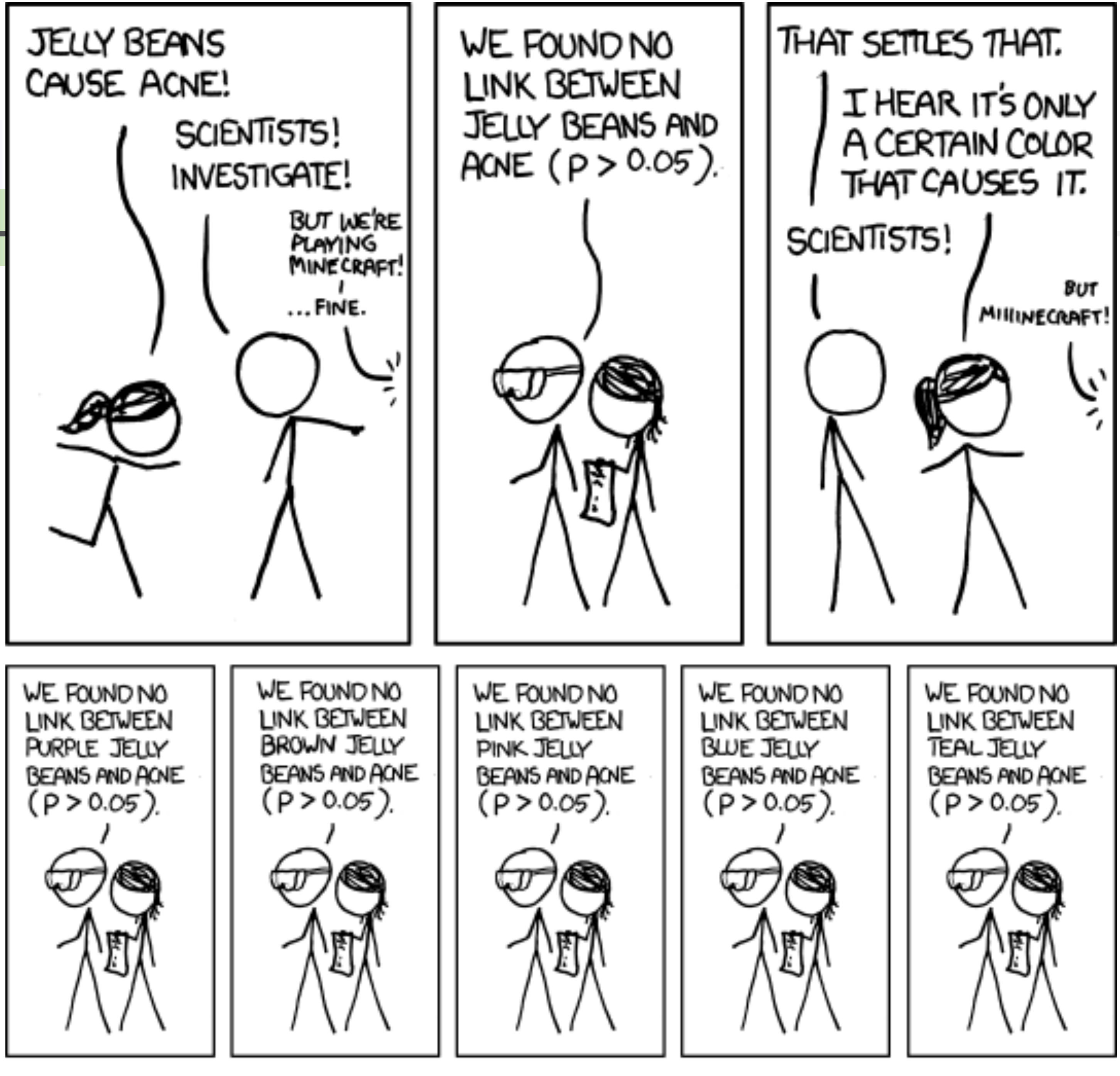
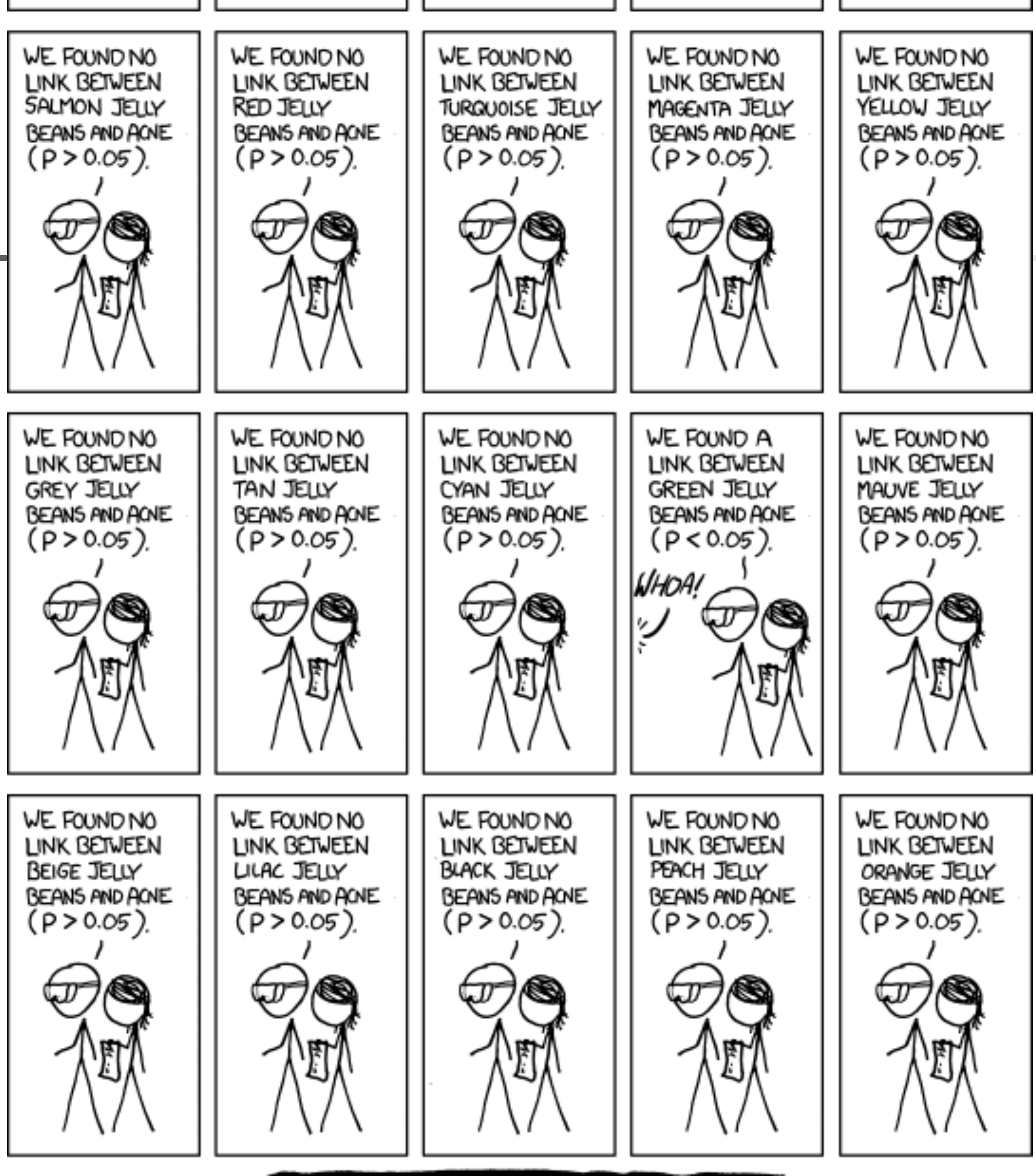
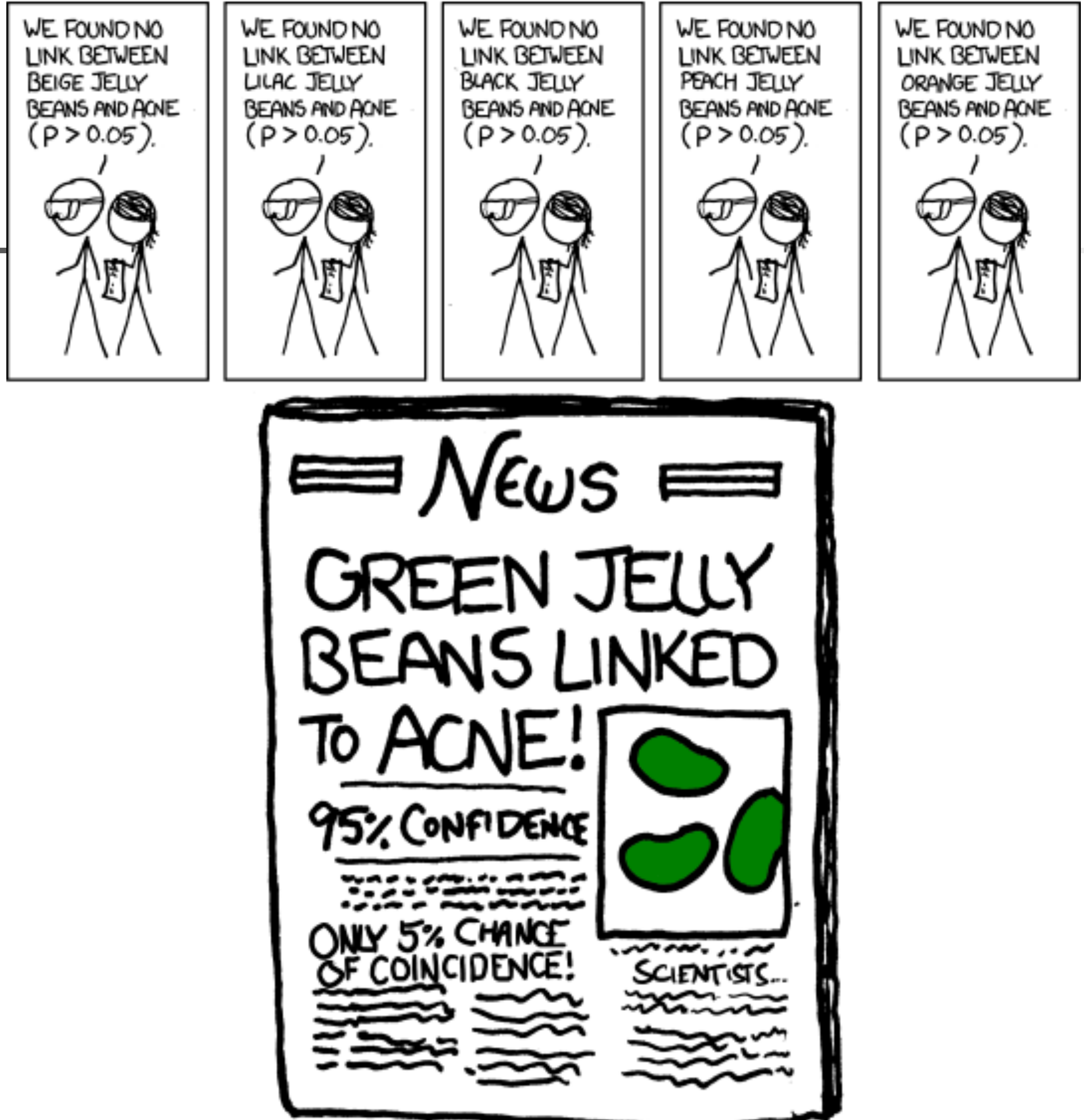
The Curse of Too Many Questions

Eli Upfal

BROWN


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Data Mining

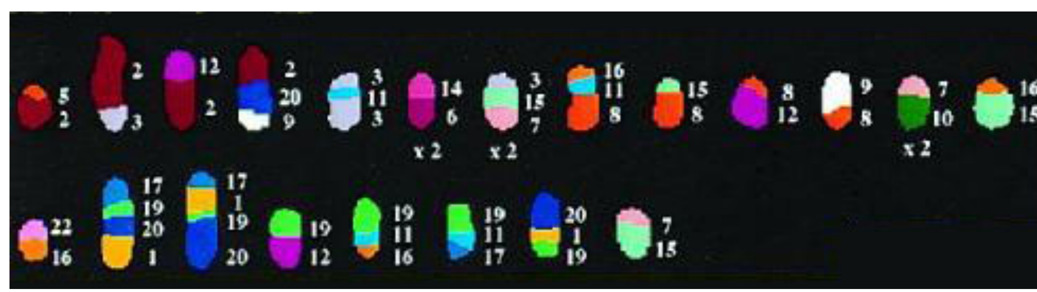
- Discover hidden patterns, correlations, association rules, etc., in large data sets
- When is the discovery interesting, important, significant?
- We develop rigorous mathematical/ statistical approach

Frequent Itemsets

- Dataset **D** of transactions $\mathbf{t_j}$ (subsets) of a base set of items **I**, ($\mathbf{t_j} \subseteq 2^{\mathbf{I}}$).
- Support of an itemsets **X** = number of transactions that contain **X**.
- **I =** set of mutations
- **T_j** = the set of mutations found in patient **J**

Frequent Itemsets

- Discover all itemsets with significant support.
- Fundamental primitive in data mining, Data Bases (association rules), network security, computational biology, …



Significance

- What support level makes an itemset significantly frequent?
  - Minimize false positive and false negative discoveries
  - Improve "quality" of subsequent analyses
- How to narrow the search to focus only on significant itemsets?
  - Reduce the possibly exponential time search

Statistical Model

- Input:
  - **D** = a dataset of **t** transactions over **|I|=n**
  - For $\mathbf{i} \in \mathbf{I}$, let **n(i)** be the support of **{i}** in **D**.
  - $\mathbf{f_i} = \mathbf{n(i)/t}$ = frequency of **i** in **D**
- $\mathbf{H_0}$ Model:
  - **D** = a dataset of **t** transactions, **|I|=n**
  - Item **i** is included in transaction **j** with probability $\mathbf{f_i}$ independent of all other events.

Statistical Tests

- $\mathbf{H_0}$: null hypothesis – the support of no itemset is significant with respect to **D**
- $\mathbf{H_1}$: alternative hypothesis, the support of itemset $\{\mathbf{X_1, X_2, \ldots, X_r}\}$ is significant. It is unlikely that this support comes from the distribution of **D**
- Significance level:

  α = **Prob(** rejecting $\mathbf{H_0}$ when it's true **)**

Naïve Approach

- Let $\mathbf{X=\{x_1,x_2,\ldots x_r\}}$,
- $\mathbf{f_x = \Pi_j\, f_j}$, probability that a given itemset is in a given transaction
- $\mathbf{s_x}$ = support of $\mathbf{X}$, distributed $\mathbf{s_x \sim B(t, f_x)}$

- Reject $\mathbf{H_0}$ if:
  $\mathbf{Prob(B(t, f_x) \geq s_x) = p\text{-value} \leq \alpha}$

Naïve Approach

- Variations:
  - **R**=support **/ E[**support in **D]**
  - **R**=support **- E[**support in **D]**
  - **Z**-value = (s-**E[**s**]**)/**σ[**s**]**
  - many more…

Measure (Symbol)	Definition
Correlation (ϕ)	$\frac{N f_{11} - f_{1+} f_{+1}}{\sqrt{f_{1+} f_{+1} f_{0+} f_{+0}}}$
Odds ratio (α)	$(f_{11} f_{00}) / (f_{10} f_{01})$
Kappa (κ)	$\frac{N f_{11} + N f_{00} - f_{1+} f_{+1} - f_{0+} f_{+0}}{N^2 - f_{1+} f_{+1} - f_{0+} f_{+0}}$
Interest (I)	$(N f_{11}) / (f_{1+} f_{+1})$
Cosine (IS)	$(f_{11}) / (\sqrt{f_{1+} f_{+1}})$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N} - \frac{f_{1+} f_{+1}}{N^2}$
Collective strength (S)	$\frac{f_{11} + f_{00}}{f_{1+} f_{+1} + f_{0+} f_{+0}} \times \frac{N - f_{1+} f_{+1} - f_{0+} f_{+0}}{N - f_{11} - f_{00}}$
Jaccard (ζ)	$f_{11} / (f_{1+} + f_{+1} - f_{11})$
All-confidence (h)	$\min\left[\frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}}\right]$
Goodman-Kruskal (λ)	$\left[\frac{\sum_j \max_k f_{jk} + \sum_k \max_j f_{jk} - \max_j f_{j+} - max_k f_{+k}}{2N - \max_j f_{j+} - \max_k f_{+k}}\right]$
Mutual Information (M)	$\frac{\sum_i \sum_j \frac{f_{ij}}{N} \log \frac{N f_{ij}}{f_{i+} f_{+j}}}{\min\left[-\sum_i \frac{f_{i+}}{N} \log \frac{f_{i+}}{N}, -\sum_j \frac{f_{+j}}{N} \log \frac{f_{+j}}{N}\right]}$
J-Measure (J)	$\frac{f_{11}}{N} \log \frac{N f_{11}}{f_{1+} f_{+1}} + \max\left[\frac{f_{10}}{N} \log \frac{N f_{10}}{f_{1+} f_{+0}}, \frac{f_{01}}{N} \log \frac{N f_{01}}{f_{0+} f_{+1}}\right]$
Gini index (G)	$\max\Big[\frac{f_{1+}}{N} \times [(\frac{f_{11}}{f_{1+}})^2 + (\frac{f_{10}}{f_{1+}})^2] + \frac{f_{0+}}{N} \times [(\frac{f_{01}}{f_{0+}})^2 + (\frac{f_{00}}{f_{0+}})^2] - (\frac{f_{+1}}{N})^2 - (\frac{f_{+0}}{N})^2,$ $\frac{f_{+1}}{N} \times [(\frac{f_{11}}{f_{+1}})^2 + (\frac{f_{01}}{f_{+1}})^2] + \frac{f_{+0}}{N} \times [(\frac{f_{10}}{f_{+0}})^2 + (\frac{f_{00}}{f_{+0}})^2] - (\frac{f_{1+}}{N})^2 - (\frac{f_{0+}}{N})^2\Big]$
Laplace (L)	$\max\left[\frac{f_{11}+1}{f_{1+}+2}, \frac{f_{11}+1}{f_{+1}+2}\right]$
Conviction (V)	$\max\left[\frac{f_{1+} f_{+0}}{N f_{10}}, \frac{f_{0+} f_{+1}}{N f_{01}}\right]$
Certainty factor (F)	$\max\left[\frac{\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}}{1 - \frac{f_{+1}}{N}}, \frac{\frac{f_{11}}{f_{+1}} - \frac{f_{1+}}{N}}{1 - \frac{f_{1+}}{N}}\right]$
Added Value (AV)	$\max\left[\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}, \frac{f_{11}}{f_{+1}} - \frac{f_{1+}}{N}\right]$

What' s wrong? – example

- **D** has 1,000,000 transactions, over 1000 items, each item has frequency 1/1000.
- We observed that a pair **{i,j}** appears 7 times, is this pair statistically significant?
- In **D** (random dataset):
  - E[support(**{i,j}**)] = 1
  - Prob(**{i,j}** has support ≥ 7) ≃ 0.0001
- p-value 0.0001 - must be significant!

What's wrong? – example

- There are 499,500 pairs, each has probability 0.0001 to appear in 7 transactions in **D**
- The expected number of pairs with support ≥ 7 in D is ≃ 50,

  not such a rare event!
- Many false positive discoveries (flagging itemsets that are not significant)
- **Need to correct for multiplicity of hypothesis.**

Multi-Hypothesis test

- Testing for significant itemsets of size **k** involves testing simultaneously for **m**=$\binom{n}{k}$ null hypothesis.
- $\mathbf{H_0(X)}$ = support of **X** conforms with **D**

  $\mathbf{s_x}$ **=** support of **X**, distributed: $\mathbf{s_x \sim B(t, f_x)}$
- How to combine **m** tests while minimizing false positive and negative discoveries?

The Statistics Approach

Correct but conservative: prefers false negative to false positive results.

Conservative - There is often nothing to report – no statistically significant discoveries

Family Wise Error Rate (FWER)

- Family Wise Error Rate (**FWER**) = probability of at least one false positive
  (flagging a non-significant itemset as significant)
- Bonferroni method (union bound) – test each null hypothesis with significance level α/m
- Too conservative – many false negative – does not flag many significant itemsets.

False Discovery Rate (FDR)

- Less conservative approach
- V= number of false positive discoveries
- R= total number of rejected null hypothesis
  = number itemsets flagged as significant

$\boxed{\textbf{FDR} = \textbf{E[V/R]} \qquad (\textbf{FDR}=\textbf{0} \text{ when } \textbf{R}=\textbf{0})}$

- Test with level of significance α : reject maximum number of null hypothesis such that **FDR ≤ α**

Standard Multi-Hypothesis test

Theorem (Benjamini and Yekutieli,'01). *Assume that we are testing for m null hypotheses.*
Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered observed p-values of the m tests. To control of FDR at level β, define

$$\ell = \max \left\{ i \geq 0 : p_{(i)} \leq \frac{i}{m \sum_{j=1}^{m} \frac{1}{j}} \beta \right\},$$

and reject the null hypotheses of tests $(1), \ldots, (\ell)$.

Standard Multi-Hypothesis test

- Less conservative than Bonferroni method:
  - $i\alpha/m$ **vs** α/m

- For **m**= $\binom{n}{k}$, still needs a very small individual p-value to reject an hypothesis

Alternative Approach

- $\mathbf{Q(k, s_i)}$ = observed number of itemsets of size $\mathbf{k}$ and support $\geq \mathbf{s_i}$
- **p-value** =
  the probability of $\mathbf{Q(k, s_i)}$ in **D**
- Fewer hypothesis
- How to compute the p-value? What is the distribution of the number of itemsets of size $\mathbf{k}$ and support $\geq \mathbf{s_i}$ in **D** ?

[JACM 2012 - Kirsch, Mitzenmacher, Pietracaprina, Pucci, U, Vandin]

Alternative Statistical Test

- Instead of testing the significance of the support of individual itemsets we test the significance of the number of itemsets with a given support
- The null hypothesis distribution is specified by the Poisson approximation result
- Reduces the number of simultaneous tests
- More powerful test – less false negatives

Test I

- Define $\alpha_1, \alpha_2, \alpha_3, \ldots$ such that $\Sigma\alpha_i \leq \alpha$
- For $\mathbf{i=0,\ldots,\log(s_{max} - s_{min}) + 1}$
  - $\mathbf{s_i = s_{min} + 2^i}$
  - $\mathbf{Q(k, s_i)}$ = observed number of itemsets of size **k** and support $\geq \mathbf{s_i}$
  - $\mathbf{H_0(k,s_i)}$ = "$\mathbf{Q(k,s_i)}$ conforms with $\mathbf{Poisson(\lambda_i)}$"

- Reject $\mathbf{H_0(k,s_i)}$ if **p-value** $< \alpha_i$

Test I

- Let **s*** be the smallest **s** such that $\mathbf{H_0 (k,s)}$ rejected by Test I
- With confidence level α the number of itemsets with support ≥ **s*** is significant
- Some itemsets with support ≥ **s*** could still be false positive

Test II

- Define $\beta_1, \beta_2, \beta_3, \ldots$ such that $\sum \beta_i \leq \beta$

- Reject $\mathbf{H_0(k,s_i)}$ if:

  p-value < α_i and $\mathbf{Q(k,s_i)} \geq \lambda_i / \beta_i$

- Let **s*** be the minimum **s** such that $\mathbf{H_0(k,s)}$ was rejected
- If we flag all itemsets with support **≥ s*** as significant, **FDR ≤** β

Proof

- $\mathbf{V_i}$ = false discoveries if $\mathbf{H_0(k,s_i)}$ first rejected
- $\mathbf{E_i}$ = "$\mathbf{H_0(k,s_i)}$ rejected"

$$
\begin{aligned}
FDR &= \sum_{i=0}^{h-1} E\left[\frac{V_i}{Q_{k,s_i}}\right] \mathbf{Pr}(E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0) \\
&\leq \sum_{i=0}^{h-1} \frac{E[X_i \mid E_i \bar{E}_{i-1}, \ldots, \bar{E}_0]}{\lambda_i/\beta_i} \mathbf{Pr}(E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0) \\
&= \sum_{i=0}^{h-1} \frac{\sum_j j \mathbf{Pr}(X_i = j, E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0)}{\lambda_i/\beta_i} \\
&\leq \sum_{i=0}^{h-1} \frac{\beta_i \lambda_i}{\lambda_i} \leq \sum_{i=0}^{h-1} \beta_i \leq \beta.
\end{aligned}
$$

$\square$

The Theoretical CS Approach

- The Vapnik / PAC Learning approach
- Uniform Convergence Samples

Uniform Convergence

- Let **C** be a collection of hypotheses (concepts).
- We want a minimum sample (training set) that includes, for each wrong concept, at least one example demonstrating that this concept is wrong.
- At least for concepts that are "significantly wrong".

Uniform Convergence

- Classification problems on a set of items **I**
- A concept is a subset of items classified **True**
- Training examples are generated by a distribution **D**
- Algorithm is measures on the same distribution **D**

Uniform Convergence

A concept class (model) is **(m, ε , δ)-PAC-learnable** iff there is an algorithm that for **any** distribution **D**

- given m random inputs from for **D**
- with probability **1-** δ , outputs a concept
- concept is **correct** with probability **1-** ε on examples drawn randomly from **D.**

Uniform Convergence

- A concept class with **VC-dimension d** is (ε, δ)-PAC-learnable with
- $\mathbf{m = \Theta((d + \log 1/\delta)/\varepsilon)}$ samples

A sample of that size is an ε **– net** -

a sample that hits any set of size (measure) $\geq \varepsilon$

Vapnik-Chervonenkis Dimension

- Combinatorial property of a collection of subsets from a domain
- Measures the "richness", "expressivity" of the subsets
- A *Range set* is a pair (X,R)
  - X – set of items
  - R – collection of subset of X
- The VC-dimension of (X,R) is the maximal set size d such that all its 2^d partitions are obtained by intersections with sets in R
- The **sample "converge uniformly"** on all concepts in the class.

ε - Sampler

- estimating the sizes of all subsets
- Given a collection of sets (a range space), an ε – Sampler is a subset of elements that, with probability 1- δ, gives an ε – estimate of the sizes of all sets.

- If the VC-dimension of the collection of sets is d, then a random sample of size f(d, ε, δ) is an ε-sampler.

Are VC-Dimension Bounds Tight?

- VC – dimension is a combinatorial bound that "ignores" the data distribution
- Often hard to compute
- Rademacher Complexity….

The Practical (AI) Approach

- Cross Validation – compare results on subsets of the sample.
- If subsets are not disjoint estimates the variance **in** the sample
- Not a good predictor for “generalization” error.