SQL query

Arend Hintze

you saw the SELECT command before…

```
SELECT * FROM employees

SELECT * FROM employees WHERE name="Amy Wong"

SELECT * FROM employees WHERE name="Amy Wong" OR salary<2.0

SELECT * FROM employees WHERE name="Amy Wong" OR salary<2.0
        ORDER BY salary

SELECT * FROM employees WHERE name="Amy Wong"
    UNION
        SELECT * FROM employees WHERE salary<2.0

SELECT * FROM employees WHERE name="Amy Wong" AND salary<2.0
```

when lists become long
LIMIT, ORDER

```
SELECT * FROM employees LIMIT 4
SELECT * FROM employees ORDER BY salary LIMIT 4

SELECT * FROM employees LIMIT 4 ORDER BY salary
```

will not work!

how much money do we spent in total?

```
SELECT SUM(salary) FROM employees
```

on office nr 1?

```
SELECT SUM(salary) FROM employees  WHERE
roomNumber=1
```

per office?

```
SELECT SUM(salary) FROM employees GROUP BY
roomNumber
```

and how many employees do we have?

```
SELECT COUNT(salary) FROM employees
SELECT COUNT(*) FROM employees
```

who earns most, least?

```
SELECT MAX(salary) FROM employees
SELECT MIN(salary) FROM employees
```

and on average?

```
SELECT AVG(salary) FROM employees
```

Distances

Arend Hintze

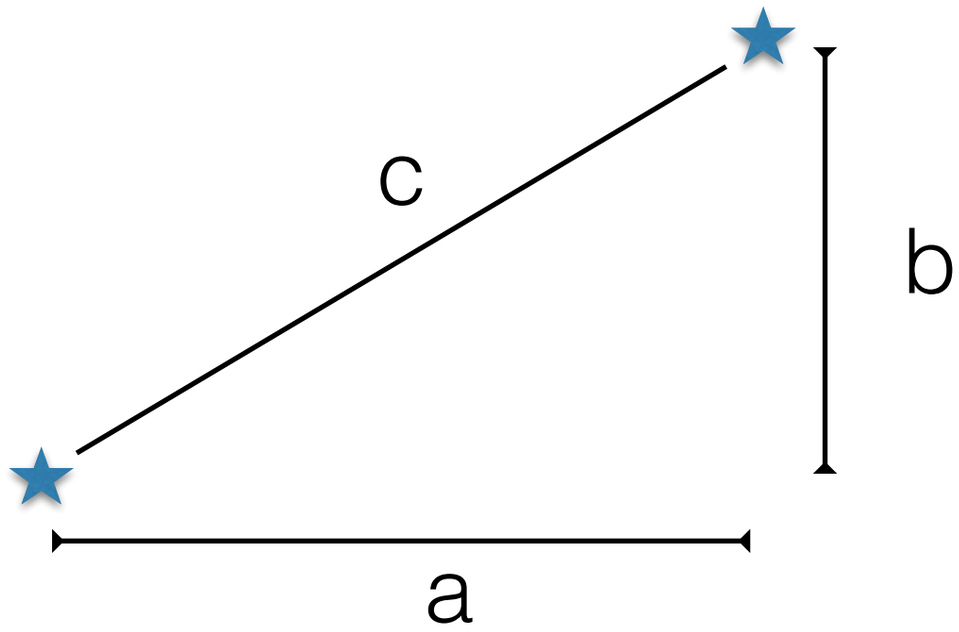
similarity dissimilarity

- similarity
  - numerical measure how alike two data objects are
  - the higher the more alike
  - [0,1]
- dissimilarity/distance
  - numerical measure how different two data objects are
  - the lower the more alike
  - minimum is 0, upper limit ?
- proximity refers to both

measures

- Dissimilarity
  - Euclidean
  - Minkowski
  - Mahalanobis
- Similarity
  - Binary data (SMC, Jaccard, cosine, Hamming)
  - continuous data (Tanimoto, correlation)

Euclidean



c=sqrt(a^2+b^2)

multiple dimensions?

p = (X,Y,Z,…)
q = (X,Y,Z,…)

$$dist(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

sqrt((Xq-Xp)^2+(Yq-Yp)^2+(Zq-Zp)^2+…)

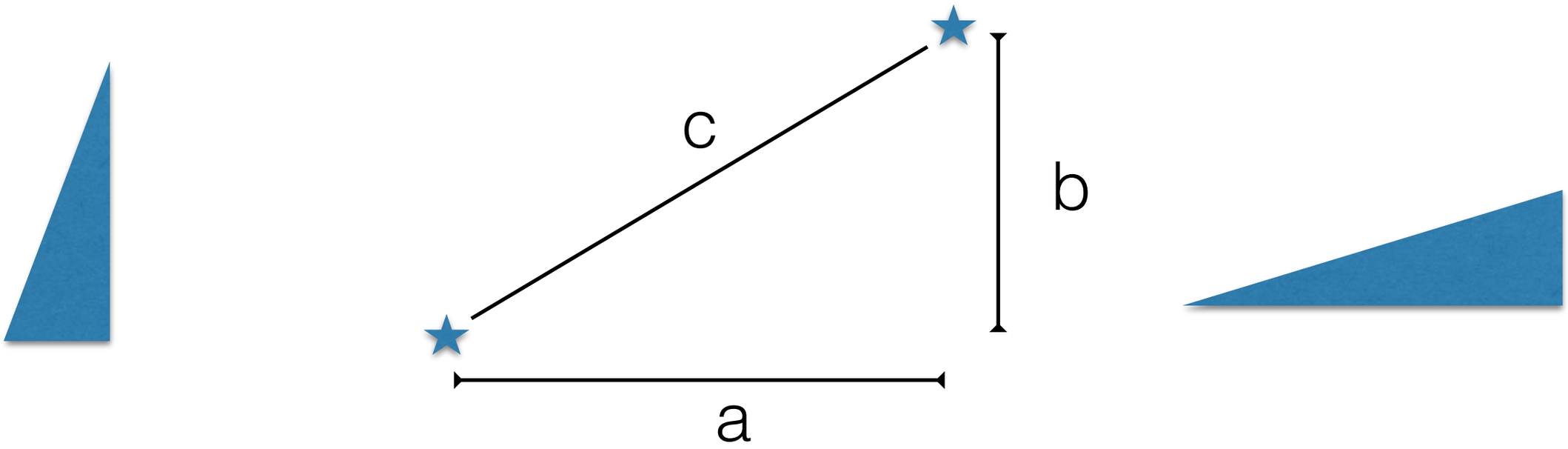
Minkowski Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

pow((|Xq-Xp|)^r+(|Yq-Yp|)^r+(|Zq-Zp|)^r+…,1/r)

- r=1 City block
- r=2 Euclidean
- r -> inf

what if data needs normalization or rescaling?



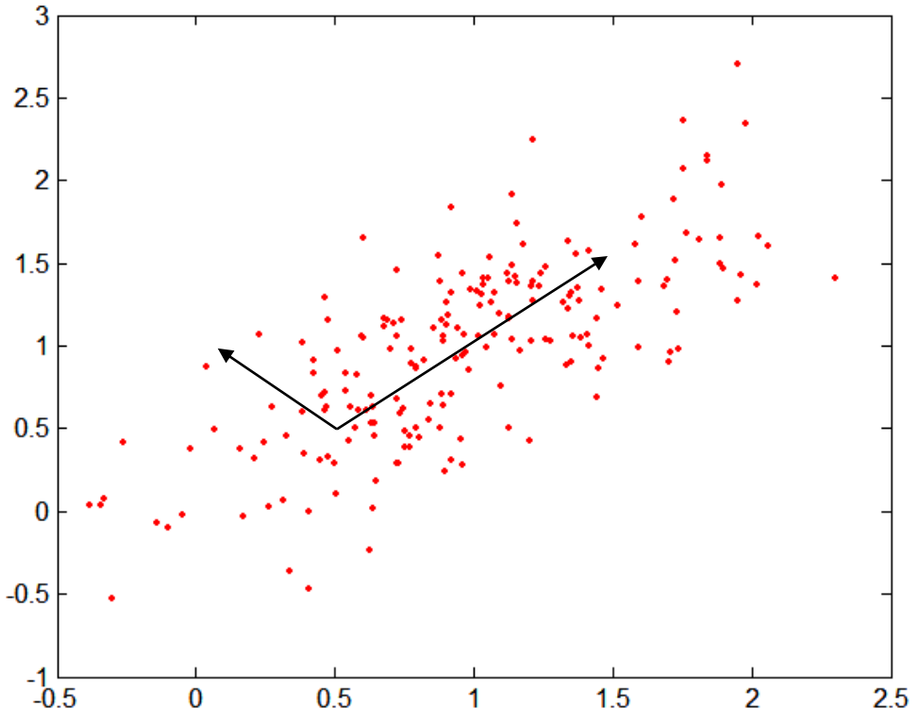
other issues binary data:

A	1	1	1	1	1	1	1	1	1
B	0	0	1	1	1	1	1	0	0

which pair is more similar?

A	1	1	0	0	0	0	0	1	1
B	0	0	0	0	0	0	0	0	0

Mahalanobis Distance



X →

A	B
1	1
2	2
1	2
3	4
…	…
5	5

Y →

sqrt(
(Ax-Ay)^2/std(A) +
(Bx-By)^2/std(B)
)

std(A)
std(B)

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}},$$

still, what to do with this:

A	1	1	1	1	1	1	1	1	1
B	0	0	1	1	1	1	1	0	0

M00 = nr where A=0 and B=0
M01 = nr where A=0 and B=1
M10 = nr where A=1 and B=0
M11 = nr where A=1 and B=1

simple matching coefficient:

SMC = (M11+M00)/(M00+M01+M10+M11)

Jaccard Coefficient:

J = M11/(M01+M10+M11)

Tanimoto

$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

where p•q is the dot product
and |p| is the magnitude (length) of vector p

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \qquad \|\mathbf{x}\| := \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

if used on binary data -> Jaccard coefficient

cosine similarity

imagine two multidimensional data points

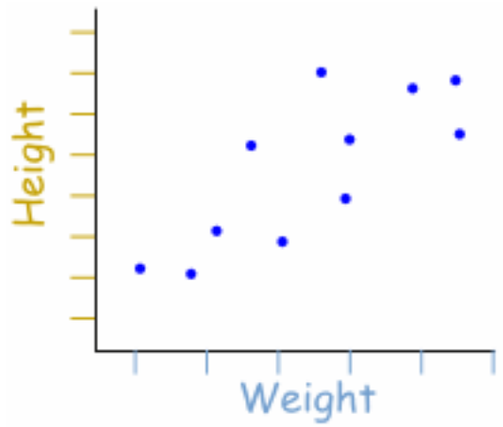
now imagine a vector from the origin to each point

the angle between both vectors is the cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

dot product divided by both their lengths multiplied

correlation



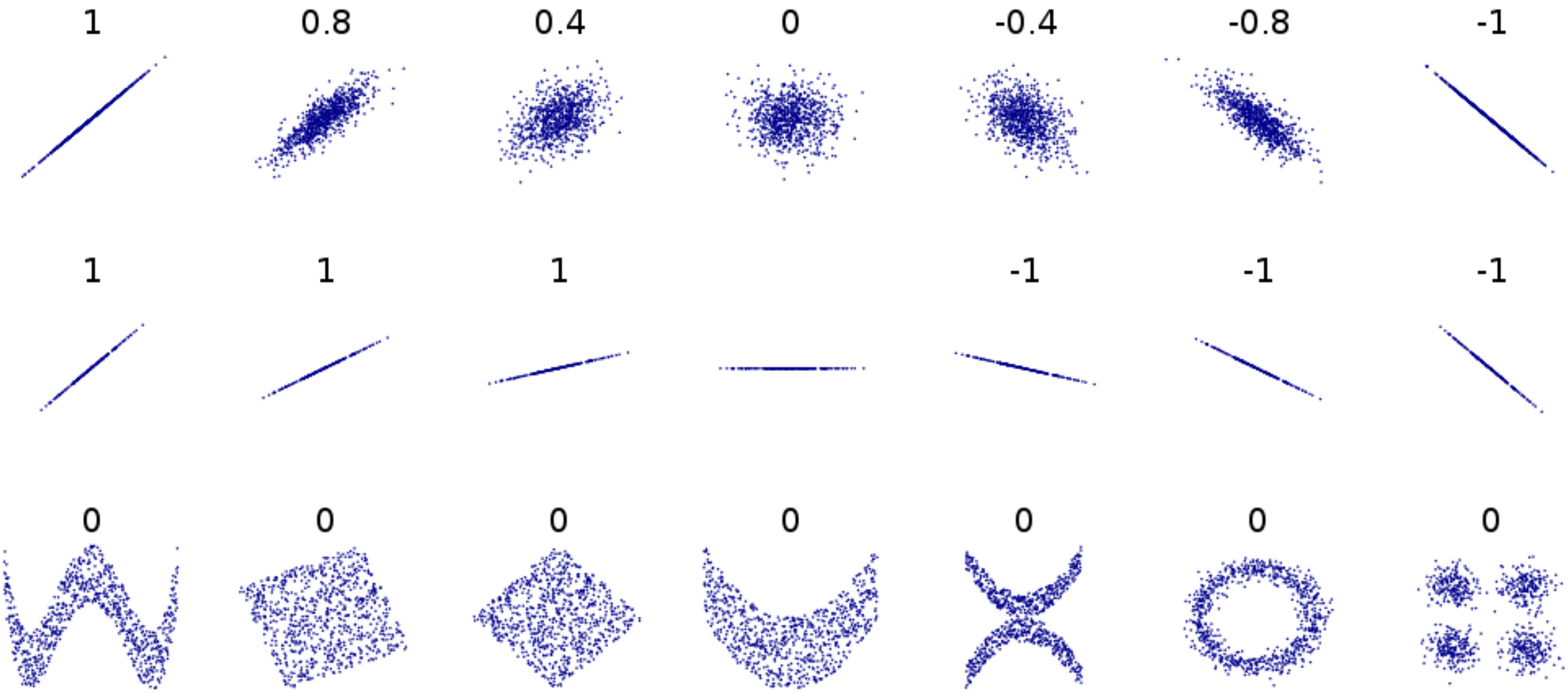
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

Ā=mean(A)

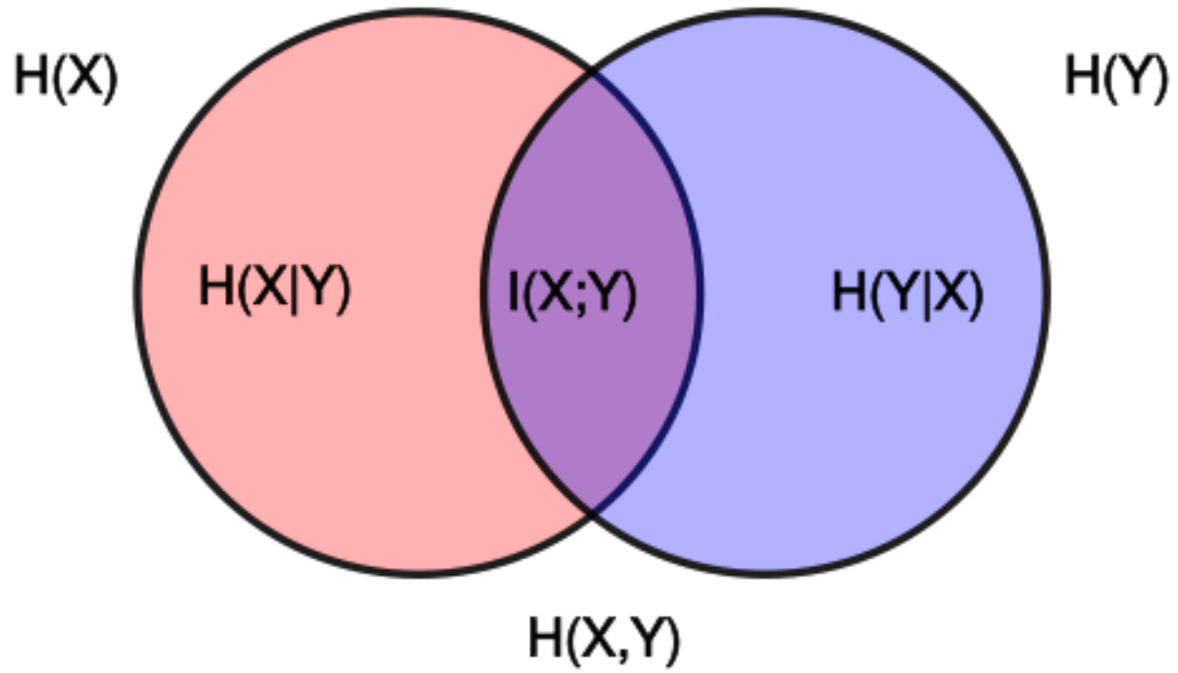
s=σ=standard deviation

$$\sigma = \sqrt{\frac{1}{N}\left[(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2\right]}, \quad \text{where} \quad \mu = \frac{1}{N}(x_1 + \cdots + x_N),$$

well, so much for that…



mutual information



$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$