survey statistician’s perspective

Frauke Kreuter

JPSM – Uni Mannheim – IAB

@fraukolos
Data Collection Design
• What problem does DP solve at the recruitment stage?
• How do we deal with error prone survey answers in DP?
• Can we afford our data collection if we design for DP?

Data Analysis
• Can we still work the way we are used to with DP data?
• Do we risk distorting the benchmark?

Research Community and Replication
INNOVATIONS IN FEDERAL STATISTICS

Combining Data Sources While Protecting Privacy

FEDERAL STATISTICS, MULTIPLE DATA SOURCES, AND PRIVACY PROTECTION

Next Steps
Data Collection I

Recruitment and answering sensitive questions
Coutts & Jann [2011, SMR]

Randomized Response Techniques are problematic because of
- limited trust
- high variance due to false negative tendency (especially for more sensitive questions)

Kirchner [2015]

No improvement of reporting accuracy with RRT compared to direct questioning (using administrative data for benchmark validation)

Caution shared by others [e.g. Holbrook and Krosnik 2010; Coutts et al. 2011; Wolter and Preisendoerfer 2013; Hoeglinger, Jann, and Diekmann 2014] though not all [e.g. Blair, Imai, Zhou 2015]
Reported Sensitivity of Survey Questions by True State and Mode of Data Collection [Kreuter, Presser, Tourangeau 2008, POQ]

Survey of UMD alumni n=1.003 randomized to modes
### Table 3.9
Types of Offender Behaviors Indicating Coercion/Lack of Consent for Past-Year Non-Penetrative Sexual Assaults, by Gender

<table>
<thead>
<tr>
<th>Question</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>They continued even when you told them or showed them that you were unwilling</td>
<td>60.75% (50.44–70.39)</td>
<td>54.15% (50.26–58.01)</td>
</tr>
<tr>
<td>They used physical force to make you comply</td>
<td>13.96% (8.08–21.88)</td>
<td>24.04% (20.63–27.72)</td>
</tr>
<tr>
<td>They physically injured you</td>
<td>5.02% (1.92–10.44)</td>
<td>4.59% (3.02–6.67)</td>
</tr>
<tr>
<td>They threatened to physically hurt you (or someone else)</td>
<td>7.94% (3.29–15.58)</td>
<td>4.69% (3.17–6.65)</td>
</tr>
<tr>
<td>They threatened you (or someone else) in some other way</td>
<td>15.52% (9.10–24.04)</td>
<td>20.36% (17.10–23.94)</td>
</tr>
<tr>
<td>They did it when you were passed out, asleep, or unconscious</td>
<td>7.12% (1.05–22.09)</td>
<td>11.64% (9.02–14.70)</td>
</tr>
<tr>
<td>They did it when you were so drunk, high, or drugged that you could not understand what was happening or could not show them that you were unwilling</td>
<td>10.12% (3.02–23.23)</td>
<td>15.61% (12.66–18.94)</td>
</tr>
</tbody>
</table>
Reported Believe – Data are Kept Confidential in the Federal Statistical System [Childs, Eggelston, Fobia 2018, BigSurv]

* Change in instruments coincided with a 4.8% decrease in reported belief.

Census data collection via Gallup Tracking Poll
Your participation is vital to our effort. Domestic terrorism preparedness transcends any single level of government, including the Federal government. It is a national issue that can only be effectively addressed through close cooperation at all levels—Federal, state, and local. The work of this Panel concerns nothing less than the security of our nation, the protection of our citizens’ civil liberties, and the ideals of our democratic society.

Your organization has been randomly selected to represent «ORG_TYPE_TEXT» throughout the United States. The survey is being

“The estimates will be the same with or without you in the data”.
Examples of open research questions ...

• How do we communicate the method?
• How do we establish sufficient trust?
Data Collection II

Random Noise and Missing Answers
**Measurement**
- Construct $\mu_i$
- Measurement $Y_i$
- Response $y_i$
- Edited Response $y_{ip}$

**Representation**
- Target Population $\bar{Y}$
- Sampling Frame $\bar{Y}_C$
- Sample $\bar{y}_a$
- Respondents $\bar{y}_r$
- Adjustments $\bar{y}_{rw}$

$\bar{y}_{prw}$ Survey Statistic
**Measurement**

Construct $\mu_i$

**Measurement** $Y_i$

Response $y_i$

Edited Response $y_{ip}$

Survey Statistic $\bar{y}_{prw}$

police reported crime

“During the last 6 months, did you call the police to report something that happened to you that you thought was a crime?”
Measurement

Construct $\mu_i$

Validity

Measurement $Y_i$

Response $y_i$

Edited Response $y_{ip}$

Survey Statistic $\overline{y}_{prw}$

Extent to which measure reflects construct over all possible trials
“During the last 6 months, did you call the police to report something that happened to you that you thought was a crime?”

“I called them about my neighbor’s car hitting my mailbox.” – Interviewer

police reported crime

Survey Statistic
**Measurement**

Construct $\mu_i$

Measurement $Y_i$

Measurement Error

Response $y_i$

Edited Response $y_{ip}$

**Survey Statistic** $\bar{y}_{prw}$

Departure of response to a measure from true value of measure for a respondent

Classical error model (Lord & Novick 1969)

$$y_i = Y_i + \varepsilon_i, \text{ with } \varepsilon \sim N(0, \sigma)$$

perfect measurement means $\sigma = 0$.

[Tourangeau, Rips, Rasinski 2000; Krosnick & Presser 2010; Biemer et al. 2013; Vannette & Krosnick 2018]
**Measurement**

- Construct $\mu_i$
- Measurement $Y_i$
- Response $y_i$
- Edited Response $y_{ip}$

**Processing Error**

Departure of data from true value due to coding or editing of response

$$y_{ip} - y_i$$

**Survey Statistic**

$\bar{y}_{prw}$
Examples of open research questions...

- What if *data linkage* is desired to reduce respondent burden and shorten the interview? [Sakshaug, Kreuter 2012, SRM]
- Should *error-prone survey answers* be considered fixed for the purpose of a DP definition? [Oberski 2019]
- Are we taking agency away from respondents by creating *values for missing data* or when we “improve” *values for misreports*?
Data Collection III

Design Decisions

Cost: $$ and time
US HH population, 12 years and over

Persons linked to housing units at listed address in sample areas

\[ \bar{y}_{prw} \quad \text{Survey Statistic} \]

\[ \bar{y}_{rw} \quad \text{Adjustments} \]

\[ \bar{y}_r \quad \text{Respondents} \]

\[ \bar{y}_a \quad \text{Sample} \]

\[ \bar{y}_C \quad \text{Sampling Frame} \]

\[ \bar{Y} \quad \text{Target Population} \]
Difference between covered and noncovered population

\[
\bar{Y}_c = \bar{Y}_N + \frac{U}{N} (\bar{Y}_C - \bar{Y}_U)
\]
US HH population, 12 years and over

Persons linked to housing units at listed address in sample areas

Multistage area probability sample of persons in sample HH

**Representation**

$\bar{Y}$ Target Population

$\bar{Y}_C$ Sampling Frame

$\bar{Y}_a$ Sample

$\bar{Y}_r$ Respondents

$\bar{Y}_{rw}$ Adjustments

$\bar{y}_{prw}$ Survey Statistic
Sampling error vs. sampling variance

Representation

\[ \bar{Y} \quad \text{Target Population} \]

\[ \bar{Y}_C \quad \text{Sampling Frame} \]

Sample

\[ \bar{y}_a \quad \text{Sample} \]

\[ \bar{Y}_r \quad \text{Respondents} \]

Adjustments

\[ \bar{y}_{rw} \quad \text{Adjustments} \]

Survey Statistic

\[ \bar{y}_{prw} \]
Target population:
Noninstitutionalized adults in contiguous U.S.

500 adults at random every month

How would you do that? What is the probability?
Many important surveys are not simple random samples but have one or more of these characteristics:

- Multistage
- Stratified
- Clustered
- Sampled with unequal probabilities
STRATIFIED SAMPLING

Population

Strata

Sample

Focus on reducing variance

CLUSTER SAMPLING

Population

Cluster

Sample

Focus on reducing costs

Effect on Sample Size – Cost (rough estimates)

Mail: $50 per case
Phone: $250 per case
Face-to-face: $1,000 per case

Complex sample design inflates SE by a design effect of 1.4

Mail: several weeks
Phone: couple of months
Face-to-face: several months

Kreuter, Valliant (2007), PISA Test scores means and confidence intervals with and without complex sample design
US HH population, 12 years and over

Persons linked to housing units at listed address in sample areas

Multistage area probability sample of persons in sample HH

Persons providing answers to interviewer-administered questionnaire

**Representation**

$\bar{Y}$ Target Population

$\bar{Y}_C$ Sampling Frame

$\bar{Y}_a$ Sample

$\bar{Y}_r$ Respondents

$\bar{Y}_{rw}$ Adjustments

$\bar{Y}_{prw}$ Survey Statistic
Values of statistic computed based only on respondent data differ from those based on entire sample

\[ B(\bar{y}_R) \approx \left( \frac{\sigma_{y\rho}}{\bar{\rho}} \right) \]
Examples of open research questions ...

- Can we still get design-unbiased estimates with DP?
- What sample size do we need to maintain desired precision with DP? Can we afford that?
- What do *interviewers* *longitudinal data households* do to the DP analysis?
- Can we justify the costs if we limit access via privacy budget?
Data Analysis I

Research Questions, Workflow, Informative outliers
Typical Research Questions

“What is the relation between experiences of discrimination and the risk of PTSD among African American adults?”
Sibrava et al. 2019, American Psychologist

“How does being integrated with poor students affect the social behaviors and academic outcomes of rich students?”
Rao, 2019, American Economic Review

“Does high immigration increase inter-ethnic tension?”
Weber, 2018, European Sociological Review
Typical methods

- Psychology
  - A/B experiments
  - ANOVA
- Economics
  - Linear regression
  - Time series analysis
- Sociology
  - Multilevel (random effects) linear regression models
- Demography
  - Time to event (survival) models

Often data collected for other purposes (research or gov. statistics)

- ICPSR data archive in the U.S.
- Essex - UK data archive
- GESIS – data archive, Germany
1. Focus on coefficients (parameters) and standard errors (statistical sampling variation)

2. Model exploration is part of the analysis

3. Mix of categorical and continuous variables

4. Complex data structures

<table>
<thead>
<tr>
<th>Parameter variance</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between countries</td>
<td>P.E. 1.60 S.E. 0.50</td>
<td>P.E. 1.58 S.E. 0.50</td>
<td>P.E. 1.56 S.E. 0.49</td>
<td>P.E. 1.88 S.E. 0.63</td>
</tr>
<tr>
<td>Within countries</td>
<td>P.E. 14.67 S.E. 0.11</td>
<td>P.E. 14.66 S.E. 0.11</td>
<td>P.E. 14.64 S.E. 0.11</td>
<td>P.E. 13.20 S.E. 0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance components</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic minority</td>
<td>P.E. 0.37 S.E. 0.19</td>
<td>P.E. 0.36 S.E. 0.19</td>
<td>P.E. 0.26 S.E. 0.15</td>
<td>P.E. 0.21 S.E. 0.14</td>
</tr>
<tr>
<td>First generation</td>
<td>P.E. 0.11 S.E. 0.09</td>
<td>P.E. 0.02 S.E. 0.08</td>
<td>P.E. 0.01 S.E. 0.08</td>
<td>P.E. 0.00 S.E. 0.08</td>
</tr>
<tr>
<td>Gender</td>
<td>P.E. 0.10 S.E. 0.04</td>
<td>P.E. 0.10 S.E. 0.04</td>
<td>P.E. 0.10 S.E. 0.04</td>
<td>P.E. 0.07 S.E. 0.03</td>
</tr>
<tr>
<td>Partner</td>
<td>P.E. 0.13 S.E. 0.05</td>
<td>P.E. 0.13 S.E. 0.05</td>
<td>P.E. 0.13 S.E. 0.05</td>
<td>P.E. 0.08 S.E. 0.04</td>
</tr>
<tr>
<td>21–35 years</td>
<td>P.E. 0.26 S.E. 0.10</td>
<td>P.E. 0.25 S.E. 0.10</td>
<td>P.E. 0.25 S.E. 0.10</td>
<td>P.E. 0.12 S.E. 0.06</td>
</tr>
<tr>
<td>50–64 years</td>
<td>P.E. 0.13 S.E. 0.06</td>
<td>P.E. 0.13 S.E. 0.06</td>
<td>P.E. 0.13 S.E. 0.06</td>
<td>P.E. 0.04 S.E. 0.03</td>
</tr>
<tr>
<td>65 years or older</td>
<td>P.E. 0.80 S.E. 0.26</td>
<td>P.E. 0.79 S.E. 0.26</td>
<td>P.E. 0.79 S.E. 0.26</td>
<td>P.E. 0.28 S.E. 0.12</td>
</tr>
<tr>
<td>Outside Europe</td>
<td>P.E. 0.28 S.E. 0.18</td>
<td>P.E. 0.26 S.E. 0.18</td>
<td>P.E. 0.26 S.E. 0.18</td>
<td>P.E. 0.15 S.E. 0.14</td>
</tr>
<tr>
<td>Ethnic discrimination</td>
<td>P.E. 0.18 S.E. 0.18</td>
<td>P.E. 1.01 S.E. 0.39</td>
<td>P.E. 0.11 S.E. 0.05</td>
<td>P.E. 0.05 S.E. 0.14</td>
</tr>
<tr>
<td>Finding it very difficult</td>
<td>P.E. 0.37 S.E. 0.19</td>
<td>P.E. 0.28 S.E. 0.22</td>
<td>P.E. 0.23 S.E. 0.10</td>
<td>P.E. 0.04 S.E. 0.04</td>
</tr>
<tr>
<td>Finding it difficult</td>
<td>P.E. 0.05 S.E. 0.14</td>
<td>P.E. 0.37 S.E. 0.19</td>
<td>P.E. 0.28 S.E. 0.22</td>
<td>P.E. 0.04 S.E. 0.04</td>
</tr>
<tr>
<td>Student</td>
<td>P.E. 0.04 S.E. 0.04</td>
<td>P.E. 0.00 S.E. 0.00</td>
<td>P.E. 0.00 S.E. 0.00</td>
<td>P.E. 0.00 S.E. 0.00</td>
</tr>
</tbody>
</table>

Source: European Social Survey, 3rd round, own calculations

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, Wald Z test
Examples of open research questions ...

• Can microdata be used for iterative analysis processes?
• Can <INSERT FAVORITE METHOD> be used?
• Can we learn enough about coefficients?
• Can outliers still lead to insights?
• What about responsive design / predictive (policy) intervention applications?
Variance Estimation

**Exact formulas**
- Only possible for ‘linear’ estimators

**Linearization (Taylor series)**
- Used for ‘nonlinear’ estimators

**Replication**
- Applies to linear and nonlinear estimators

In practice often ignored! [West, Sakshaug, Aurelien 2016, PlusOne]
Replicate Weights

General procedure
1. Divide full sample into subsamples (replicates)
2. Repeat weight computation for each subsample
   - base weight
   - adjustment for subsampling
   - nonresponse adjustment
   - calibration
3. Each sample element has a full sample weight and a series of replicate weights
4. Uses receive (large) file with data and all weights
GREG – Nonresponse Adjustment

Categorical and continuous variables can be used
Estimator of total is

$$\hat{T}_{y, GREG} = \hat{t}_y + (t_x - \hat{t}_x)^T \hat{B}$$

$\hat{t}_y$ is estimated total using input weights (base or NR-adjusted)
$t_x$ is vector of pop totals of $x$'s
$\hat{t}_x$ is vector of estimated pop totals of $x$'s using input weights
$\hat{B}$ is (input weighted) slope of $y$ on $x$

Underlying model for GREG is

$$y_i = x_i^T \beta + \epsilon_i, \epsilon_i \sim (0, \nu_i)$$

$$\hat{T}_{y, GREG} = \sum_{i \in s} \left[ 1 + (t_x - \hat{t}_x)^T (X^T D V^{-1} X)^{-1} x_i / \nu_i \right] d_i y_i$$

$$= \sum_{i \in s} g_i d_i y_i$$
Examples of open research questions ...

• What about replicate weights in DP?
• What if population benchmarks for nonresponse adjustments are DP? [Dever & Valliant 2016, JSSM; Lee & Valliant 2015, JOS; Liao & Valliant 2012]
• How do get fixed privacy guarantee with need for valid variance estimation (multiple synthetic data sets)
Example: IAB – SMART [Kreuter et al. 2019]

Sample of households with at least one welfare benefit recipient (at reference date)
- Refreshed annually
- Surveyed annually

Random household sample of resident population
- Refreshed annually
- Surveyed annually
Examples of open research questions ...

- Focus on input or output control?
- How to automate either or both?
- How to automate the rich context documentation?
- How to scale the use?