Effects of Disruptions on Total Survey Error

Alan F. Karr
karr@niss.org

National Institute of Statistical Sciences
Research Triangle Park, NC 27709 USA

ITSEW 2009, Tällberg
June 15, 2009

NISS
The Statistics Community Serving the Nation

Two Linked But Distinct Problems

SD Statistical and cost issues arising from disruptions to data collection process and infrastructure

SDC Statistical and cost issues associated with implementation of "special data collections" to assess the scope and consequences of national emergencies, or to evaluate the effectiveness of responses

Members

Alan Karr NISS
Myron Katzoff US National Center for Health Statistics
Meena Khare US National Center for Health Statistics
Saki Kinney NISS
Daniel Nussbaum Naval Postgraduate School

Exemplars

Hurricane Ike Natural disaster, broad but sub-national geographical extent, substantial population displacement, few "contagion" effects, major disruption to communication and transportation infrastructure

Flu Pandemic National scale, major "contagion" effects, relatively little dislocation, little disruption to physical infrastructure, significant disruption to data collection infrastructure

9/11 Major effects of one sort on a limited geographical scale, major and different effects on a national scale, complex dynamics

Dimensions

Event Dimensions

- Duration
- Geographical extent
- Severity
- Infrastructural consequences

Problem Dimensions

- Sector (supply = health care providers or demand = health care recipients)
- Direct vs. indirect effects (direct = immediate consequences of the event, e.g., injury, contamination of food or water; indirect = exacerbation of chronic conditions)
- Detection vs. estimation
- Description vs. intervention

The Big Picture

Very abstract Problem = *cost–data quality tradeoffs driven by decision maker needs*

Less abstract, v.1 For **SD**

- Costs arise from identifying, locating and contacting sample units
- Principal data quality consequences are uncollected data and flawed responses

For **SDC**

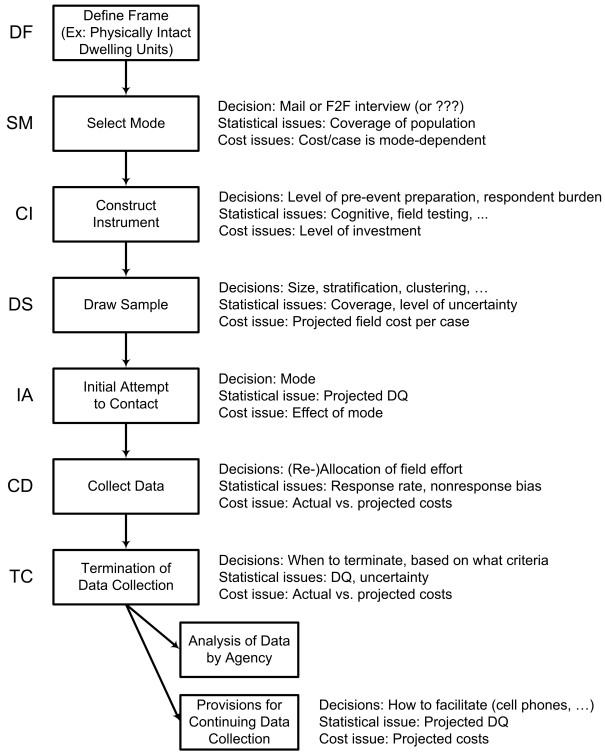
- Costs run full gamut (frame to analysis)
- Data quality consequences run full gamut

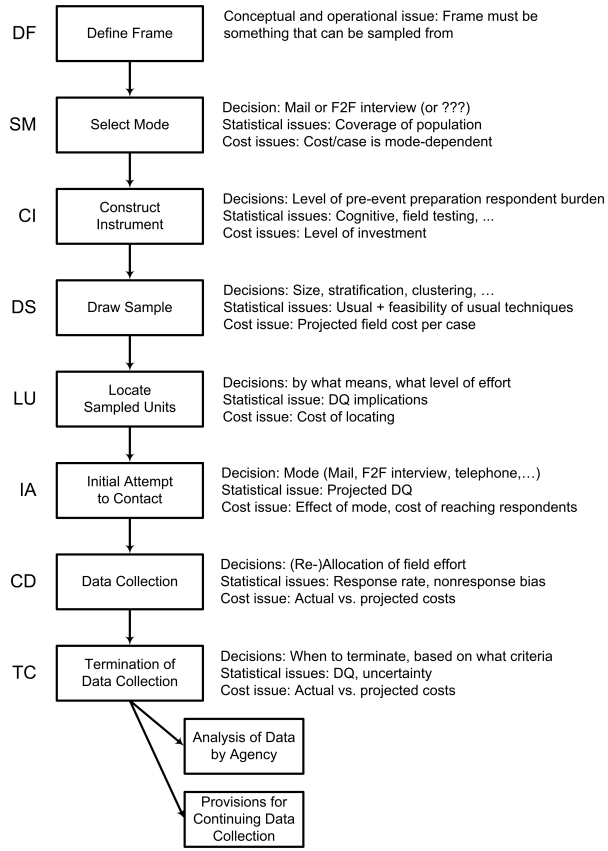
The problem Is this all too abstract to be useful?

Important Distinctions

- Directly impacted region vs. elsewhere
- Remaining population vs. displaced population (not a clean distinction)

Remaining Population on Left—Displaced on Right





Key Data Quality Effects

Faulty frame, not constituting the population of interest

Instrument deficiencies, e.g., unreasonable burden on stressed respondents, cognitive disconnects

Nonresponse resulting from inability to contact, refusal, . . .

Nonresponse bias not handleable by conventional methods

Flawed responses, stemming from lying/exaggeration, poor memory, subjects' being asked to make judgements that they are not qualified to make, . . .

Lack of timeliness, which is especially problematic in a setting of health assessment

And ?????

Two Models

Model 1

$$\text{Total cost} = \text{Fixed cost} + \text{Sample size} \times \text{Variable cost per data case}$$

Model 2

$$\begin{aligned} \text{Total cost}(S) &= \text{Fixed cost} + f_0(S, \text{Other factors}) \\ &\quad + C_{\text{case}}(\text{Other factors}) \times S, \end{aligned}$$

where $f_0(\cdot, \text{Other factors})$ is a step function

Issue: Dependence on scenario characteristics

Distance-Based Formulation

Responses HD = Health Determinants and HO = Health Outcomes[1]

Model for Responses With $r =$ distance from "center" of event:

$$\begin{aligned} \mathrm{HD}(r) &= \mathrm{HD}(0) - b_{\mathrm{D}} e^{-\alpha_{\mathrm{D}} r} + \varepsilon \\ \mathrm{HO}(r) &= \mathrm{HO}(0) - b_{\mathrm{O}} e^{-\alpha_{\mathrm{O}} r} + \varepsilon \end{aligned}$$

Cost Model $c(r) = c_0 + c_1 e^{-\beta r}$

DQ Model

$$\begin{aligned} P\{\text{Response at } r\} &= p_1 + (p_0 - p_1)(1 - e^{-\gamma r}) \\ \text{Reported value at } r &= \text{True value} + e^{-\delta r}\varepsilon \end{aligned}$$

[1] *Wisconsin County Health Rankings 2007*

Then Want to Construct

Survey instrument designed to collect the data necessary to calculate the indices from sampled units

Sample Setting aside frame and related issues, a sample would consist of units $U_1, \ldots, U_N$ at distances $r_1, \ldots, r_N$

Cost function $C(U_1, \ldots, U_N) = \sum_{i=1}^{N} c(r_i)$ (separability problematic)

Value function $V(U_1, \ldots, U_N)$ quantifying value of sample units in fitting the models. Prime example: reduction in uncertainty.

With these, informed tradeoffs are possible

Challenges

- How to do cost-quality tradeoffs at the error source level
- How to get credible (para)data on costs
- How to close the gap with decision makers