How Many Ways Can Things Be The Same?

Set Theory For Multiple Site Surveys.

Richard Clegg (richard@manor.york.ac.uk)

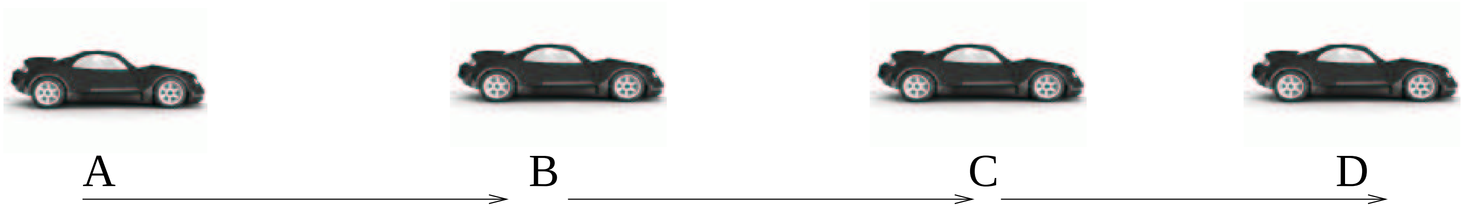
Networks and Nonlinear Dynamics Group,
Department of Mathematics,
University of York

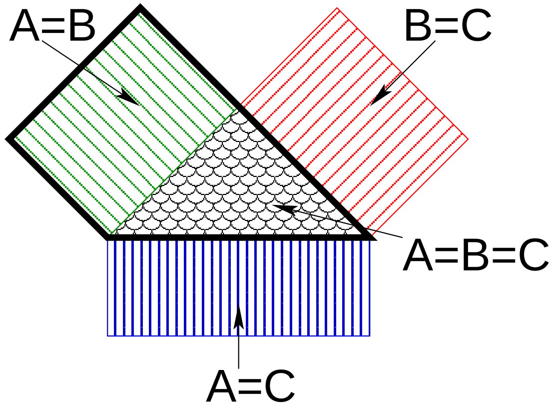
Slides prepared using the Prosper package and LaTeX

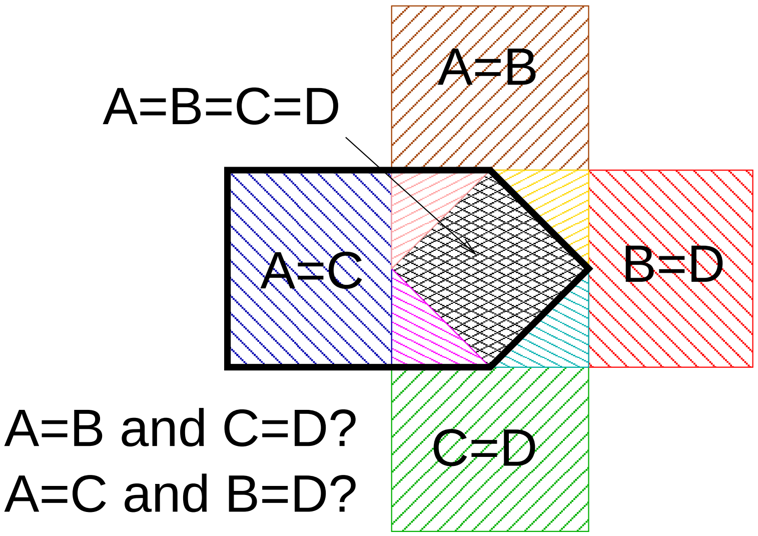
Summary of Talk

- This talk is about a general framework for multiple site surveys in any context.
- This talk is about car number plates.
- This talk is about set theory.
- This talk is about a generalisation of the game of snap.
- This talk comes with a special offer.

Visualising the Problem







It seems that there are different types of match.

Problem Statement

1. Formalise the notion of a type of match.
2. Enumerate the types of match.
3. Formalise the concept of a false match.
4. Create an algorithm for removing false matches from real data.
5. Test this algorithm on simulated data and real data.

An n-Dimensional Game Of Snap

Type of match formalised with equivalence classes. An n-point observation represented as n-tuple: $\mathbf{x} = (x_1, \ldots x_n)$. Two n-tuples $\mathbf{x}$ and $\mathbf{y}$ are equivalent ($\mathbf{x} \sim \mathbf{y}$) iff:

$$(x_i = x_j) \iff (y_i = y_j)$$

$$(1, 4, 7, 1) \sim (0, 10, 7, 0)$$

$$(\heartsuit, \heartsuit, \spadesuit, \heartsuit, \spadesuit) \sim (\diamondsuit, \diamondsuit, \spadesuit, \diamondsuit, \spadesuit)$$

$$(\mu, \mu, \pi, \phi) \not\sim (\mu, \pi, \phi, \phi)$$

$$(elephant, rhino, hippo, elephant) \sim (\bullet, \bullet, \bullet, \bullet)$$

$$(\texttt{A154FDE}, \texttt{A154FDE}, \texttt{B232DSR}) \not\sim (\texttt{A154FDE}, \texttt{A154FDE}, \texttt{A154FDE})$$

The Set $\mathcal{M}_n$ of All Types of Match

An n-tuple $\mathbf{x} \in \mathcal{M}_n$ iff $x_i \in \mathbb{N}$ and:

$$x_i = \begin{cases} 1 & i = 1 \\ x_j \text{ for some } j < i & i > 1 \quad \text{or} \\ 1 + \max_{j<i}(x_j) & i > 1 \end{cases}$$

$$(1, 4, 7, 1) \sim (1, 2, 3, 1)$$
$$(\heartsuit, \heartsuit, \spadesuit, \heartsuit, \spadesuit) \sim (1, 1, 2, 1, 2)$$

The set $\mathcal{M}_n$ is a transversal of all n-tuples under the relation defined by $\sim$.

Enumerating and Ordering $\mathcal{M}_n$

$\mathcal{M}_n$ can be set in one-to-one correspondance with the set $\mathcal{P}_n$ of partitions of $(1, 2, \ldots, n)$.

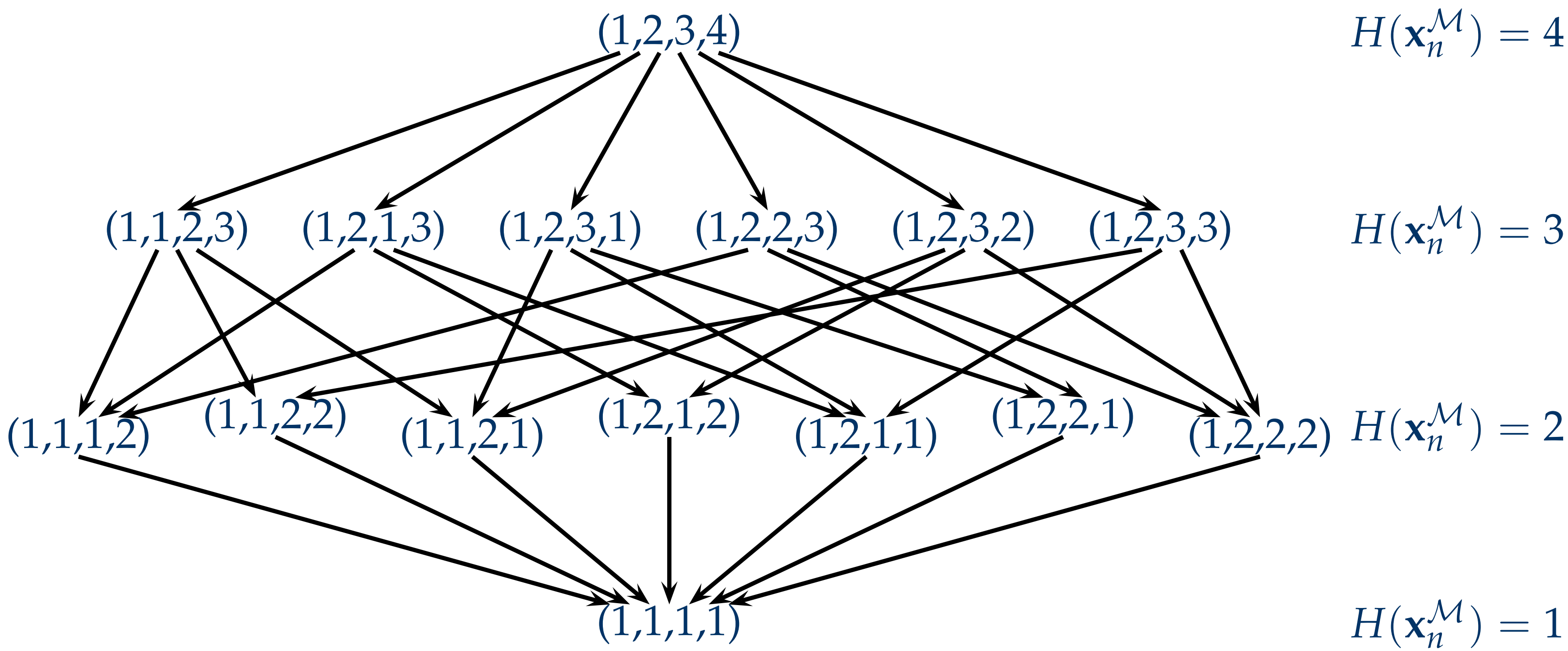
$$(1,1,2,3,1) \sim \{\{1,2,5\},\{3\},\{4\}\}$$
$$(1,2,2,1) \sim \{\{1,4\},\{2,3\}\}$$

$\mathcal{P}_n$ can be counted using Stirling numbers.

The next step is to introduce a partial ordering on $\mathcal{M}_n$. If $\mathbf{x}_n^{\mathcal{M}}, \mathbf{y}_n^{\mathcal{M}} \in \mathcal{M}_n$ then:

$$\mathbf{x}_n^{\mathcal{M}} \gtrsim \mathbf{y}_n^{\mathcal{M}} \text{ iff } (x_i = x_j) \Longrightarrow (y_i = y_j).$$

Visualising the Set $\mathcal{M}_n$



Hasse diagram for $\mathcal{M}_4$.

Relating this to False Matches

The censoring function $C(x)$ represents observation of only part of a plate. If $\mathbf{y} = C(\mathbf{x})$ then:

$$(x_i = x_j) \Longrightarrow (y_i = y_j).$$

The partial ordering now relates to the censoring function. If $\mathbf{z}$ is an n-tuple of observations and $\mathbf{x}_n^{\mathcal{M}}, \mathbf{y}_n^{\mathcal{M}} \in \mathcal{M}_n$ then:

$$(\mathbf{x}_n^{\mathcal{M}} \sim \mathbf{z}, \mathbf{y}_n^{\mathcal{M}} \sim C(\mathbf{z})) \Longrightarrow (\mathbf{y}_n^{\mathcal{M}} \precsim \mathbf{x}_n^{\mathcal{M}}).$$

Probability and Height

- The Height of $\mathbf{x}_n^{\mathcal{M}} \in \mathcal{M}_n$ is the maximal element. $H(1, 2, 2, 1, 3) = 3$.
- The Height of $\mathbf{x}_n^{\mathcal{M}} \sim \mathbf{y}$ is the number of distinct elements observed in the n-tuple $\mathbf{y}$.
- Define $p(n)$ as the probability that n distinct observations are observed to be the same in the censored data.
- The probability that $\mathbf{x}$ is a match is $p(H(\mathbf{y}_n^{\mathcal{M}}))$ where $\mathbf{y}_n^{\mathcal{M}} \sim \mathbf{x}$. Hence construct an algorithm for false matches.

Results and Problems

- Tests have been made on simulated data (see paper).
- In general the results are good – the estimator seems to be unbiased (as claimed).
- Variance on estimates is high.
- In real surveys estimating $p(n)$ can be difficult.
- In real surveys, the number of false matches can be huge.

Conclusion and a Request

- This framework provides new methods for surveys over more than two sites.
- Next: extend the method to provide confidence limits and deal with errors.
- The EPSRC has agreed to fund further work on this method to correct these problems and apply it to new data sets.
- I need to find data sets which people want analysing which might benefit from this method. (richard@manor.york.ac.uk).