the permutation scheme**
- – often not clear (but critical), how to permutate data!
- – computationally heavy (b: efficiency vs. quality trade-off)
- – How to apply during search??

Note: Randomization test vs. Fisher's exact test

When testing significance of $X \rightarrow A$

- a natural permutation scheme fixes $N = n$, $N_X = fr(X)$, $N_A = fr(A)$
- randomization test generates some random contingency tables with these constraints
- full permutation test = Fisher's exact test studies all contingency tables
  - faster to compute (analytically)
  - produces more reliable results

$\Rightarrow$ No need for randomization tests, here!

Part I Contents

3.2 Sampling models

= defining the distribution under H_0

$\leftarrow$ What do we assume fixed?

- Variable-based dependencies: classical sampling models (Statistics)
- Value-based dependencies: several suggestions (Data mining)

Basic idea

Given a sampling model $\mathcal{M}$
$\mathcal{T}$=set of all possible contingency tables.

1. Define probability $P(T_i|\mathcal{M})$ for contingency tables $T_i \in \mathcal{T}$
2. Define an **extremeness relation** $T_i \succeq T_j$
   - T_i contains at least as strong dependency $X \rightarrow A$ as T_j does
   - depends on the strength measure, e.g. δ (var-based) or γ (val-based)
3. Calculate $p = \sum_{T_i \succeq T_0} P(T_i|\mathcal{M})$
   (T_0=our table)

Sampling models for variable-based dependencies

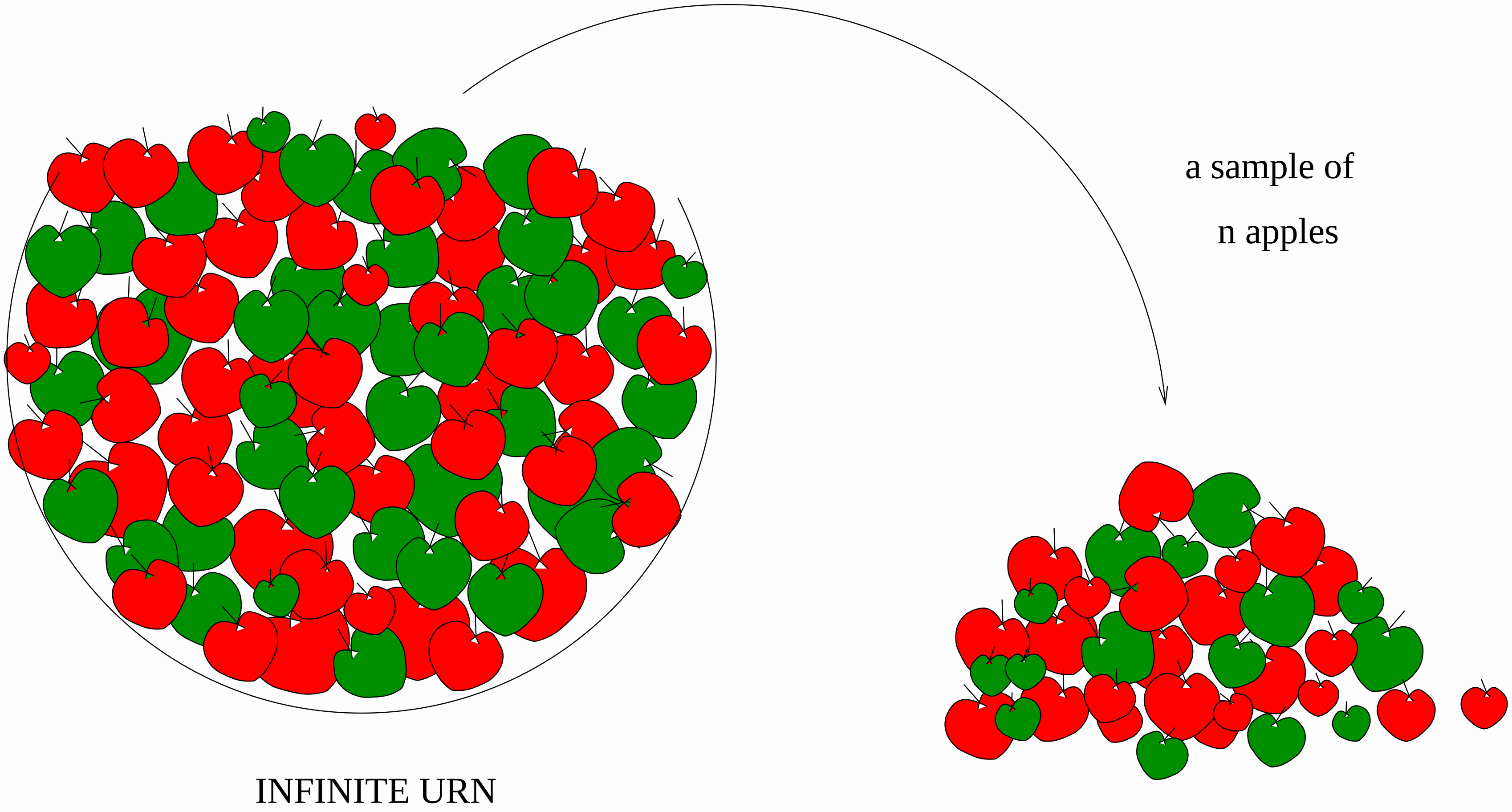
3 basic models:

1. Multinomial ($N = n$ fixed)
2. Double binomial ($N = n$, $N_X = fr(X)$ fixed)
3. Hypergeometric ($\rightarrow$ Fisher's exact test)
   ($N = n$, $N_A = fr(A)$, $N_X = fr(X)$ fixed)

+ asymptotic measures (like χ^2)

Multinomial model

Independence assumption: In the infinite urn, $p_{XA} = p_X p_A$.
(p_{XA}=probability of red sweet apples)



Multinomial model

T_i is defined by random variables N_{XA}, $N_{X\neg A}$, $N_{\neg XA}$, $N_{\neg X\neg A}$

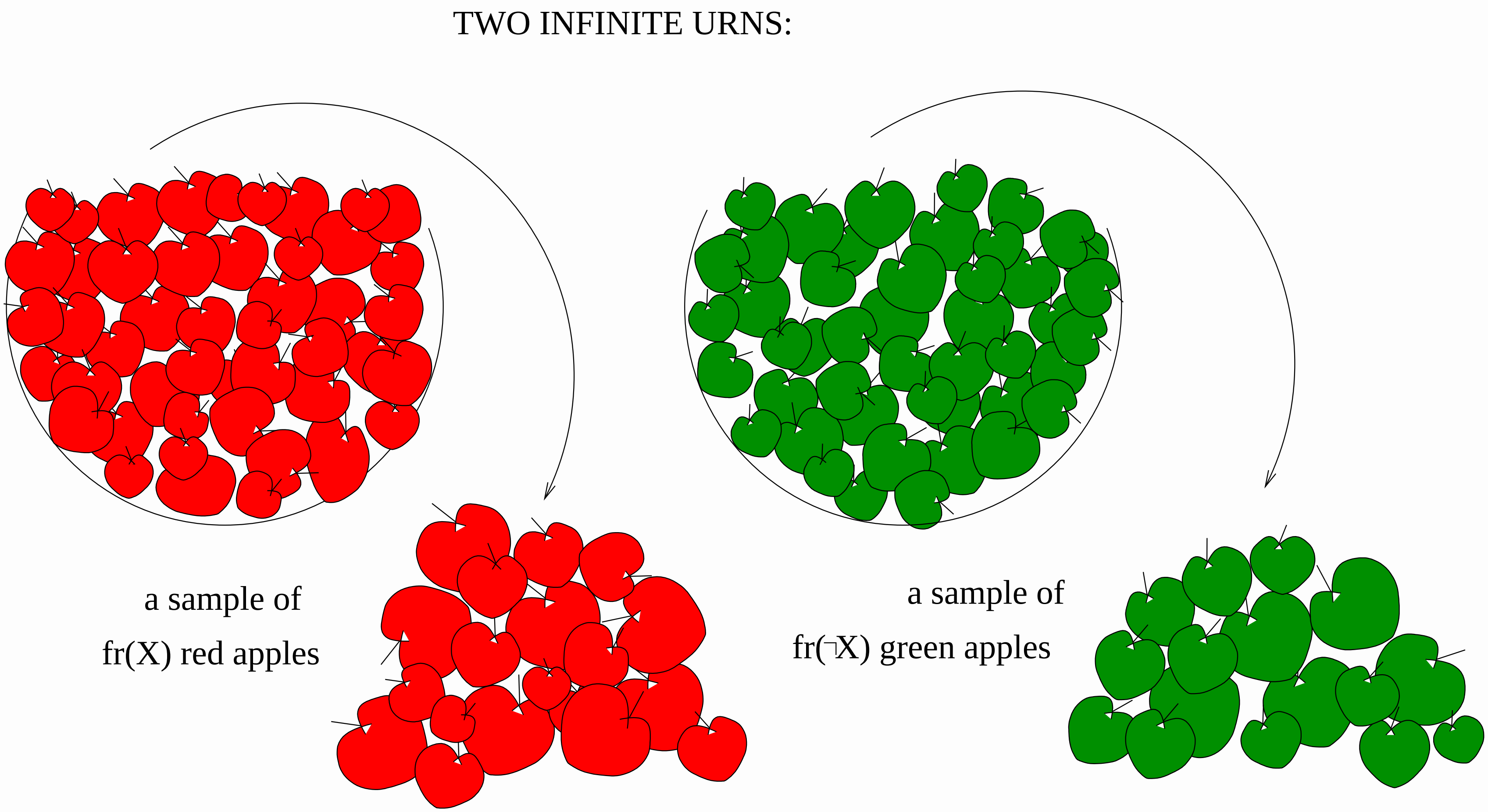
$$P(N_{XA}, N_{X\neg A}, N_{\neg XA}, N_{\neg X\neg A} | n, p_X, p_A) =$$
$$\binom{n}{N_{XA}, N_{X\neg A}, N_{\neg XA}, N_{\neg X\neg A}} p_X^{N_X}(1-p_X)^{n-N_X} p_A^{N_A}(1-p_A)^{n-N_A}.$$

$$p = \sum_{T_i \succeq T_0} P(N_{XA}, N_{X\neg A}, N_{\neg XA}, N_{\neg X\neg A} | n, p_X, p_A)$$

- p_X and p_A can be estimated from the data

Double binomial model

Independence assumption: $p_{A|X} = p_A = p_{A|\neg X}$



Double binomial model

Probability of red sweet apples:

$$P(N_{XA}|fr(X), p_A) = \binom{fr(X)}{N_{XA}} p_A^{N_{XA}} (1 - p_A)^{fr(X) - N_{XA}}$$

Probability of green sweet apples:

$$P(N_{\neg XA}|fr(\neg X), p_A) = \binom{fr(\neg X)}{N_{\neg XA}} p_A^{N_{\neg XA}} (1 - p_A)^{fr(\neg X) - N_{\neg XA}}$$

Double binomial model

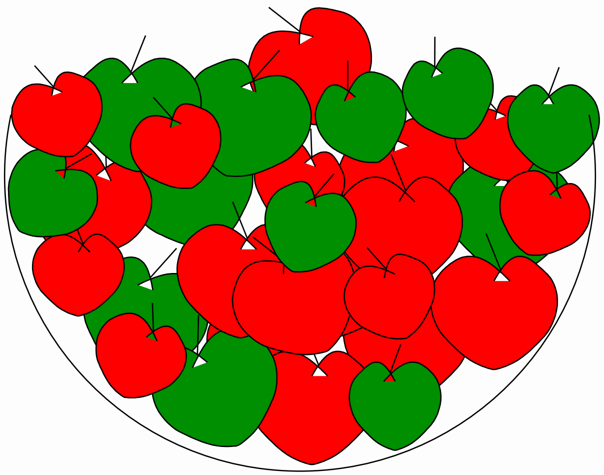
T_i is defined by variables N_{XA} and $N_{\neg XA}$.

$$P(N_{XA}, N_{\neg XA} | n, fr(X), fr(\neg X), p_A) = \binom{fr(X)}{N_{XA}}\binom{fr(\neg X)}{N_{\neg XA}} p_A^{N_A}(1-p_A)^{n-N_A}$$

$$p = \sum_{T_i \succeq T_0} P(N_{XA}, N_{\neg XA} | n, fr(X), fr(\neg X), p_A)$$

Hypergeometric model (Fisher's exact test)

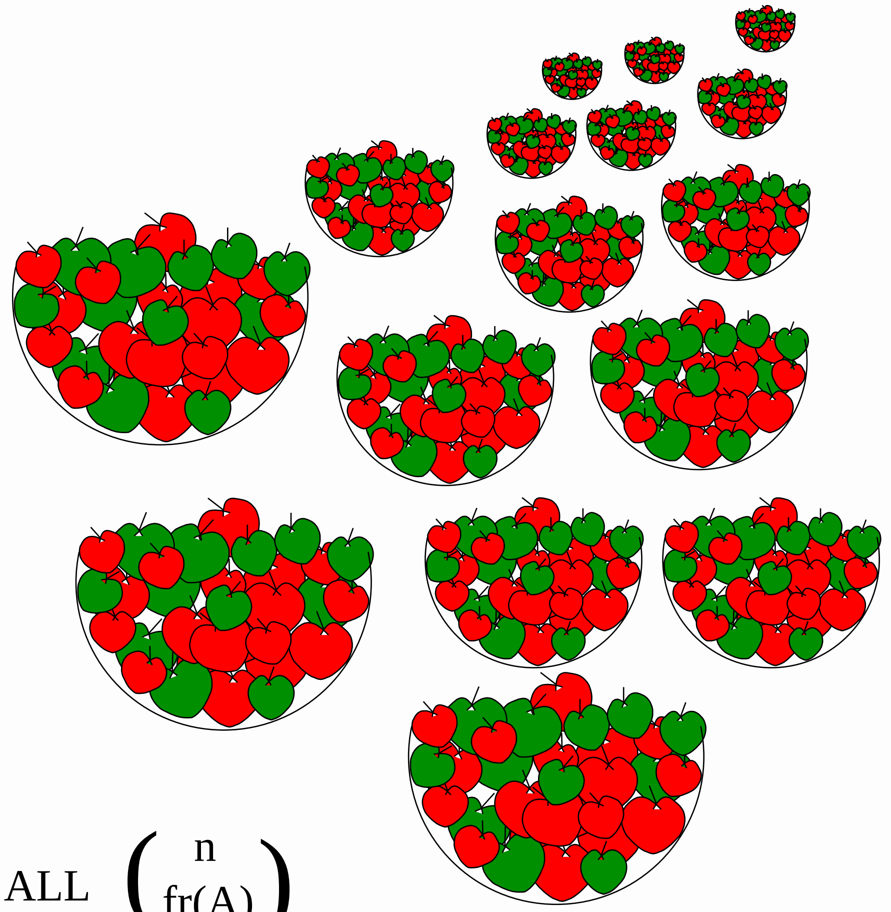
How many other similar urns have
at least as strong dependency as ours?



OUR URN   n apples

fr(A) sweet + fr(¬A) bitter

fr(X) red + fr(¬X) green



ALL $\binom{n}{fr(A)}$

SIMILAR URNS

Like in a full permutation test

	X						¬X			
	1	2	3	4	5	6	7	8	9	10
urn1	A	A	A	¬A	¬A	¬A	¬A	¬A	¬A	¬A
urn2	A	A	¬A	A	¬A	¬A	¬A	¬A	¬A	¬A
	A	A	¬A	¬A	A	¬A	¬A	¬A	¬A	¬A
⋮				⋮					⋮	
urn120	¬A	¬A	¬A	¬A	¬A	¬A	¬A	A	A	A

n=10

fr(X)=6

fr(A)=3

Hypergeometric model (Fisher's exact test)

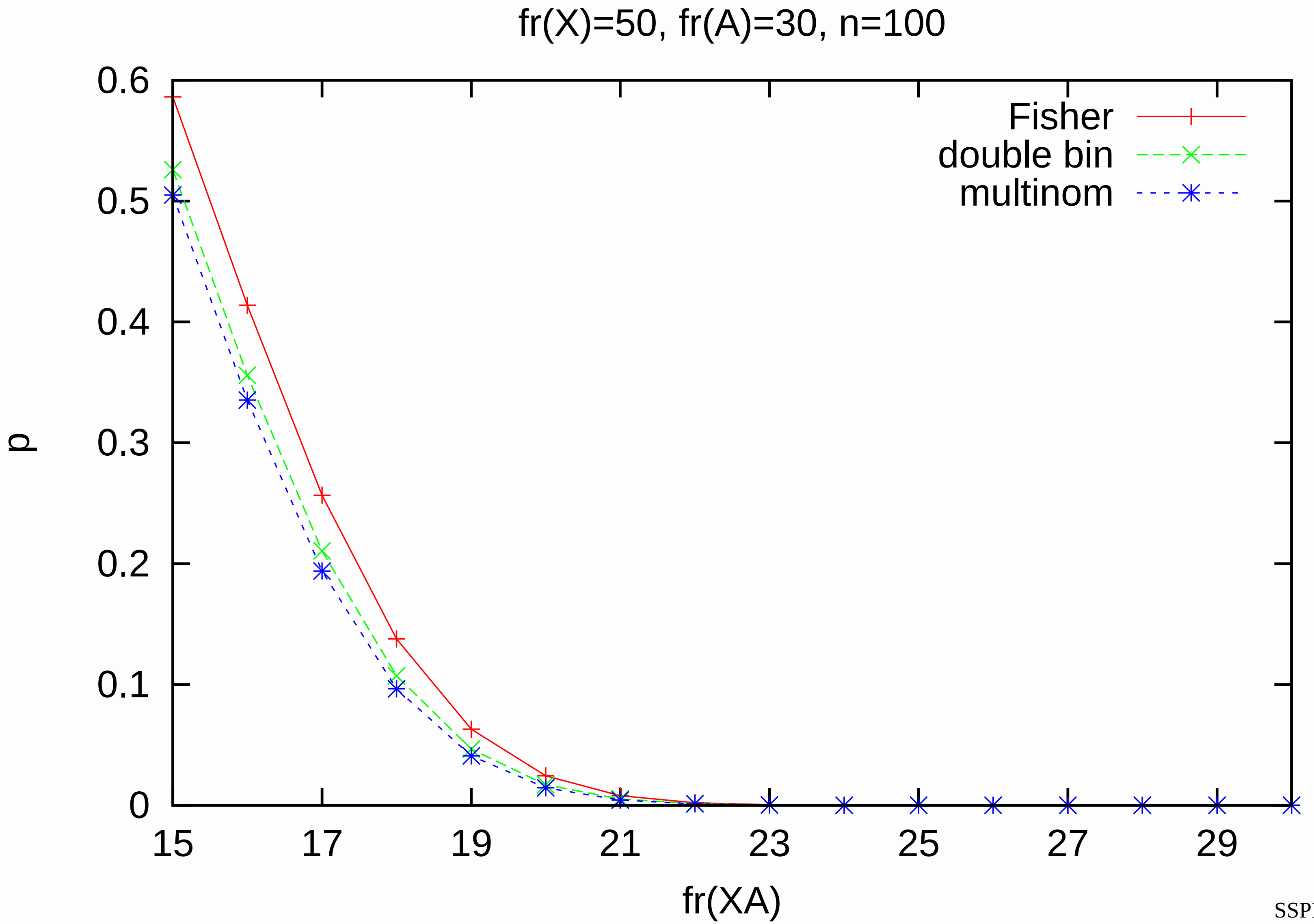
The number of all possible similar urns (fixed $N = n$, $N_X = fr(X)$ and $N_A = fr(A)$) is

$$\sum_{i=0}^{fr(A)} \binom{fr(X)}{i} \binom{fr(\neg X)}{fr(A) - i} = \binom{n}{fr(A)}$$

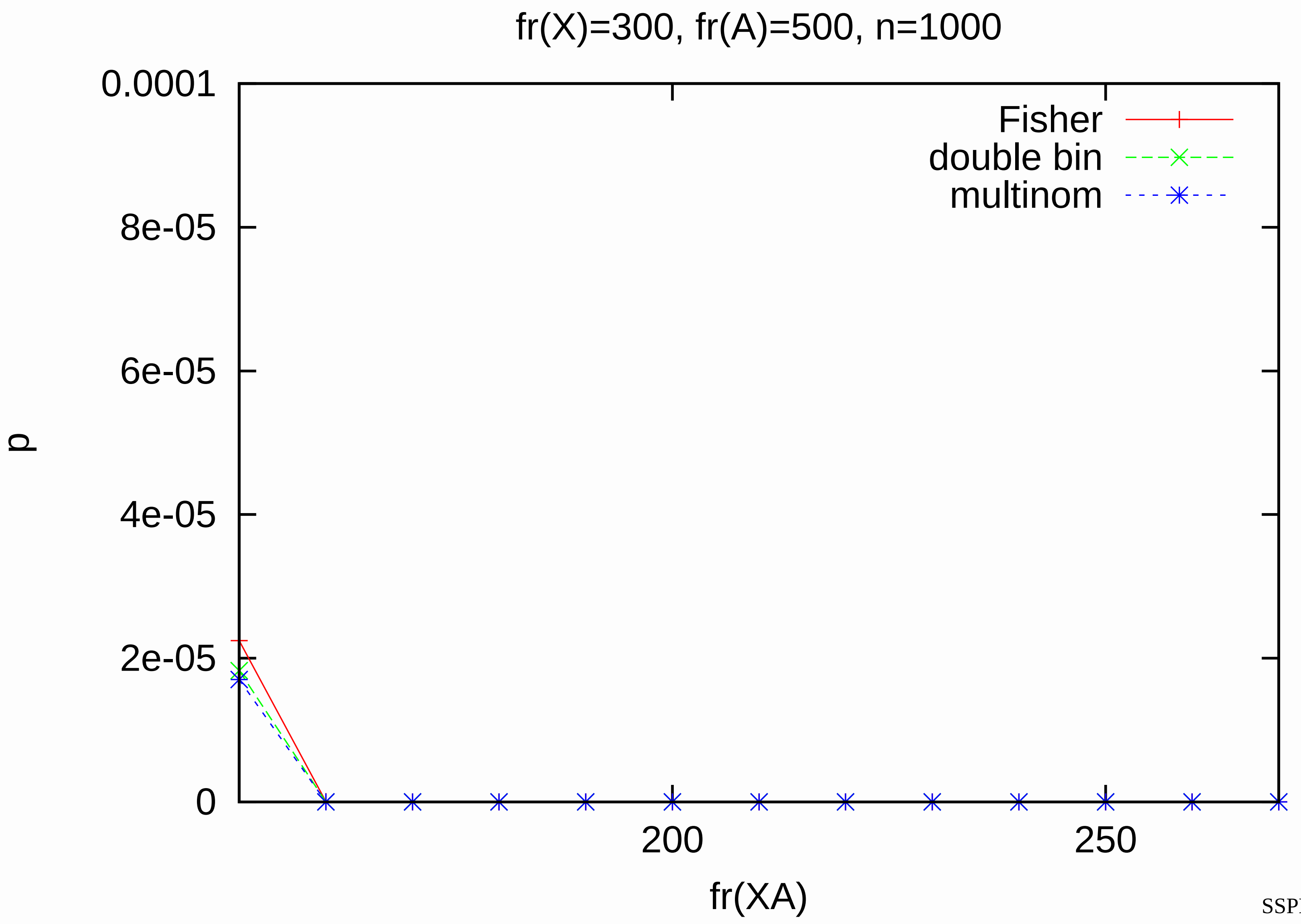
Now $(T_i \geq T_0) \equiv (N_{XA} \geq fr(XA))$. Easy!

$$p_F = \sum_{i=0}^{\infty} \frac{\binom{fr(X)}{fr(XA)+i} \binom{fr(\neg X)}{fr(\neg X \neg A)+i}}{\binom{n}{fr(A)}}$$

Example: Comparison of p-values



Example: Comparison of p-values



Example: Comparison of p-values

fr_{XA}	multi-nomial	double binomial	Fisher (hyperg.)
180	1.7e-05	1.8e-05	2.2e-05
200	2.3e-12	2.2e-12	3.0e-12
220	1.4e-22	7.3e-23	1.1e-22
240	2.9e-36	3.0e-37	4.4e-37
260	1.5e-53	4.2e-56	3.5e-56
280	1.3e-74	2.9e-80	1.6e-81
300	9.3e-100	3.5e-111	2.5e-119

Asymptotic measures

Idea: p-values are estimated indirectly

1. Select some “nicely behaving” measure M
   - e.g. M follows **asymptotically** the normal or the χ^2 distribution
2. Estimate $P(M \geq val)$, where $M = val$ in our data
   - Easy! (look at statistical tables)
   - But the accuracy can be poor

The χ^2-measure

$$\chi^2 = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{n(P(X=i,A=j) - P(X=i)P(A=j))^2}{P(X=i)P(A=j)}$$

$$= \frac{n(P(X,A) - P(X)P(A))^2}{P(X)P(\neg X)P(A)P(\neg A)} = \frac{n\delta^2}{P(X)P(\neg X)P(A)P(\neg A)}$$

- very sensitive to underlying assumptions!
- all $P(X=i)P(A=j)$ should be sufficiently large
- the corresponding hypergeometric distribution shouldn't be too skewed

Mutual information

$$MI = \log \frac{P(XA)^{P(XA)}P(X\neg A)^{P(X\neg A)}P(\neg XA)^{P(\neg XA)}P(\neg X\neg A)^{P(\neg X\neg A)}}{P(X)^{P(X)}P(\neg X)^{P(\neg X)}P(A)^{P(A)}P(\neg A)^{P(\neg A)}}$$

- $2n \cdot MI$=log likelihood ratio
- follows asymptotically the χ^2-distribution
- usually gives more reliable results than the χ^2-measure

Comparison: Sampling models for variable-based dependencies

- Multinomial: impractical but useful for theoretical results
- Double binomial: **not exchangeable**
  $p(X \rightarrow A) \neq p(A \rightarrow X)$ (in general)
- Hypergeometric (Fisher's exact test): recommended, enables efficient search, reliable results
- Asymptotic: often sensitive to underlying assumptions
  - χ^2 very sensitive, not recommended
  - MI reliable, enables efficient search, approximates p_F

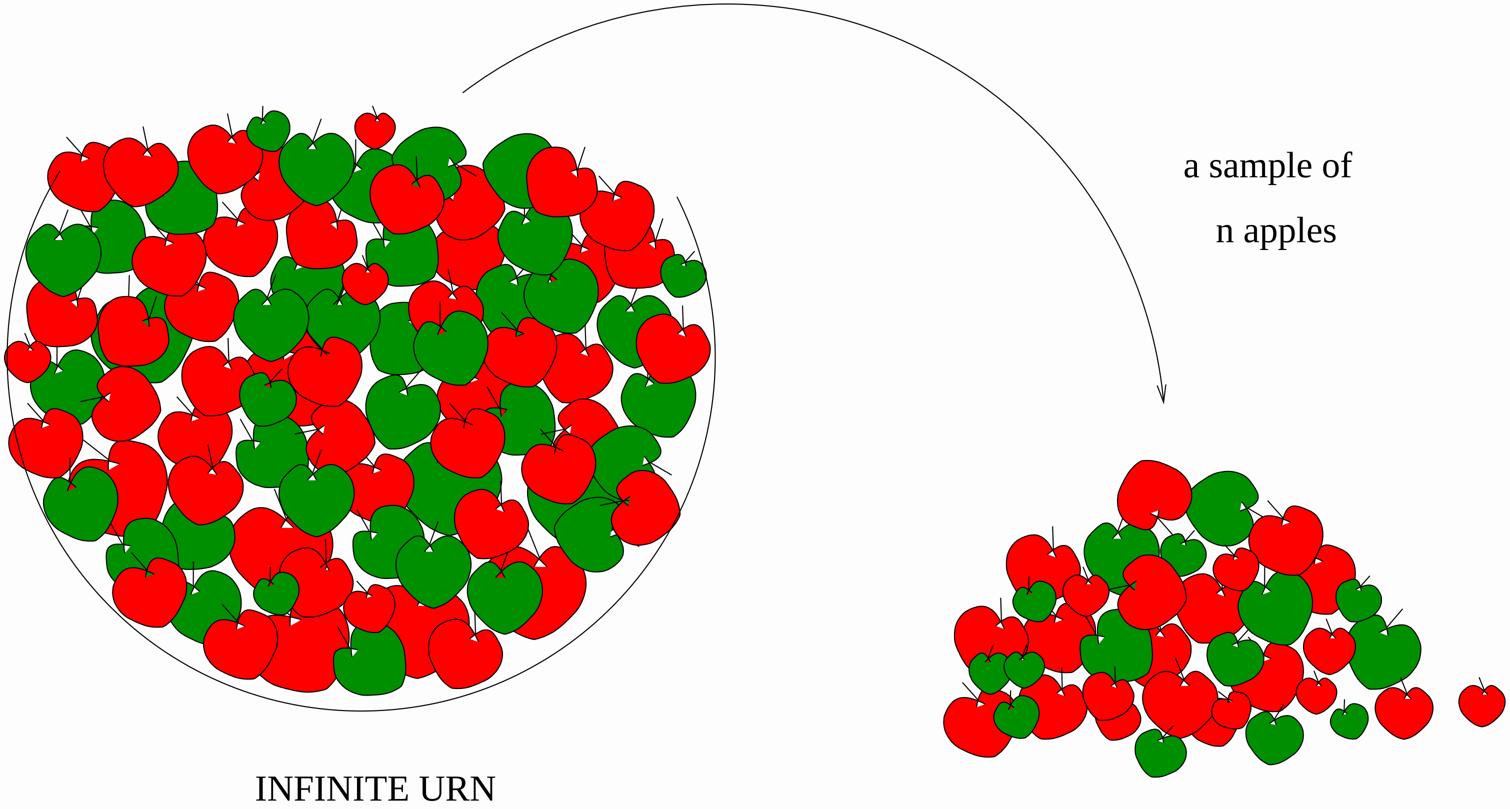
Sampling models for value-based dependencies

Main choices:

1. Classical sampling models but with a different extremeness relation
   - use lift γ to define a stronger dependency
   - Multinomial and Double binomial: can differ much from var-based
   - Hypergeometric: leads to Fisher's exact test, again!
2. Binomial models + corresponding asymptotic measures

Binomial model 1 (classical binomial test)

Probability of sweet red apples is $p_{XA} = p_X p_A$. If a random sample of n apples is taken, what is the probability to get $fr(XA)$ sweet red apples and $n - fr(XA)$ green or bitter apples?



Binomial model 1 (classical binomial test)

Probability of getting exactly N_{XA} sweet red apples and $n - N_{XA}$ green or bitter apples is

$$p(N_{XA}|n, p_{XA}) = \binom{n}{N_{XA}} (p_{XA})^{N_{XA}} (1 - p_{XA})^{n - N_{XA}}$$

$$p(N_{XA} \geq fr(XA)|n, p_{XA}) = \sum_{i=fr(XA)}^{n} \binom{n}{i} (p_{XA})^{i} (1 - p_{XA})^{n-i}$$

(or $i = fr(XA), \ldots, \min\{fr(X), fr(A)\}$)

- Use estimate $p_{XA} = P(X)P(A)$
- Note: N_X and N_A unfixed

Corresponding asymptotic measure

z-score:

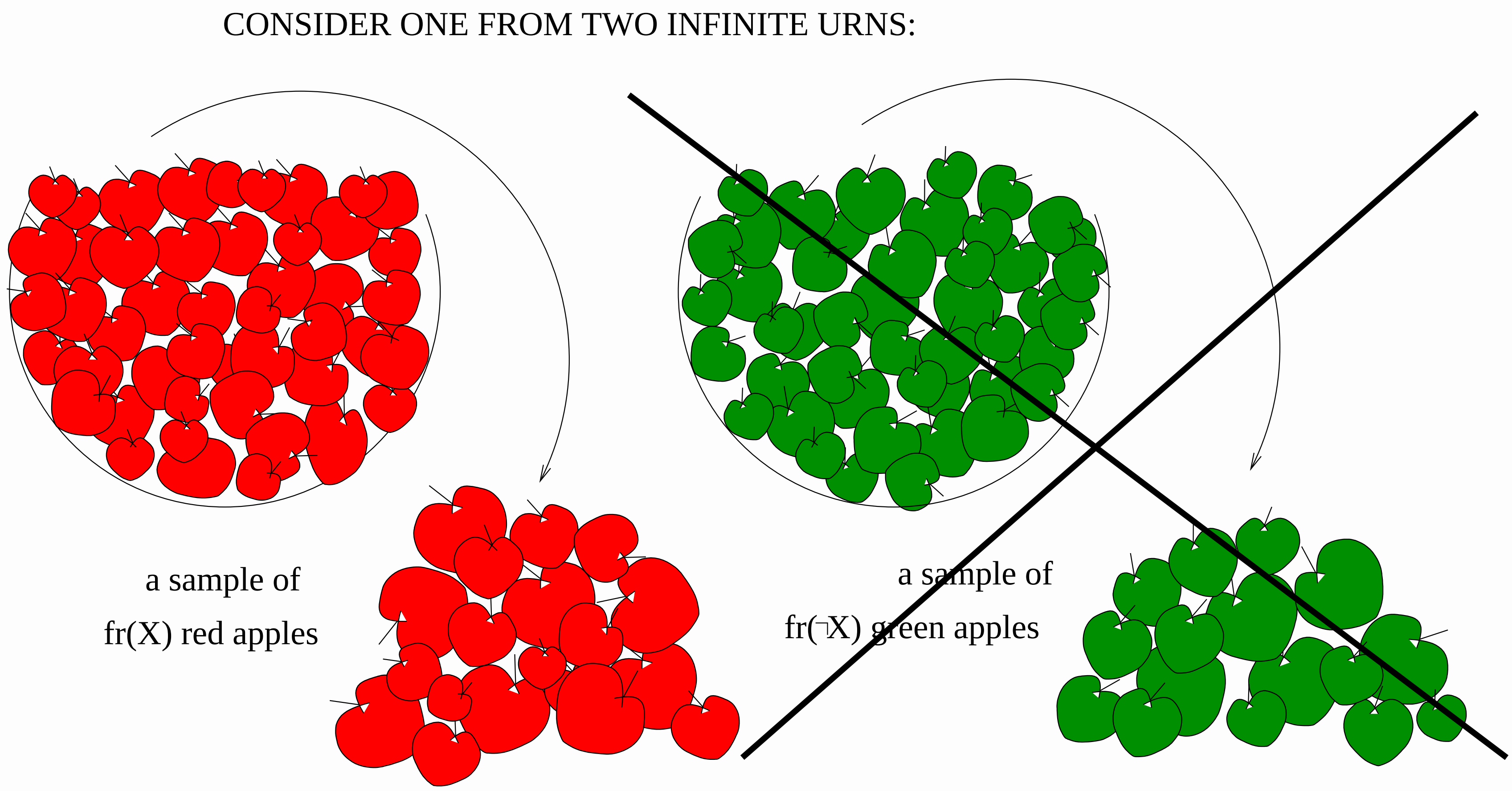
$$z_1(X \rightarrow A) = \frac{fr(X,A) - \mu}{\sigma} = \frac{fr(X,A) - nP(X)P(A)}{\sqrt{nP(X)P(A)(1 - P(X)P(A))}}$$

$$= \frac{\sqrt{n}\delta(X,A)}{\sqrt{P(X)P(A)(1 - P(X)P(A))}} = \frac{\sqrt{nP(XA)}(\gamma(X,A) - 1)}{\sqrt{\gamma(X,A) - P(X,A)}}.$$

- follows asymptotically the normal distribution

Binomial model 2 (suggested in DM)

Like the double binomial model, but forget the other urn!



Binomial model 2

$$p(N_{XA} \geq fr(XA) | fr(X), P(A)) = \sum_{i=fr(XA)}^{fr(X)} \binom{fr(X)}{i} P(A)^i P(\neg A)^{fr(X)-i}$$

Corresponding z-score:

$$z_2 = \frac{fr(XA) - \mu}{\sigma} = \frac{fr(XA) - fr(X)P(A)}{\sqrt{fr(X)P(A)P(\neg A)}}$$

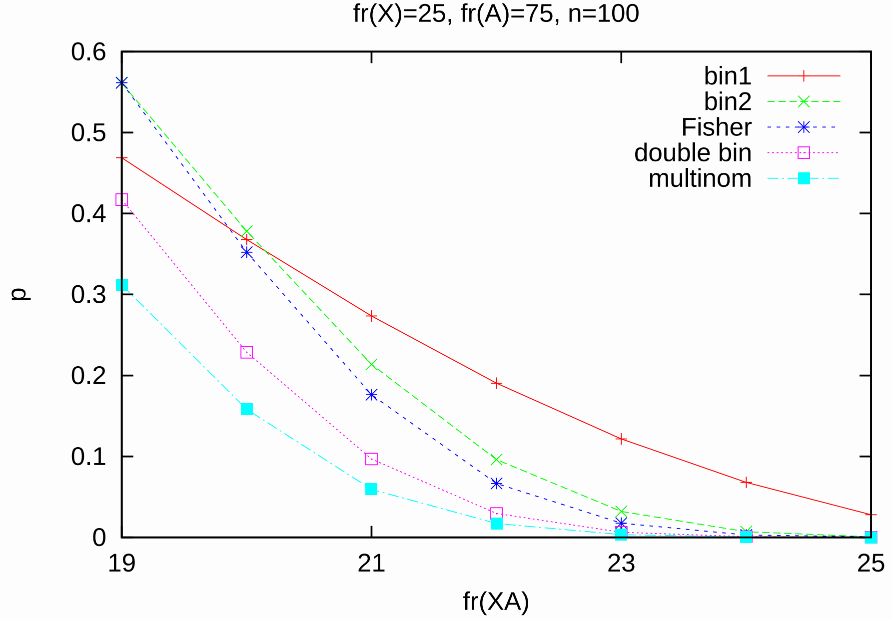
$$= \frac{\sqrt{n}\delta(X, A)}{\sqrt{P(X)P(A)P(\neg A)}} = \frac{\sqrt{fr(X)}(P(A|X) - P(A))}{\sqrt{P(A)P(\neg A)}}$$

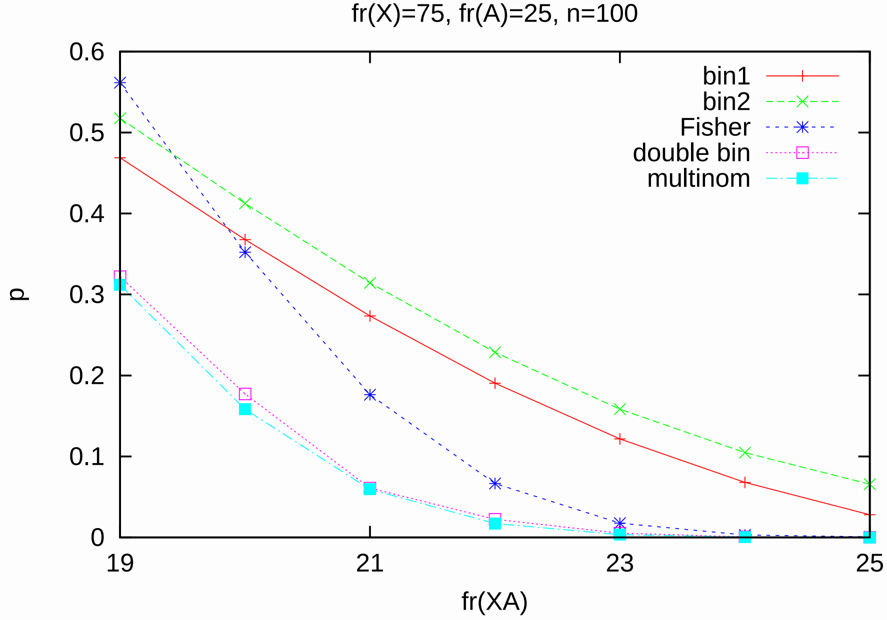
J-measure

$\approx$ one urn version of MI

$$J = P(XA)\log\frac{P(XA)}{P(X)P(A)} + P(X\neg A)\log\frac{P(X\neg A)}{P(X)P(\neg A)}$$

Example: Comparison of p-values





Comparison: Sampling models for value-based dependencies

- Multinomial, Hypergeometric, classical Binomial + its z-score: $p(X \to A) = P(A \to X)$
- Double binomial, alternative Binomial + its z-score: $p(X \to A) \neq P(A \to X)$ (in general)
- The alternative Binomial, its z-score and J can disagree with the other measures (only the X-urn vs. whole data)
- z-score easy to integrate into search, but may be unreliable for infrequent patterns $\to$ (classical) Binomial test in post-pruning improves quality!

Part I Contents

3.3 Multiple testing problem

The more patterns we test, the more spurious patterns we are likely to accept.

- If threshold $\alpha = 0.05$, there is 5% probability that a spurious dependency passes the test.
- If we test 10 000 rules, we are likely to accept 500 spurious rules!

Solutions to Multiple testing problem

1. **Direct adjustment approach**: adjust α (stricter thresholds)
    - easiest to integrate into the search
2. **Holdout approach**: Save part of the data for testing $\rightarrow$ Webb
3. **Randomization test approaches**: Estimate the overall significance of all discoveries or adjust the individual p-values empirically
   $\rightarrow$ e.g. Gionis et al., Hanhijärvi et al.

Contingency table for m significance tests

	spurious rule H_0 true	genuine rule H_1 true	All
declared significant	V false positives	S true positives	R
declared insignificant	U true negatives	T false negatives	$m - R$
All	m_0	$m - m_0$	m

Direct adjustment: Two approaches

(i) Control **familywise error rate** = probablity of accepting at least one false discovery

$$FWER = P(V \geq 1)$$

(ii) Control **false discovery rate** = expected proportion of false discoveries

$$FDR = E\left[\frac{V}{R}\right]$$

	spurious rule	genuine rule	All
decl. sign.	V	S	R
decl. insign	U	T	$m - R$
All	m_0	$m - m_0$	m

(i) Control familywise error rate FWER

Decide $\alpha^* = FWER$ and calculate a new stricter threhold α.

- If tests are mutually independent: $\alpha^* = 1 - (1 - \alpha)^m$
  $\Rightarrow$ Šidák correction: $\alpha = 1 - (1 - \alpha^*)^{\frac{1}{m}}$
- If they are not independent: $\alpha^* \leq m \cdot \alpha$
  $\Rightarrow$ **Bonferroni correction**: $\alpha = \frac{\alpha^*}{m}$

- conservative (may lose genuine discoveries)
- How to estimate m?
  - may be explicit and implicit testing during search
- **Holm-Bonferroni** method more powerful
  - but less suitable for the search (all p-values should be known, first)

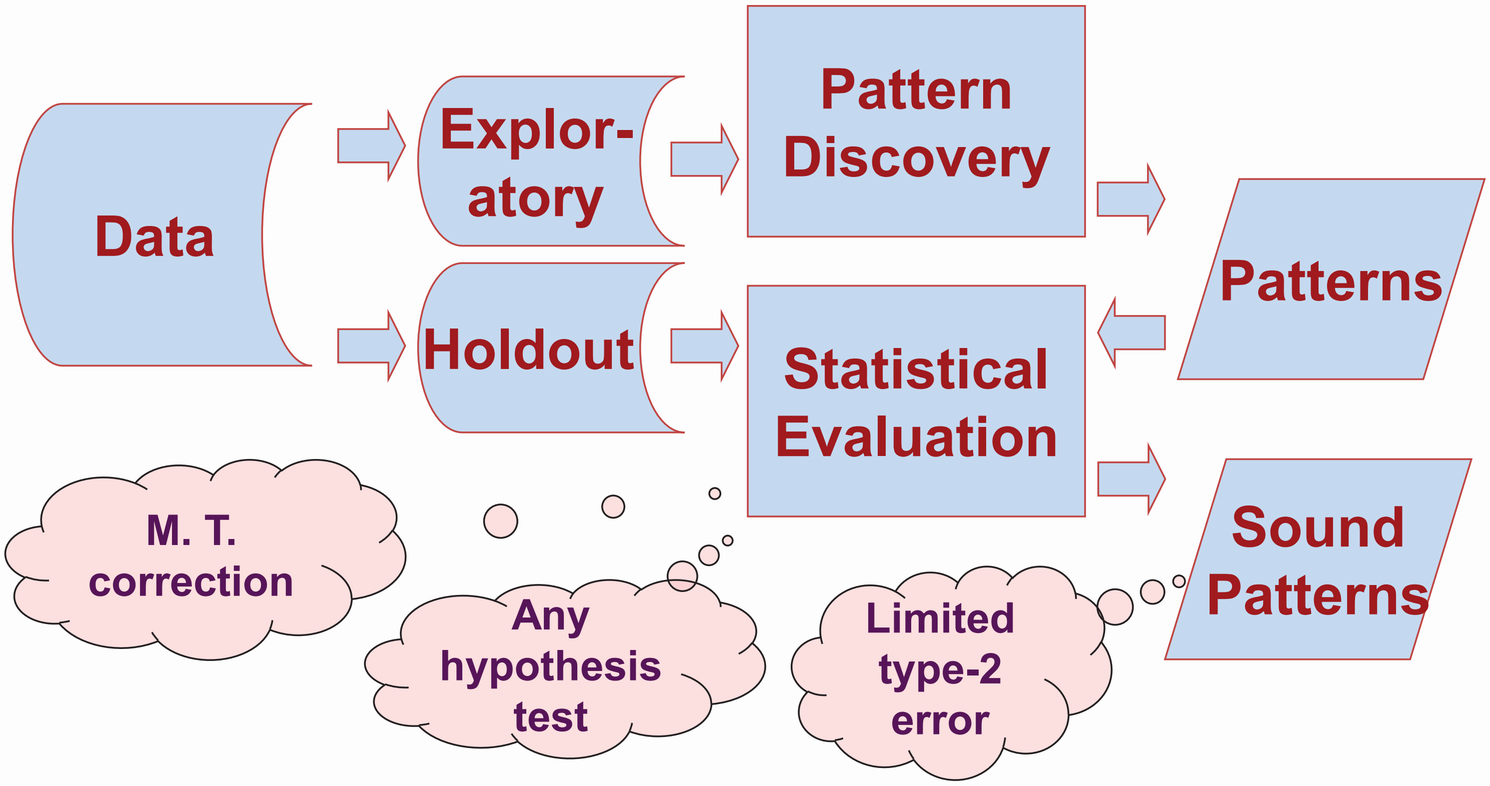
(ii) Control false discovery rate FDR

Benjamini–Hochberg–Yekutieli procedure

1. Decide $q = FDR$
2. Order patterns r_i by their p-values
   Result $r_1, \ldots, r_m$ such that $p_1 \leq \ldots \leq p_m$
3. Search the largest k such that $p_k \leq \frac{k \cdot q}{m \cdot c(m)}$
   - if tests mutually independent or positively dependent, $c(m) = 1$
   - otherwise $c(m) = \sum_{i=1}^{m} \frac{1}{i} \approx \ln(m) + 0.58$
4. Save patterns $r_1, \ldots, r_k$ (as significant) and reject $r_{k+1}, \ldots, r_m$

Hold-out approach

Powerful because m is quite small!



Randomization test approaches

1. Estimate the overall significance of discoveries at once
   - e.g. What is the probability to find K_0 dependency rules whose strength is at least min_M?
   - Empirical p-value

$$p_{emp} = \frac{|\{d_i \mid K_i \geq K_0\}| + 1}{b + 1}$$

d_0 original set
$d_1, \ldots, d_b$ random sets
$K_1, \ldots, K_b$ numbers of discovered patterns from set d_i

$\rightarrow$ Gionis et al.

Randomization test approaches (cont.)

2. Use randomization tests to correct individual p-values
   - e.g., How many sets contained better rules than $X \rightarrow A$?

$$p' = \frac{|\{d_i | (\mathcal{S}_i \neq \emptyset) \wedge (\min p(Y \rightarrow B\,|d_i) \leq p(X \rightarrow A\,|d_0)\}|}{b+1},$$

d_0 original set
$d_1, \ldots, d_b$ random sets
$\mathcal{S}_i$=set of patterns returned from set d_i

$\rightarrow$ Hanhijärvi

Randomization test approaches

- **+** dependencies between patterns not a problem $\rightarrow$ more powerful control over $FWER$
- **+** one can impose extra constraints (e.g. that a certain pattern holds with a given frequency and confidence)
- **–** most techniques assume *subset pivotality* $\approx$ the complete hypothesis and all subsets of true null hypotheses have the same distribution of the measure statistic

Remember also points mentioned in the single hypothesis testing

Part I Contents

1. Statistical dependency rules
2. Variable- and value-based interpretations
3. Statistical significance testing
   3.1 Approaches
   3.2 Sampling models
   3.3 Multiple testing problem
4. **Redundancy and significance of improvement**
5. Search strategies

4. Redundancy and significance of improvement

When $X \rightarrow A$ is redundant with respect to $Y \rightarrow A$ ($Y \subsetneq X$)? Improves it significantly?

Examples of redundant dependency rules:

- *smoking, coffee* → *atherosclerosis*
  coffee has no effect on *smoking* → *atherosclerosis*
- *high cholesterol, sports* → *atherosclerosis*
  sports makes the dependency only weaker
- *male, male pattern baldness* → *atherosclerosis*
  adding *male* hardly any significant improvement

Redundancy and significance of improvement

- Value-based: $X \rightarrow A$ is **productive** if $P(A|X) > P(A|Y)$ for all $Y \subsetneq X$
- Variable-based: $X \rightarrow A$ is **redundant** if there is $Y \subsetneq X$ such that $M(Y \rightarrow A)$ is better than $M(X \rightarrow A)$ with the **given goodness measure** M
  $\Leftrightarrow X \rightarrow A$ is **non-redundant** if for all $Y \subsetneq X$ $M(X \rightarrow A)$ is better than $M(Y \rightarrow A)$
- When the improvement is significant?

Value-based: Significance of productivity

Hypergeometric model:

$$p(YQ \rightarrow A|Y \rightarrow A) = \frac{\sum_i \binom{fr(YQ)}{fr(YQA)+i}\binom{fr(Y\neg Q)}{fr(Y\neg QA)-i}}{\binom{fr(Y)}{fr(YA)}}$$

$\approx$ probability of the observed or a stronger conditional dependency $Q \rightarrow A$, given Y, in a **value-based** model.

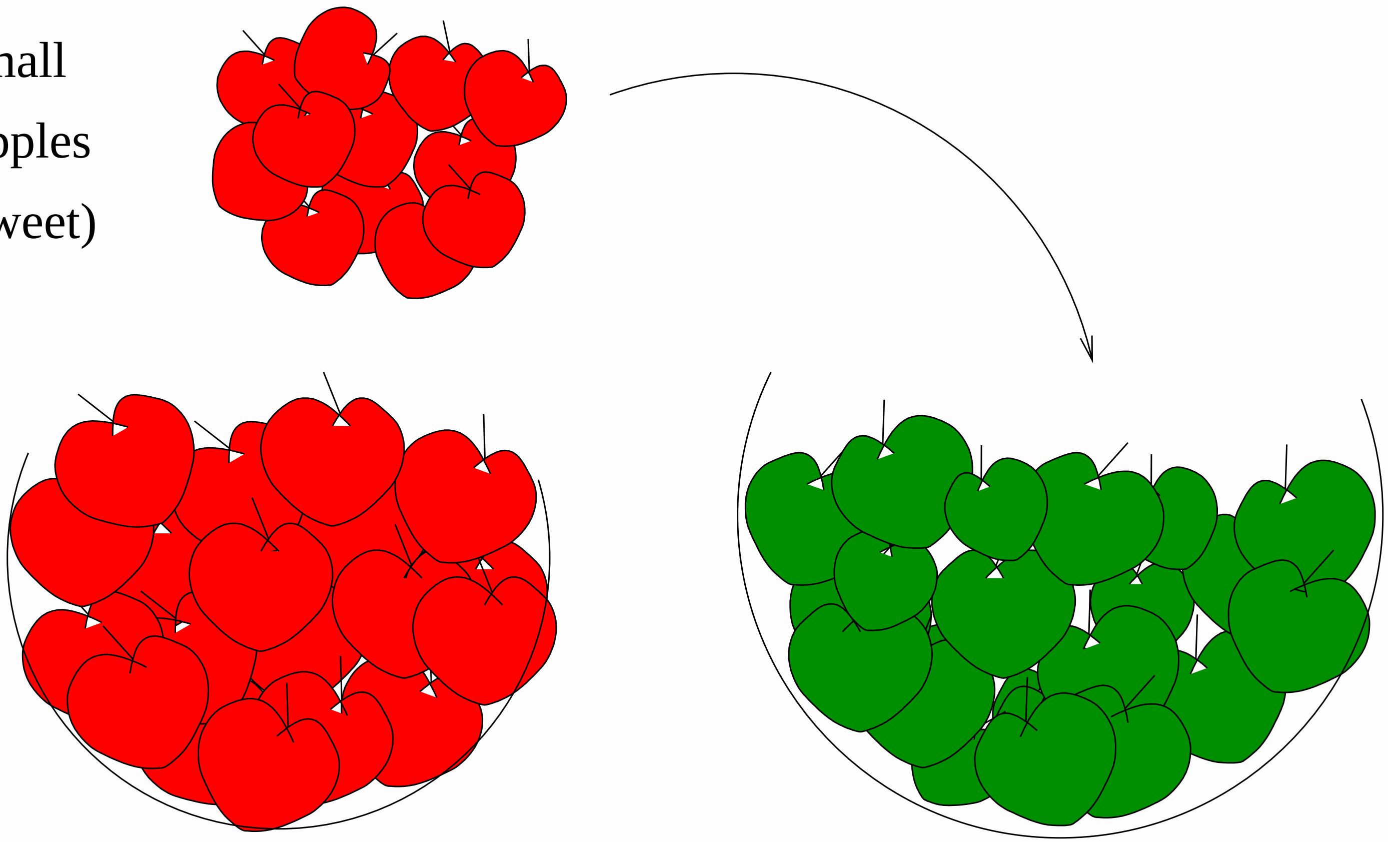
- also asymptotic measures (χ^2, MI)

Apple problem: value-based

$$p(YQ \to A | Y \to A) = 0.0029$$

Y=red, Q=large

20 small
red apples
(15 sweet)



Basket 1

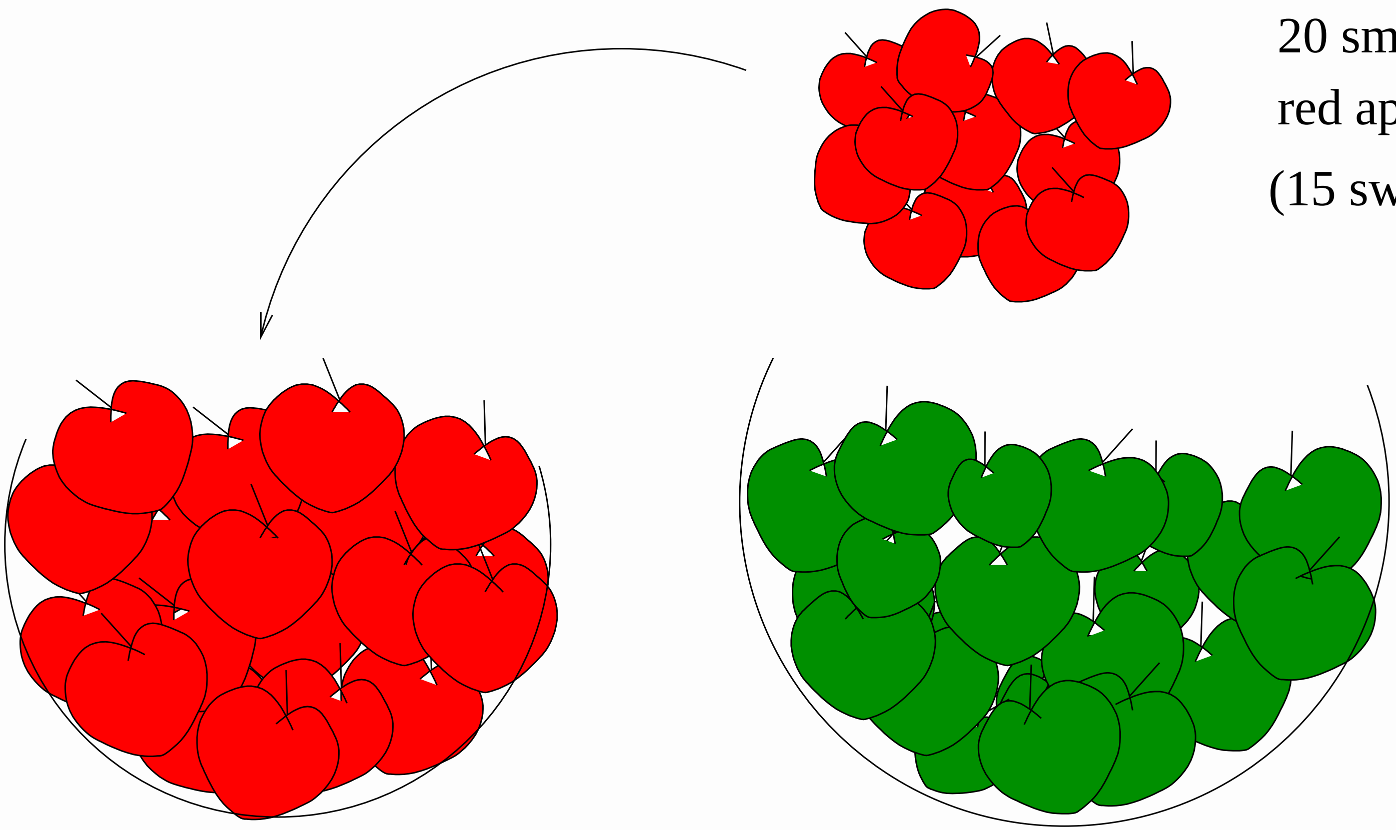
40 large red apples
(all sweet)

Basket 2

40 green apples
(all bitter)

Apple problem: variable-based?

$$p(\neg Y \rightarrow \neg A | \neg(YQ) \rightarrow \neg A) = 2.9e-10 << 0.0029$$



Observation

$$\frac{p(\neg Y \to \neg A | \neg(YQ) \to \neg A)}{p(YQ \to A | Y \to A)} \approx \frac{p_F(Y \to A)}{p_F(YQ \to A)}$$

Thesis: Comparing productivity of $YQ \to A$ and $\neg Y \to \neg A \equiv$ redundancy test with $M = p_F$!

Part I Contents

5. Search strategies

1. Search for the strongest rules (with γ, δ etc.) that pass the significance test for productivity
   $\rightarrow$ MagnumOpus (Webb 2005)
2. Search for the most significant non-redundant rules (with Fisher's p etc.)
   $\rightarrow$ Kingfisher (Hämäläinen 2012)
3. Search for frequent sets, construct association rules, prune with statistical measures, and filter non-redundant rules??
   - No way!
   - closed sets? $\rightarrow$ redundancy problem
   - their minimal generators?

Main problem: non-monotonicity of statistical dependence

- $AB \rightarrow C$ can express a significant dependency even if A and C as well as B and C mutually independent
- In the worst case, the only significant dependency involves all attributes $A_1 \ldots A_k$ (e.g. $A_1 \ldots A_{k-1} \rightarrow A_k$)

$\Rightarrow$ 1) A greedy heuristic does not work!

$\Rightarrow$ 2) Studying only simplest dependency rules does not reveal everything!

ABCA1-R219K $\rightarrow$ ¬alzheimer
ABCA1-R219K, female $\rightarrow$ alzheimer

End of Part I

Questions?