Summary of Survey Cost White Paper

Alan F. Karr

National Institute of Statistical Sciences

NISS Affiliates Survey Day, Chicago
April 22, 2010

NISS
The Statistics Community Serving the Nation

Outline

1. Summary of the Workshop
   - Background
   - Findings and Caveats
2. Cost–Quality Tradeoffs
   - Deterministic Case
   - Stochastic Case
3. The Research Agenda
   - Survey Models
   - Statistical Modeling as a Means of Reducing Costs
   - Survey Integration
   - Other Research Issues
4. Conclusion
   - One Final Point

NISS
The Statistics Community Serving the Nation

The Facts

Dates and Venue April 18–19, 2006; NCES, Washington

Organizing Committee Lawrence Cox (NCHS), John Eltinge (BLS), Graham Kalton (Westat), Alan Karr (NISS), Daniel Kasprzyk (MPR), Myron Katzoff (NCHS), Partha Lahiri (University of Maryland), Judy Lessler (Chatham Research Consultancy), Marilyn Seastrom (NCES), Alan Tupek (Census), Doug Williams (Williams Consulting)

Purposes

- Articulate clearly the survey cost problems faced by the federal statistical agencies
- Construct a research agenda for survey cost methodology and modeling that addresses these problems
- Identify the key next steps in pursuing the research agenda

Program: Day 1

9:00 AM	Tutorial I: John Eltinge, BLS: Survey Basics, including Costs
10:15	Break
10:45	Tutorial II: David Banks, Duke University: Decision Theory and Simulation
12:00 N	Lunch
1:00	Welcome and Introductions Alan Karr, NISS
1:15	Robert Groves, University of Michigan: In-Process Adaptation Discussion Leader: Alan Karr, NISS
3:15	Break
3:45	Judith Lessler, Chatham Research Consultancy: Leveraging Multiple Data Collections Discussion Leader: Partha Lahiri, University of Maryland
5:45	Adjourn for the Day

Program: Day 2

8:30 AM	Three-Minute Madness: All Attendees May Speak
9:30	Break
10:00	Alan Karr, NISS: Principled Trade-offs between Costs and Quality Discussion Leader: Myron Katzoff, NCHS
12:00 N	Breakout Discussions Over Lunch
2:00 PM	Final Panel Discussion Breakout Discussion Leaders Workshop Organizers
3:30	Workshop Adjourns

The Breakout Discussions

- Paradata
- Real-Time Monitoring
- Survey Models
- Budget Reductions
- Borrowing Strength

Principal Findings

1. Existing theory, methodology and software tools are not adequate to address current, let alone emerging, problems associated with survey costs
2. The need for research to address the gaps is pressing and increasing
3. The three most promising, and clearly inter-related, research thrusts, in terms of significant and immediate impact, are
    - Agent-based or other simulation models of surveys
    - Statistical modeling as a means of reducing costs
    - Integrating data across (and beyond!) surveys

Caveats

- Access to sufficient and sufficiently high quality paradata on survey costs and effort
- Access to qualitative expertise about survey administration that resides in program managers, field managers and interviewers
- Capturing a full spectrum of fixed and variable costs
- Engagement by data collecting organizations
- High-level drivers of the problems, including federal statutes, Congressional mandates to collect data, data confidentiality and scientific generalizability

Big Questions

1. What is cost?
   - Out-of-pocket, opportunity, time (?), . . .
   - To whom?
2. 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

Deterministic Case: Formulation

Actions a

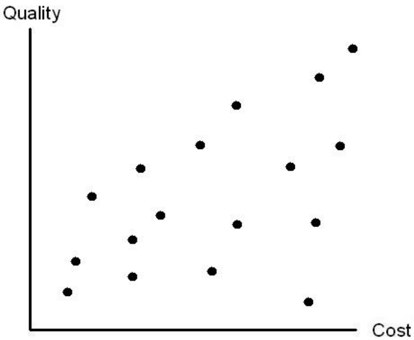
- Very complex structure
- Large, unstructured set of actions

Known costs $C(a)$

Known data qualities $Q(a)$

Duh Issue Higher quality costs more

Discrete Case



Optimization Approaches

Maximize quality given upper bound on cost

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

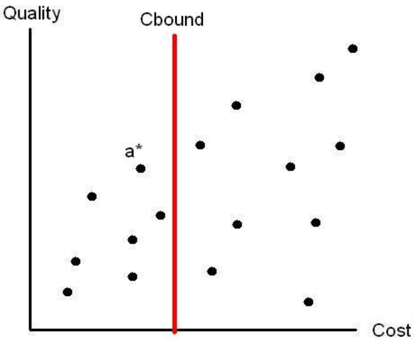
Minimize cost given lower bound on quality

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

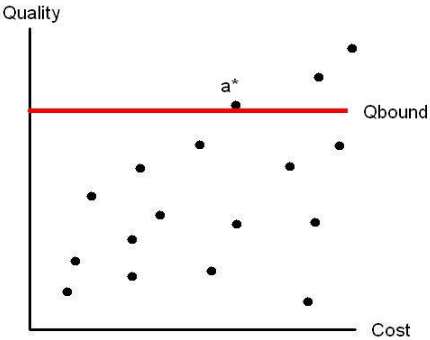
Maximize utility function

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

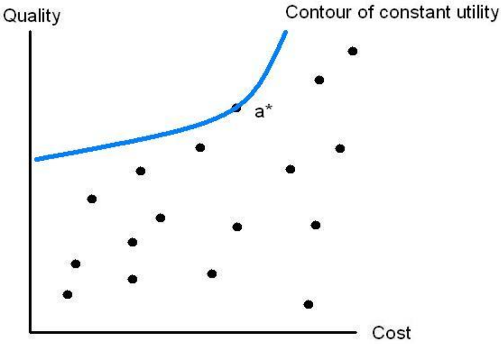
Approach 1



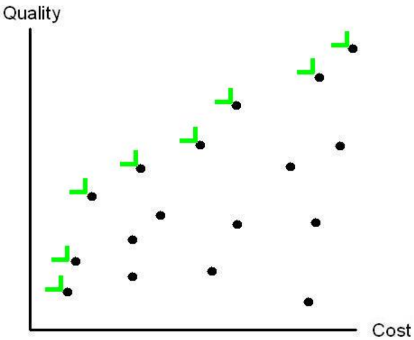
Approach 2



Approach 3



Cost–Quality Frontier



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

But with lots of dimensions, the frontier is not necessarily "small," and utility functions are not clear

Setting

Formulation Given a, $C(a)$ and $Q(a)$ are dependent random variables with distribution $F_a(x, y)$

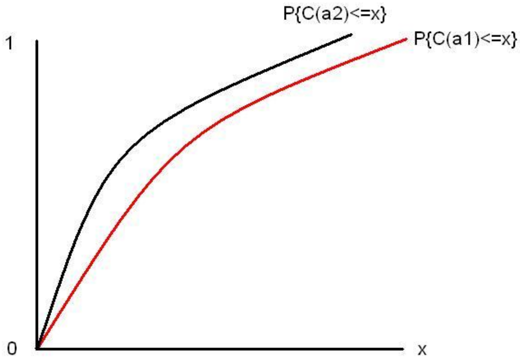
Issues

- 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?

Approaches: Simple to Hard

1. Reduce to deterministic case by using only means
2. Combine means and standard deviations:
   $C'(a) = E[C(a)] + \lambda \mathrm{SD}(C(a)),$
   $Q'(a) = E[Q(a)] - \mu \mathrm{SD}(Q(a))$
3. Use means and standard deviations to define partial order (á la 1960's portfolio analysis)
4. Use stochastic ordering to define a partial order

Stochastic Ordering



Re-setting the Context

- Existing theory, methodology and software tools are not adequate to address current, let alone emerging, problems
- The need for research to address the gaps is pressing and increasing
- The right research would have significant and immediate impact
- The research challenges are substantial, and of interest not only to statistical scientists, but also to those interested in such fields as operations research, optimization and agent-based simulation
- Without high-quality paradata and access to qualitative information about survey adminstration, the research will be speculative and its impact diminished

The Big Need

Agent-based or other simulation models of surveys in order to support

Prospective evaluation of policies and interventions, as an alternative to costly and/or infeasible "experimentation"

Sensitivity analyses, to determine which controllable factors actually do affect survey costs and data quality, and to quantify the effects

Clarification of paradata needs, which is especially important in light of difficulties in obtaining high quality paradata

and given

Intractability of analytical approaches

Challenges

- Constructing quantifiable indicators of cost and quality
- Controlling the complexity of models
- Availability of high quality paradata, especially on cost and effort, at sufficient levels of detail
- Proper role for domain knowledge
- Validation of models, given limited “real world” data

The Question

How and to what extent can statistical modeling address cost-related issues? Examples:

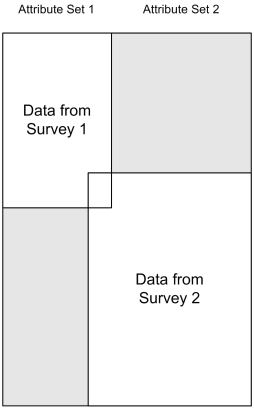
1. Can improved modeling of nonresponse bias compensate for lower response rates resulting from cost considerations?
2. Can statistical modeling produce valid, useful statistical estimates at high geographical resolution without correspondingly high-resolution sample sizes?
3. Can design concepts such as matrix sampling in educational assessment be applied to collect more information without increasing respondent burden?
4. Can statistical modeling allow borrowing strength across surveys in order to reduce costs?

The Concept

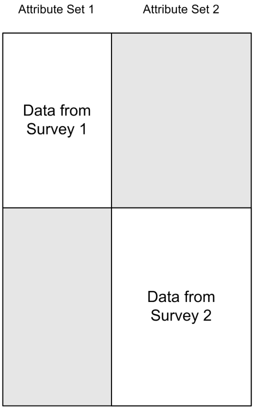
Reduce costs by

- “Borrowing strength”
- “Leveraging existing data”
- “Integrating data across surveys”
- *Added subsequently:* Using administrative data

Case 1: Duplicated Attributes and Cases



Disjoint Surveys



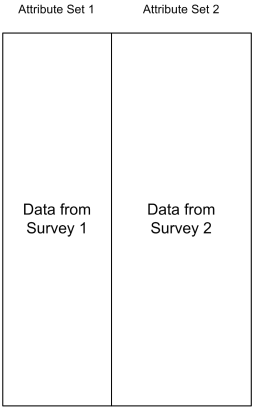
Attribute Set 1

Attribute Set 2

Data from Survey 1

Data from Survey 2

Identical Cases, No Duplicated Attributes



Attribute Set 1

Attribute Set 2

Data from Survey 1

Data from Survey 2

Challenges

- Need powerful modeling to characterize relationships between sets of attributes in Set 1 and those in Set 2
- Practical impediments
    - Example: agency conducting Survey 2 may be unable to learn what respondents are in Survey 1
- Temporal aspects
    - Example: if Survey 2 occurs five years after Survey 1, attributes in Survey 1 may no longer be correct
- Data quality differences between surveys

Quick Summary

- Paradata
  - Example: do traditional concepts of uncertainty apply?
- Multiple stakeholders

And Lingering Concerns

- Availability and quality of paradata
- Capturing a full spectrum of costs
- Engagement by data collection organizations
- Whether data required by agent-based survey models exist, or can be expected to exist
- High-level drivers, including federal statues
  - Examples: Congressional mandates to collect data at certain geographical resolutions, regulations that prohibit data sharing, data confidentiality
- Understanding data uses
  - Example: “granularity” of policy decisions has direct implications for survey design, cost and analysis
- Generalizability

And the Biggest Challenge

Linking cost–quality tradeoffs to the decisions that will be made based on analyses of the data