Tutorial I:
Survey Basics, Including Costs

John L. Eltinge

Bureau of Labor Statistics

Eltinge.John@bls.gov

April 22, 2010

Annual Meeting - NISS Affiliates

Disclaimer:

The views expressed in this paper are those of the author and do not necessarily reflect the policies of the Bureau of Labor Statistics.

Acknowledgements: Many thanks to David Chapman, Steve Cohen, Larry Cox, Pat Getz, Bob Groves, Rachel Harter, Steve Haslett, Mike Horrigan, Larry Huff, Sylvia Leaver, Judy Lessler, Bill Mockovak, Barbara O’Hare, Steve Pedlow, Jay Ryan, Marilyn Seastrom, Joe Sedransk, Clyde Tucker, Rick Valliant, Doug Williams, Kirk Wolter and Ibrahim Yansaneh for helpful discussions

Overview:

I. Sample Surveys and Administrative Record Systems

II. Components of Data Quality and Risk

III. Literature on Survey Costs

IV. Two Classes of Methodological Questions

I. Sample Surveys and Administrative Record Systems

A. Goal of Government Statistical Agencies and Other Large Survey Organizations:

Provide the best available information on a given topic for the lowest reasonable cost

B. Information: Point ests, inference?

Cost: To agency? To data user?

C. Traditional View of Sample Surveys

Superpopulation model $\xi(\theta)$

generates a finite population U

of size N with characteristics

$$(Y_i, X_i), i = 1, \ldots, N$$

1. Goal: Estimation and inference for

$\gamma = g(\theta)$ superpopulation quantity

or the corresponding finite pop quantity defined through an estimating function

$$E_N(\theta_U) = \sum_{i \in U} f(Y_i, X_i; \theta_U) - v(\theta_U) = 0$$

e.g., Binder (1983, *Int. Stat. Rev.);* Scott and Wild (1986, Biometrics)

Examples: Finite population means, totals, quantiles, regression coefficients, parameters of a generalized linear model

Historical focus of most statistical agencies:

Simple descriptive quantities (means, totals, ratios) for large aggregates (full population or large subpopulations)

Ex: Current Employment Survey:

Estimated total employment and one-month change:

- Essentially all non-agricultural U.S. employers

- Eleven large industrial “supersectors”

2. Ideally, we would take a census (100% sample) of all units in U compute the desired quantities, and publish results.

3. Seven practical constraints that make (2) unfeasible:

a. Direct use of information from administrative record system not entirely feasible:

- Definitional or aggregation issues

- Diminishing returns (as measured by inferential quality) from very large sample sizes

- Constraints on processing systems

Solution: Base estimation and inference on a sample of units

b. Candidate frames (specification of prospective sample units): incomplete

Example: New construction

Example: Aggregation

Solution: Use multiple frames, some with nesting (area frames, list frames) and sample separately from each frame

c. Nested structure of population:

May not be able to identify units of interest directly from the available frames, or cost may be prohibitive

Solution: Use cluster sampling or other forms of multistage sampling

Ex: Sample counties, then neighborhoods, then houses

d. Subpopulation membership (possibly rare) not reflected in frame

Solution: Two-phase sampling

Large sample with cheap measures

Follow-up smaller sample of “interesting” units

Epidemiological variant: Case-control studies

e. Membership in rare subpopulation not reflected in frame

and

significant network structure in subpopulation membership

Example:

Wildlife sampling, some human social networks

Solution: Adaptive or network sampling

f. Heterogeneity across population units:

Example: Sizes of establishments

Solution: Sample units with unequal probabilities (e.g., probability proportional to size)

g. Heterogeneity across identifiable subpopulations:

Examples:

Industry, size class, occupation

Solution: Stratified sampling (partition into subpopulations and sample separately from each subpopulation)

4. Resulting complications:

a. Generally impossible or inefficient to draw a simple random sample from U

Alternative: Select a sample S of size n through a complex sample design that involves the use of one or more of:

- Stratification
- Unequal selection probabilities
- Clustering or other forms of dependent selection (two-phase, adaptive)

b. Consequently, observations are not iid

c. Multiple stakeholders: No uniform consensus on basis for estimation and inference

Model $\xi(\theta)$ generally not truly known and often the subject of controversy (esp. regarding appropriate conditioning)

3. Criteria for estimator performance:

a. At a minimum, we want good properties when performance is evaluated with respect to the sample design:

$$E_p(\hat{\theta}_S) \cong \theta_U$$

i.e., performance “in repeated sampling under the specified design”

b. Note minimalist approach:

i. Limited assumptions:

How we drew the sample

- Reduced (eliminated?) risk of model failure

ii. (Almost) no assumptions on population U

iii. Modest claim for performance:

wrt repeated sampling from *this* population

- Should be minimally acceptable to a wide range of stakeholders

c. In its most pure form, effectively ignores issues with:

- Nonresponse

- Measurement error

- Loss of efficiency (under specified model constraints)

Thus, need to introduce some amount of modeling into any serious discussion of performance, but this generally is done with considerable caution

d. Ideally, prefer good properties when performance is evaluated wrt either the sample design, or the underlying superpopulation model, or both

$$E_{p\xi}(\hat{\theta}_S) \cong \theta$$

as well as under moderate deviations (via sparse effect models?) from specified superpopulation model

Similarly for variance ests, inference methods

- Asymptotics usually through triangular-array type arguments: increasing N, n, conditions

4. Primary approach for statistical agencies: Point estimation method through solution of weighted estimating equation:

$$\hat{E}_n(\hat{\theta}_S) = \sum_{i \in S} w_i f(Y_i, X_i; \hat{\theta}_S) - v(\hat{\theta}_S) = 0$$

where weights w_i are proportional to the inverse of selection probabilities (with modifications for auxiliary information)

5. Examples:

Population total: $\hat{Y} = \sum_{i \in S} w_i Y_i$

Mean of subpopulation (domain) D:

$$\hat{\bar{Y}}_D = \left(\sum_{i \in S \cap D} w_i \right)^{-1} \sum_{i \in S \cap D} w_i Y_i$$

6. Justification of a given procedure (sample design, collection method and estimation method) generally involves a combination of:

a. Optimization of formal criterion (loss function, weighted likelihood function)

b. Performance evaluated with respect to:

   - Sample design
   - Specified model, and deviations therefrom

c. Compatibility with production systems

D. Related Comment on Costs and Risks Related to Modeling

1. Costs:

a. Labor for model fitting and monitoring

b. Access to, and use of, auxiliary data X
(Ex: Multistate metropolitan areas)

c. Modification of production systems

d. Dissemination of results and exposition of risks for stakeholders

2. Risks (beyond standard measures of error)

a. Model failure: Greatest interest by stakeholders may coincide with conditions under which models may be most problematic

   - Change-points in economic conditions

   - Special subpopulations

b. Misinterpretation by stakeholders

   - Highly exploratory data analysis, implicit multiple inference (FDR, other risk measures)

c. Reduction in perceived value for stakeholders

d. Resulting reputational risk for statistical agency

E. Parallel Developments on Costs and Data Quality Related to Design of:

1. Instruments

2. Fieldwork

3. Microdata review

4. Production systems

5. Dissemination

II. Components of Data Quality and Risk

A. Strong Links Between Perceptions
of Quality and Utility

B. (Brackstone, 1999; many other variants)

Accuracy
Timeliness
Accessibility

Relevance
Interpretability
Coherence

C. Risk: Failure in one or more quality components
Implicitly reflect costs to some data users

III. Literature on Survey Costs

A. Broad Overviews

Sudman, S. (1967). *Reducing the Costs of Surveys.* Chicago: Aldine.

Pearson, R.W. and R.F. Boruch (1986). *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits.* New York: Springer.

Groves, R.M. (1986). *Survey Errors and Survey Costs*. New York: Wiley

United Nations Statistical Division (2005) http://unstats.un.org/unsd/hhsurveys/

Karr, A. and M. Last (2006). *Survey Costs: Workshop Report and White Paper.*

B. Specific Case Studies:

Bibliography available

1. Tend to be very focused on one specific cost component

2. Consequently, any one study is of limited benefit for broad discussion of cost-benefit trade-offs

C. Important Limitations on Available Survey Cost Information

1. Large fixed costs, often not well-identified

   a. Human/intellectual capital investment

      cf. “capacity building” in UNSD (2005)

   b. Legacy systems (sample, instrument, field, production)

2. Aggregation effects

a. Operational constraints

b. Filters imposed by project management procedures, incentives

c. Reporting constraints

3. Side comment:

   Incorporate more detailed variable cost accounting into OMB 83-I process?

IV. Two Classes of Methodological Questions on Survey Cost Structures and Optimization Thereof

A. Empirical Evidence on Survey Costs and Survey Efficiency

1. Gaps in current information

Conceptual Model for Cost

$$Y = A + B X + e$$

Goals:

Incremental improvement or large proportional reduction in total cost?

Characterize cost-quality trade-offs?

Empirical Issues:

a. Scope of model?
b. “Curse of the Intercept”: May dominate
c. Extent to which one may control or observe *X* in real time
d. Predictive power of model?

2. Extent of generalizability of available cost information

a. Global cost structures (simple dominant factors, consistent with underlying theory)

   - Customary scientific ideal

b. Local cost structures (survey or module specific)

Cf. Ongoing Workshops on Total Survey Error

B. Improved Methods to Optimize Survey Cost Effectiveness

1. Methods to collect and analyze cost information – within and outside accounting system

2. Characterize and quantify linkage among cost, information capacity, and data quality

3. Tools for cost optimization of survey procedures subject to complex and uncertain cost structures (cf. Karr – today)

   Ex: Leaver (2005) – Consumer Price Index

   Ex: Adaptive sampling-based data review?

   Ex: Drill-down data review

   Ex: Elicit priors from field personnel?

4. Optimize overall procedure design, in light of:

a. Uncertain and spotty cost information (Critical question: extent to which we should condition on, or integrate over, components of uncertainty?)

b. Previously absorbed fixed costs

(cf. Lessler, 2006)

c. Constraints on data collection and processing that are often cost-driven (Constraints often also involve a substantial component of uncertainty.)

V. Summary

A. Classical sample design and randomization inference

B. Role of models

C. Components of data quality & risk

D. Previous literature on survey costs

E. Two classes of methodological questions