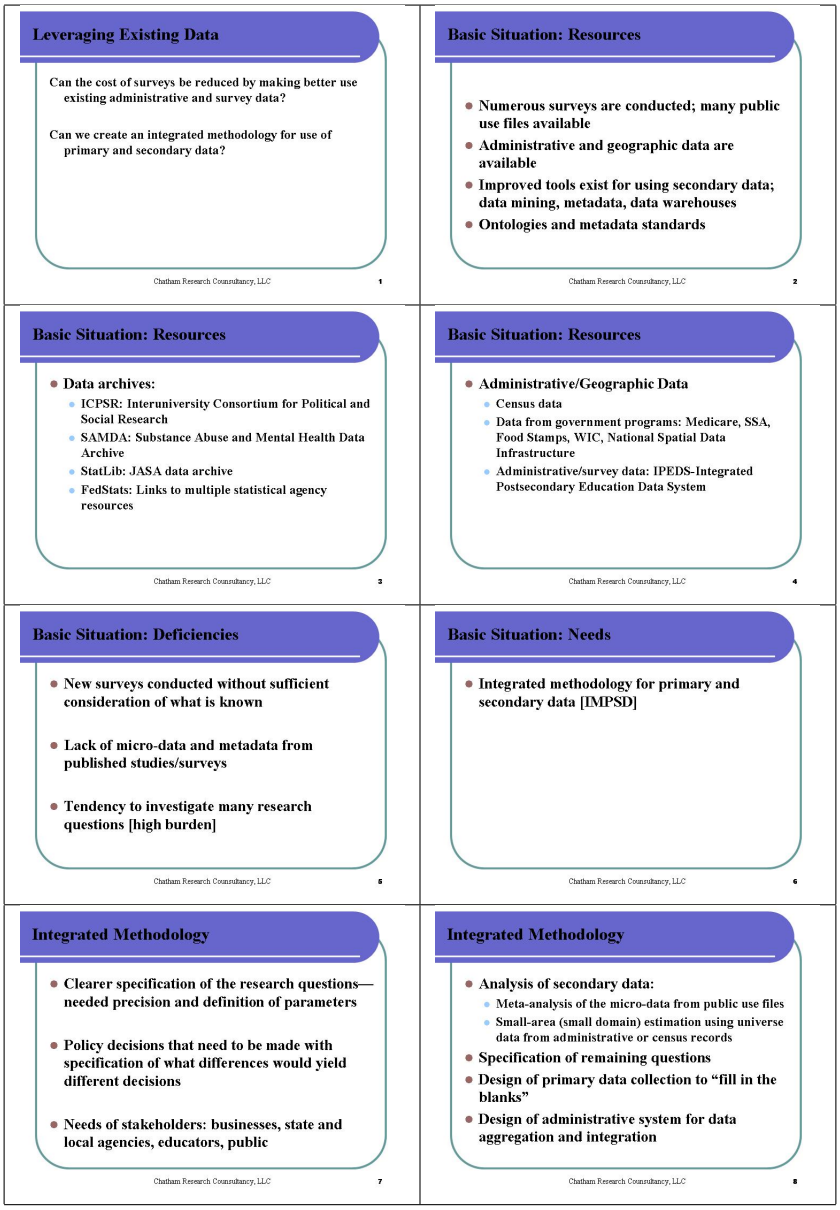
2.5 Judith Lessler: Leveraging Existing Data



Initial Data Analysis

- Set of common variables—basic demographics
- Creation of analysis variables that create more overlap—categorical variables for quantitative data
- Modeling (imputation) for blocks of missing data

Chatham Research Consultancy, LLC 9

Initial Data Analysis

Survey	Set A	Set B	Set C	Set D	Set E
1					
2					
3					
4					
5					

Chatham Research Consultancy, LLC 10

Initial Data Analysis

Survey	Set A	Set B	Set C	Set D	Set E
1					
2					
3				model	
4					
5					

Chatham Research Consultancy, LLC 11

Assess Needs for Primary Data

- More precision needed for some sets of variables
- Confirm modeled relationships between sets of variables
- Increase representation of subgroups

Chatham Research Consultancy, LLC 12

Optimum Design

- What is most cost-effective design for obtaining additional information
  - Supplemental data collections using follow-up surveys
  - New survey of selected subgroups
  - Analysis of existing administrative data

Chatham Research Consultancy, LLC 13

Needed Infrastructure

- Alignment of operational definitions of some basic variables across surveys
- Universal adoption of NIH data sharing policy for all government sponsored surveys
- Software that facilitates data integration

Chatham Research Consultancy, LLC 14

NIH Data Sharing Plan

- Publishing
- Distribution of data by researcher
- Data enclave [secure environment]
- Data archive
- Mixed mode

Chatham Research Consultancy, LLC 15

NIH Data Sharing Plan

- Schedule for data sharing
- Format of final data set
- Documentation to be provided
- Provision of analytic tools, if any
- Need for sharing agreement
- Mode of data sharing

Chatham Research Consultancy, LLC 16

IMPSD Summary

- Specification of questions and quality parameters
- Meta-analysis of existing survey and administrative data
- Design of primary data collection to supplement or confirm conclusions from secondary data analysis
- Required provision of survey results to other researchers
- Data integration software

Chatham Research Consultancy, LLC 17

Implementation of IMPSD

- Advantages of such an approach
- Disadvantages
- Barriers
- Facilitators

Chatham Research Consultancy, LLC 18

2.6 Alan Karr: Principled Cost-Quality Tradeoffs

NISS

Principled Cost-Quality Tradeoffs

Alan F. Karr
National Institute of Statistical Sciences
karr@niss.org
April 19, 2006

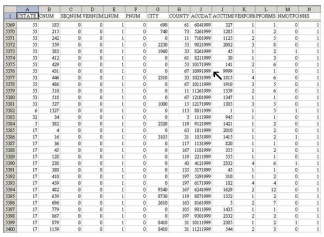
Outline: Simple to Not-So-Simple

- Analytical approaches
  - Deterministic case
    - Discrete case
    - Parametric case
  - Stochastic case
- Modeling approaches: a sketch
- Some ugly questions
- DISCUSSION

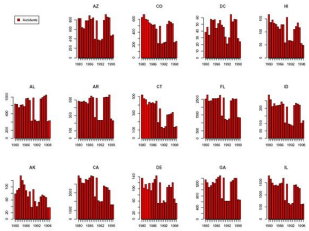
Big Questions

- What is cost?
  - Out-of-pocket
  - Opportunity
  - Time (?)
  - To whom?
- What is quality?
  - Accuracy
  - Timeliness
  - Response rate
  - Accessibility, relevance, coherence, interpretability, …
  - Reference: A. F. Karr, D. L. Banks, A. P. Sanil (2006). Data quality: A statistical perspective. *Statistical Methodology* **3(2)** 137-173.

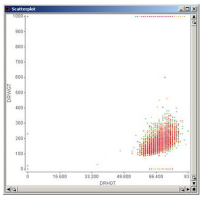
FARS-A 1999 Excerpt



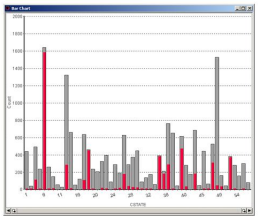
FARS-A File Sizes



FARS-D Driver Height/Weight



FARS-A: Missing EMS Arrival Time



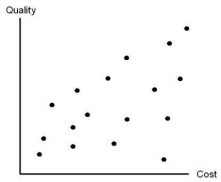
TRI Excerpt

TRIFID	Air	SurfH2O	Land	OnSite	OffSite	Total	Year
38109RFNDM257WE	15300	0	0	15300	222163	237463	1988
38109RFNDM257WE	15300	0	0	15300	148151	163451	1989
38109RFNDM257WE	10200	0	440000	450200	19444	469644	1990
38109RFNDM257WE	3800	0	0	3800	76850	80650	1991
38109RFNDM257WE	1900	0	0	1900	313750	315650	1992
44706THTMK4511F	1200	21	0	1221	240090	241311	1988
44706THTMK4511F	1300	19	0	1319	265110	266429	1989
44706THTMK4511F	1400	13	0	1413	280026	281439	1990
44706THTMK4511F	1500	1	0	1501	290030	291531	1991
44706THTMK4511F	1800	0	0	1800	51025	52825	1992
44706THTMK4511F	1800	2	0	1802	29	1831	1993
44706THTMK4511F	540	2	0	542	32	574	1994
44706THTMK4511F	604	1	0	605	5480	6085	1995
44706THTMK4511F	550	2	0	552	106380	106932	1996
44706THTMK4511F	540	1	0	541	330380	330921	1997
44706THTMK4511F	200	1	0	201	282800	283001	1998

Deterministic Case

- Formulation
  - Actions *a*
    - Actions may have very complex structure
    - May be many actions
    - Discrete case: finite, not necessarily structured set of actions
    - Parametric case: actions qualitatively identical, but depend on some parameter, such as sample size
  - Known costs *C(a)*
    - Initially, one-dimensional
  - Known resulting data qualities *Q(a)*
    - Initially, one dimensional

Discrete Case



Optimization Approaches

- Maximize quality subject to upper bound on cost

$$a^* = \arg\max_a Q(a) \quad \text{s.t.} C(a) \le C_{\text{bound}}$$

- Minimize cost subject to lower bound on quality

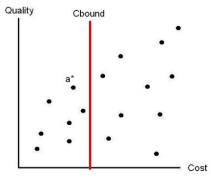
$$a^* = \arg\min_a C(a) \quad \text{s.t. } Q(a) \ge Q_{\text{bound}}$$

Optimization Approaches—2

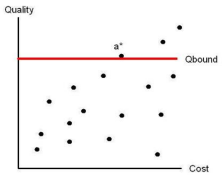
- Maximize utility: given utility function *U(c,q)*,

$$a^* = \arg\max_a U(C(a), Q(a))$$

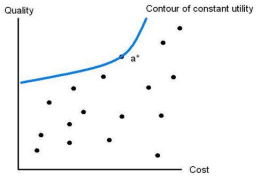
Maximizing Quality s.t. Cost Bound



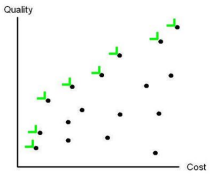
Minimizing Cost s.t. Quality Bound



Maximizing Utility



The Cost-Quality Frontier



Computational Considerations

- Can "always" restrict to frontier: actions *a* for which there is no *a'* such that $C(a') \le C(a)$ and $Q(a') \ge Q(a)$
  - This may help a lot
  - There are efficient algorithms for calculating the frontier
- Maximizing quality subject to cost bound and minimizing cost subject to quality bound may be solvable only by enumeration
- Utility maximization solvable by bisection search, but this is still effectively enumeration
- Size of problem = size of frontier

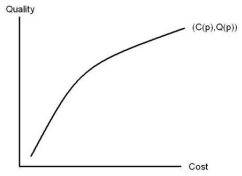
Other Computational Technologies

- Simulated annealing
  - Requires some concept of "neighboring" actions
- Genetic algorithms
  - Requires that actions have "separable" components

"Parametric" Case

- Actions are parametrized by a (numerical) parameter p
  - Example: parameter = sample size
- Then
  - $a = a(p)$
  - $C(p) = C(a(p))$
  - $Q(p) = Q(a(p))$
- Everything else is more or less the same

Pictorial View



Multi-dimensional Cost and Quality

- Frontier carries over: set of actions is partially ordered by
  $a_1 \preceq a_2$ if and only if
  - $C_j(a_2) \leq C_j(a_1)$ for all cost measures C_j
  - $Q_k(a_2) \geq Q_k(a_1)$ for all quality measures Q_k
- Problem: with lots of dimensions, the frontier is not necessarily "small"
- Utility functions not clear

Stochastic Costs and Quality

- Formulation
  - Given a, $C(a)$ and $Q(a)$ are (dependent!) random variables with distribution $F_a(x,y)$
- Issues
  - How much of the deterministic framework carries over?
  - From data are the F_a estimated? Or are they based on expert opinion? Or some combination?
  - How are uncertainties in the estimated F_a to be incorporated in the framework?

"Elementary" Approaches

- Use only means
  - $C(a) = E[C(a)]$
  - $Q(a) = E[Q(a)]$
- Combine means and standard deviations. Example:

$$C'(a) = E[C(a)] + \lambda \mathsf{StdDev}(C(a))$$
$$Q'(a) = E[Q(a)] - \mu \mathsf{StdDev}(Q(a))$$

Not Quite So Elementary Approaches

- Use means and standard deviations to define the partial order
  $a_1 \preceq a_2$ if and only if

$$E[C(a_2)] \leq E[C(a_1)]$$
$$\mathsf{StdDev}(C(a_2)) \leq \mathsf{StdDev}(C(a_1))$$
$$E[Q(a_2)] \geq E[Q(a_2)]$$
$$\mathsf{StdDev}(Q(a_2)) \leq \mathsf{StdDev}(Q(a_1))$$

- Related approaches in the portfolio analysis literature

More Complicated Approaches

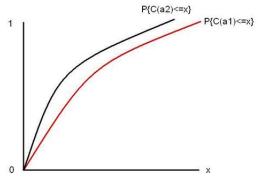
- Use stochastic ordering to define the partial order. For example,

  $a_1 \preceq a_2$ if and only if

$$P\{C(a_2) \leq x\} \geq P\{C(a_1) \leq x\} \text{ for all } x$$
$$P\{Q(a_1) \leq y\} \leq P\{Q(a_2) \leq y\} \text{ for all } y$$

- Lots of applications in the literature
  - Example: Reliability

Pictorial View of Stochastic Ordering



What About Optimization?

- Constraints: replace by probabilistic versions:

$$P\{C(a) \leq C_{\text{bound}}\} \geq p_C$$
$$P\{Q(a) \geq Q_{\text{bound}}\} \geq p_Q$$

- Utility functions: ?????
  - One approach: derive from loss functions using concepts from statistical decision theory

Back to the Really Hard Questions

- From data are the F_a estimated? Or are they based on expert opinion? Or a combination?
- How are uncertainties in the estimated F_a to be incorporated in the framework?
  - Bayesian methods can account for this
- Are there other implications?
  - Example: confidentiality
- Temporal evolution of surveys

Dynamics

- Formulation is inherently stochastic:
  - At time t, given a, $C_t(a)$ and $Q_t(a)$ are (dependent!) random variables with distribution $F_a(t,x,y)$ that reflects all information available at t
- Issues
  - Ongoing incorporation of information: Bayesian techniques can do this, but need a model for that information
  - Interventions seem very problematic to model analytically

Problems with the Analytical Approach

- Getting hopelessly complicated, especially once dynamics are present
- Sensitivity analyses hard
- Disconnected from the science, therefore limited insight
  - Example: if a involves adaptive increase in sampling or follow-up, a la Groves, how does this get represented in F_a?
  - Example: not clear how to differentiate different kinds of surveys (e.g., drug use vs. educational experience) or populations

What About a Modeling Approach?

- Technological basis: agent-based simulation:
  - Represent agents (e.g., interviewers and respondents) explicitly
  - Specify agent-level behavior and interactions
  - Specify "context" and "interventions"
- Key {point/hope/blind faith}s
  - Context and interventions are fairly literal
  - Sensible models of agent produce "correct" emergent, macroscopic behavior
- Variability characterized via replication of simulations

Example: TRANSIMS (LANL)

- Metro-area simulator of traffic
  - Context:
    - Population (synthesized from Census PUMS)
    - Activities (synthesized from survey)
    - Geography: Locations of houses, businesses, …, road network
  - Interventions
    - Signal settings, …
- TRANSIMS models traffic "realistically" based on simple car-following rules
  - Example: speed up if there is enough space between you and the car in front of you

What Would It Mean for Surveys?

- Set of abstractions
  - Agents
    - Respondents
    - Interviewers
  - Responses
    - Cost
    - Quality
  - Interventions
    - Increase incentives
    - Increase intensity of non-respondent follow-up

More Implications

- "Simple" models that relate agent behavior to interventions
  - Example: $P\{\text{response}\} = f(\text{agent characteristics, item, intervention})$
- Some intra-agent interactions may be ignorable, especially to start

And More …

- How to validate agent-based models?
  - Does bias matter if interventions are only being compared?
- Can models be simple enough that replication is possible?
- Is every survey a special case?
  - There is sobering evidence that there is not a science of data quality.

Some Ugly Questions

- Is one person's cost another's quality?
- Shouldn't quality be measured by inferential uses of the data?

The End

- Thank you

•Let's discuss!