Module 6: Data quality monitoring
Video 4 of 6: Monitoring substantive data

Sharan Sharma
Possible data sources to monitor interviewers

1. GPS coordinates
2. Reinterviews
3. Listening-in
4. Mystery respondents
5. Substantive data
6. Paradata
7. Audio recordings (CARI)
5. Data-based monitoring: Monitor the actual substantive data

• Do the collected data exhibit *unusual* patterns?
  – Mean (‘average’)
  – Standard deviation (‘spread of the data’)
  – Relationships between variables e.g. regression (useful but uncommonly used in quality control)

• Basic idea : How do different groups compare?
  – “Groups” are often interviewers in quality control but could also be supervisors, fieldwork agencies, geographies, time of the day, fieldwork period (trends) etc.

• Some methods based on analyzing substantive data :
A. Descriptive statistics

• How much do data collected by an interviewer deviate from the overall average/expected data?

• Murphy et al. (2004) computed a score for each interviewer based on deviations of substance use rates within age, sex, and ethnicity strata.
  – Response deviations for three known falsifiers were the highest among the overall pool of 10 interviewers.
Verification calls: first 2 non-interviews and first 2 interviews in each quarter + at least 5% of each interviewer’s screenings + at least 15% of each interviewer’s completed interviews

- Any suspicion e.g. missing/refused telephone number → more verification

### Example from National Survey on Drug Use and Health, USA, 2002.

<table>
<thead>
<tr>
<th>Interviewer / Cases</th>
<th>Number of Interviews</th>
<th>Lifetime Use (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cigarettes</td>
</tr>
<tr>
<td>All valid cases</td>
<td>1,188</td>
<td>56.9</td>
</tr>
<tr>
<td>Falsifier 1’s fraudulent cases</td>
<td>92</td>
<td>26.1\textsuperscript{a}</td>
</tr>
<tr>
<td>Falsifier 2’s fraudulent cases</td>
<td>119</td>
<td>48.7</td>
</tr>
<tr>
<td>Falsifier 3’s fraudulent cases</td>
<td>77</td>
<td>29.9\textsuperscript{a}</td>
</tr>
</tbody>
</table>

\textsuperscript{a} rate significantly different from valid case rate at p<.05 level

Source: Murphy et al, 2004
What do you think is going on here?

(Each dot is the average for an interviewer’s workload)
Data entry errors?

Kilograms of rice consumed last month

November 2020 Module 6 – Remote training on Phone Surveys 7
What do you think is going on here?

(Distribution of responses for 10 interviewers, one for each panel)

Kilograms of rice consumed last month
Trying too hard not to be detected?

Kilograms of rice consumed last month
B. Straightlining

• The tendency to use an identical response category for all items in a series of questions e.g. obtaining ten “Agree” responses to a series of 10 opinion questions that have a 5-point response ranging from “Totally agree” to “Totally disagree”

• Phone surveys more susceptible to such behaviors than face-to-face surveys [Holbrook et al (2003)]

• For a battery of binary questions, compute %Yes (or No) for each interviewer. Compare interviewers on this metric.
  – If possible, phrase questions so that one expects a mix of Yes/No
  – Kim et al (2018) summarize different ways to measure straightlining
C. Rounding

“What was your gross income in 2019?”

Responses collected by Interviewer 1 (in $): 40235, 110683, 23568, 15000, 89675, …. 

Responses collected by Interviewer 2 (in $): 40000, 100000, 20000, 15000, 89000, …. 

• Another term used for variables like age is that of “heaping”.
D. Responses to Roster questions

• Obtaining a list of household members and their characteristics is often the first (and critical) component.

• In countries with large household sizes, interviewers may fail to include everyone.
  – Could be due to definitional issues → “My husband works in the city and comes home once in 15 days”. To include him?
  – Could be motivated underreporting → many questionnaires have questions (e.g. education levels) that loop through all household members for each member. Some interviewers may also be tempted to exclude members for whom a special instrument needs to be administered e.g. an ‘eligible woman’ questionnaire.
E. Responses to Screener questions

- Keep a watch on household screening rates
  Tourangeau et al (2012) find that interviewers with higher screener completion rates tended to find fewer households with eligible members.

- At the question level: how many interviewers obtain a negative response to gate questions.
  e.g. Obtaining “No” to “Do you own any farmland?” will skip the entire battery of 25 questions on farmland ownership, leasing, crop cultivation, etc.
F. Benford’s law

- Turns out that the leading digits for many numbers occurring in real life are not uniformly distributed i.e the number 1 (or 2, 3…9) does not occur 11.1% of the time (1/9) but follows the ‘leading digit distribution’.

- Useful since a falsifying interviewer expects a uniform distribution.
Benford’s law…


- They ran this analysis for each interviewer….
<table>
<thead>
<tr>
<th>Leading Digit (d)</th>
<th>CEQ's Nationwide Distribution (n=734,684)</th>
<th>A Typical FR (θ = 10.39) (n=1,143)</th>
<th>An Unusual FR (θ = 102.43) (n=1,132)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.5</td>
<td>31.4</td>
<td>28.9</td>
</tr>
<tr>
<td>2</td>
<td>19.3</td>
<td>19.7</td>
<td>18.0</td>
</tr>
<tr>
<td>3</td>
<td>12.3</td>
<td>11.6</td>
<td>8.1</td>
</tr>
<tr>
<td>4</td>
<td>9.0</td>
<td>9.5</td>
<td>8.5</td>
</tr>
<tr>
<td>5</td>
<td>10.4</td>
<td>8.3</td>
<td>17.2</td>
</tr>
<tr>
<td>6</td>
<td>6.8</td>
<td>6.4</td>
<td>10.5</td>
</tr>
<tr>
<td>7</td>
<td>4.8</td>
<td>4.7</td>
<td>4.2</td>
</tr>
<tr>
<td>8</td>
<td>4.4</td>
<td>5.2</td>
<td>3.2</td>
</tr>
<tr>
<td>9</td>
<td>2.5</td>
<td>3.2</td>
<td>1.3</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

FR → Field representative

Benford’s law…

Some successes in the survey world e.g. Schräpler (2010) finds it effective to spot falsifier interviewers. But not widely used. One reason is that its applicable to specific data:

- Consist only of positive values with a unimodal distribution.
- Have a positive skew.
- Do not contain a built-in maximum.
- Raw data in the sense of not being based on averages or other summaries.
G. Duplicates

• Can occur due to technical reasons e.g. Interviewer tries to upload the data when the internet fails. Uploads again even though went through the first time.

• But can also occur due to office-level falsification
  – Blasius and Thiessen (2012): 36 questions from the World Values Survey (2005 – 2008). One country had 25% duplicate cases! Except for one country: “...for the remaining countries, the only plausible conclusion is that sample sizes were augmented through copy-and-paste procedures”

• Records need not be complete duplicates: Robbins (2015) finds 178 of 1200 interviews to have a match rate of 80%+ in the 3rd wave of the Arab Barometer survey. One specific interviewer had a match rate of 80%+ in 122 of 123 interviews.
H. Field-check tables

- Computes several group-level summaries from incoming collected data

- Typically updated weekly/fortnightly or when some threshold number of interviews (e.g. 200 more interviews) have been completed.
  - Can produce indicators based on only the incremental data and compare it to previous data. Provides a trend of indicators.

- Please see ICF Macro (2009) and Dupuis (2018) for extensive examples.
e.g. Are female ages being artificially pushed below 15 to avoid interviews?

<table>
<thead>
<tr>
<th>Team</th>
<th>Age of women in years (N)</th>
<th>Age ratio (15/14)</th>
<th>Extended age ratio (15+16)/(13+14)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Team 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All teams</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Dupuis et al (2018)
One key issue in many substantive data-based checks: Are the phenomena due to the interviewer or respondent?

• Quick check: assuming sufficient variation in respondent characteristics, summarize straightlining, rounding, etc. measures by each interviewer and rank.

• Control respondent characteristics in statistical analyses…
Adjusting for respondent profiles…

- Non-response: West and Groves (2013) evaluate cooperation rates by interviewers after adjusting for the predicted difficulty of cases.

- Measurement error: Sharma and Elliott (2019) use multilevel models that adjust for respondent characteristics to detect interviewers who might be involved in falsification.
I. Trends

Why is this important?

• Interviewer behaviors change over time [Olson and Smyth (2020)]
  – Also, interviewer behaviors are related to interview length.

• While interviewers become more efficient over-time, fatigue can also set-in. With more experience, interviewer may also resort to tactics like preempting questions.
I. Trends…

Weekly food consumption data used along with known calorie value.

Black -> ok
yellow -> somewhat small/large
Red -> outside bounds

Source: https://github.com/arthur-shaw/ehcvm-rapport-hebdomadaire
Enquête Harmonisée sur les Condition des Vie des Ménages (EHCVM) 2018-19
J. Missing data

Especially useful to check for:

- Questions that are sensitive
- Questions that involve more than usual probing
- Questions involving more work ....
GPS data used to get accurate land measurements. Track missing %

<table>
<thead>
<tr>
<th>Team</th>
<th>Parcels (N)</th>
<th>Measured (%)</th>
<th>Why not measured (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Team 1</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Team 2</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>...</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Team N</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Source: Draft field check table for the 50x30 initiative.
https://www.50x2030.org/
Possible data sources to monitor interviewers

1. GPS coordinates
2. Reinterviews
3. Listening-in
4. Mystery respondents
5. Substantive data
6. Paradata
7. Audio recordings (CARI)
6. Paradata

• Was introduced in Module 4: Auxiliary data collected in a survey that describe the data collection process.

• Survey software (e.g. Survey Solutions, BLAISE) records keystroke events made by the interviewer during the interview and their associated time-stamps. These records are then parsed to be analyzed.

• Paradata will be covered in more detail in Video 5; Kreuter (2013) has a lot of detail.
Uses of paradata for different components of the Total Survey Error framework.

Source: Kreuter and Casas-Cordero (2010)
Possible data sources to monitor interviewers

1. GPS coordinates
2. Reinterviews
3. Listening-in
4. Mystery respondents
5. Substantive data
6. Paradata
7. Audio recordings (CARI)
7. Computer audio-recorded interviewing (CARI)

- Modern computer-assisted interviews can digitally record interviews or segments of interviews.

- Unobtrusive.

- Substantive data/paradata can be used to select cases to listen to.

- Can tell us about the live interaction between the interviewer and respondent.

- More on using CARI will be covered in Video 5.
END OF VIDEO 4