The Interplay Between Research Innovation and Federal Statistical Practice

Stephen E. Fienberg
Department of Statistics
Center for Automated Learning & Discovery
Carnegie Mellon University
Pittsburgh PA

Some Themes

- Methodological innovation comes from many places and in different colors and flavors.
- Importance of cross-fertilization.
- Cycle of innovation.
- Examples: Bayesian inference, Log-linear models, Missing data and imputation, X-11-ARIMA.

Some Themes (cont.)

- Statistical thinking for non-statistical problems:
  - e.g., disclosure limitation.
- Putting several innovations to work simultaneously:
  - New proposal for census adjustment.
- New challenges

Methodological Innovation in Statistics

- Comes from many places and in different colors and flavors:
  - universities:
    - statisticians.
    - other methodologists.
  - government agencies.
  - Private industry.
  - nonprofits, e.g., NISS.
  - committees at the NRC.

Role of Statistical Models

- "All models are wrong, but some models are useful." John Tukey?
- Censuses and surveys attempt to describe and measure stochastic phenomena:
  - 1977 conference story.
  - U.S. census coverage assessment:
    - "Enumeration".
    - Accuracy and Coverage Evaluation (ACE) survey.
    - Demographic Analysis.

Importance of Cross-Fertilization

- Between government and academia.
- Among fields of application:
  - e.g., agriculture, psychology, computer science
- Between methodological domains e.g., sample surveys and experimental design:

  Jan van den Brakel Ph.D. thesis, EUR
- Using other fields to change statistics e.g., cognitive aspects of survey design:

  Krosnick and Silver; Conrad and Shober

Cycle of Innovation

- Problem need
- Formal formulation and abstraction
  - implementation on original problem
- Generalizations across domains
- General statistical theory
- Reapplication to problem and development of extensions
  - reformulation

Example 1: Bayesian Inference

- Laplace
  - role of inverse probability.
  - ratio-estimation for population estimation.
- Empirical Bayes and shrinkage (borrowing strength) - hierarchical models.
- NBC election night prediction model.
- Small area estimation.

Elliott and Little on census adjustment

Singh, Folsom, and Vaish

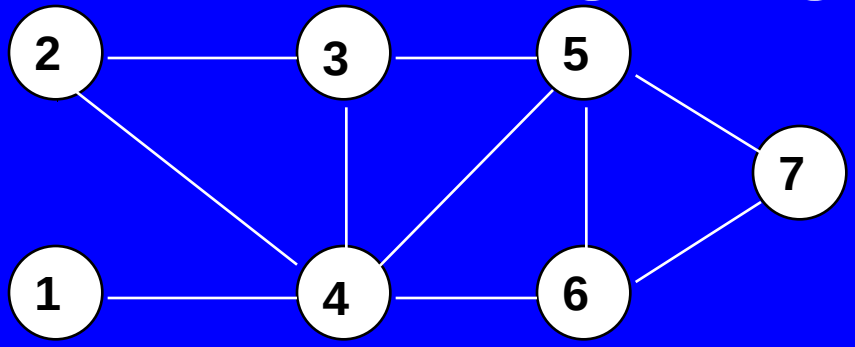
Example 2: Log-linear Models

- Deming-Stephan iterative proportional fitting algorithm (raking).
- Log-linear models and MLEs:
  - margins are minimal sufficient statistics.
  - implications for statistical reporting.
- Some generalizations: generalized IPF, graphical models, exponential families.
- Model-based inference for raking.

Graphical & Decomposable Log-linear Models

- Graphical models defined in terms of simultaneous conditional independence relationships or absence of edges in graph.

Example:



- Decomposable models correspond to triangulated graphs.

Example 3: Missing Data and Imputation

- Missing data problems in survey and censuses.
  - solutions: cold deck and hot deck imputation
- Rubin 1974 framework representing missingness as random variable.
- 1983 CNSTAT Report, Madow et al (3 vol.).
- Rubin (1997) - Bayesian approach of multiple imputation.
  - NCHS and other agency implementations

session on imputation and missing data

Example 4: X-11-ARIMA

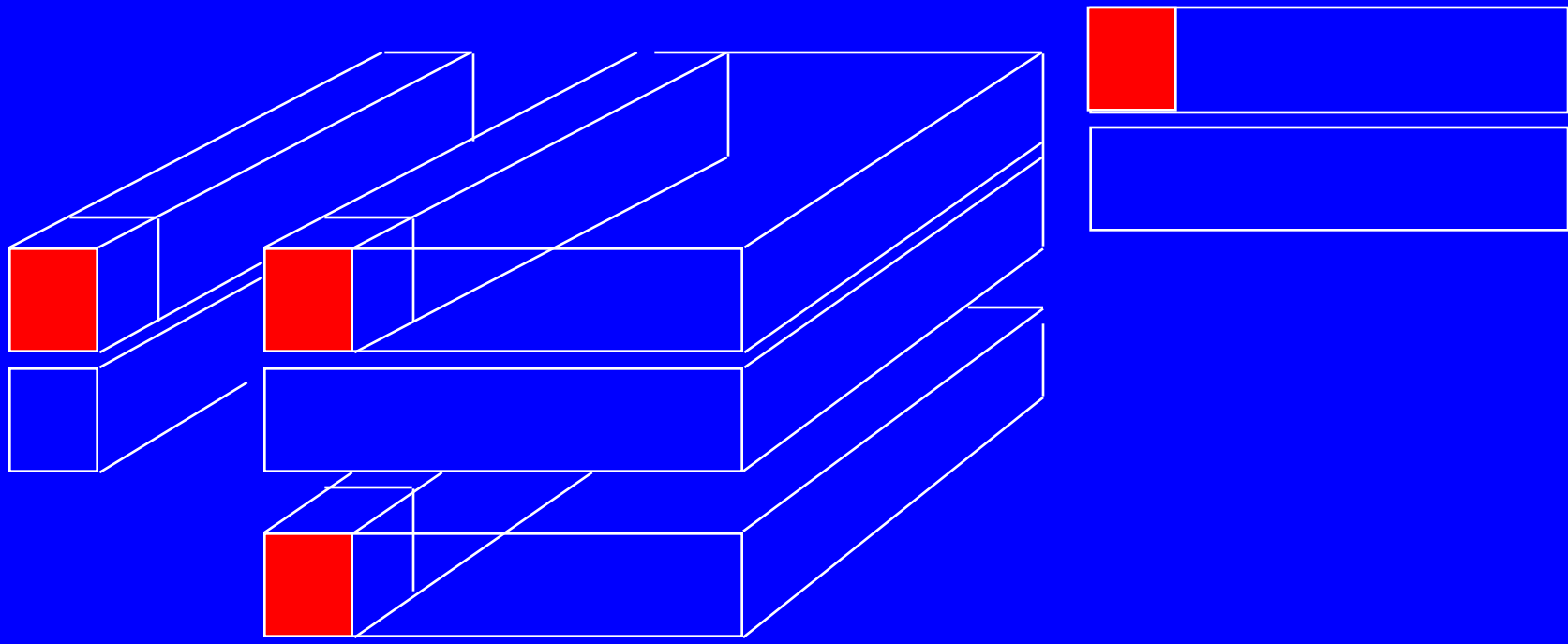
- BLS seasonal adjustment methods:
  - X-11 and seasonal factor method (Shishkin et al., 1967)
- ARIMA methods a la Box-Jenkins (1970):
  - Cleveland and Tiao (1976).
  - X-11 + ARIMA =X-11/ARIMA (Dagum, 1978).
- Levitan Commission Report (1979).
- X-12-ARIMA.

Burck and Salama

Statistical Thinking for Non-Statistical Problems

- Disclosure limitation was long thought of as non-statistical problem.
- Duncan and Lambert (1986, 1989) and others explained statistical basis.
- New methods for bounds for table entries as well as new perturbation methods are intimately linked to log-linear models.

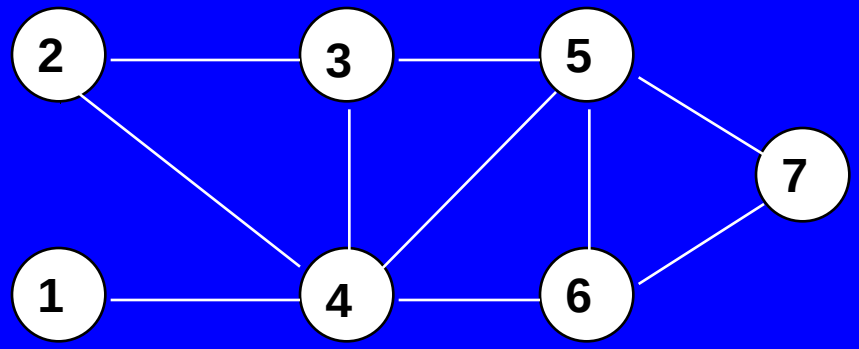
Query System Illustration (k=3)



Challenge: **Scaling up approach for large k.**

Bounds for Entries in *k*-Way Contingency Table

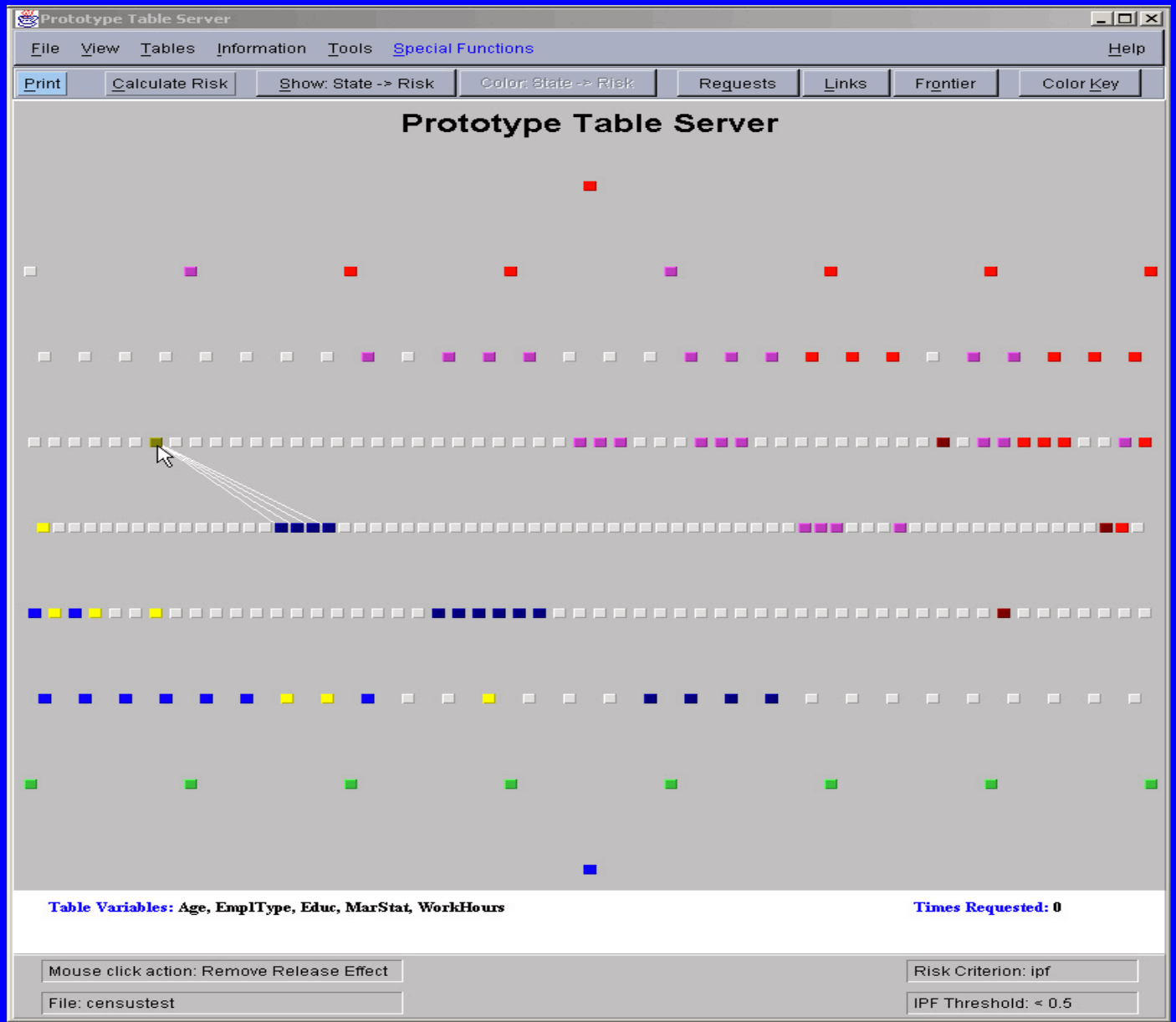
- Explicit formulas for sharp bounds if margins correspond to cliques in decomposable graph. **(Dobra and Fienberg , 2000)**
  - ***Upper bound:*** **minimum of relevant margins.**
  - ***Lower bound:*** **maximum of zero, or sum of relevant margins minus separators.**
- **Example:**
  - **cliques are {14}, {234}, {345}, {456}, {567}**
  - **separators are {4}, {34}, {45}, {56}**



More Bound Results

- Log-linear model related extensions for
  - reducible graphical case.
  - if all (k-1) dimensional margins fixed.
- General "tree-structured" algorithm. **Dobra (2000)**
  - Applied to 16-way table in a recent test.
- Relationship between bounds results and methods for perturbing tables.
- Measures of risk and their role.

NISS Table Server: 8-Way Table



Census Estimation

- 2000 census adjustment model uses variant of dual systems estimation, combining census and ACE data, to correct for
  - omissions.
  - erroneous enumerations (subtracted out before DSE is applied).
- DSE is based on log-linear model of independence.

UK and US papers on DSE adjustment methods

Dual Systems Setup

		Sample		
		In	Out	Total
Census	In	a	b	n_1
	Out	c	d??	$N - n_1$
	Total	n_2	$N - n_2$	N??

$$\hat{N} = n_1 n_2 / a$$

Putting Several Innovations to Work Simultaneously

- Third system for each block based on administrative records - ***StARS/AREX***
- Log-linear models allow us to model dependence between census and ACE:
  - in all ***K*** blocks have census and AREX data.
  - in sample of ***k*** blocks also have ACE data.
  - missing data on ACE in remaining ***K-k*** blocks.

Triple System Approach

- Example of small minority stratum from 1988 St. Louis census dress rehearsal.

		Administrative List				
		P		A		
		Sample		Sample		
		P	A	P	A	
Census	P	58	12	69	41	$n = 268$
	A	11	43	34	-	

New Bayesian Model

- Bayesian triple-systems model:
  - correction for EEs in census and administrative records.
  - missing ACE data in *K*-*k* blocks.
  - dependencies (heterogeneity) among lists.
  - smoothing across blocks and/or strata using something like logistic regression model.
- Revised block estimates to borrow strength across blocks *and* from new source.

New Challenges

Challenges of Multiple-Media Survey Data

- **Example:** Wright-Patterson A.F. Base Sample Survey
  - Demographic data PLUS three 3-dimensional full-body laser surface scans per respondent.
  - Interest in “functionals” measured from scans.
- Ultimately will have 50 MB of data for each of 10,000 individuals in U.S. and The Netherlands/Italy.

Some Methodological Issues With Image Data

- Data storage and retrieval.
- Data reduction, coding, and functionals:
  - design vs. model-based inference.
- Disclosure limitation methodology.

Standing Statistics On Its Head