Probabilistic Cleaning for Messy Data

M. Rita Thissen
RTI International

FedCASIC
May 3-4, 2016
Suitland, MD

Data Cleaning

- Social science studies, especially surveys, produce “noisy” data with varying amounts of item-level error.
  - Respondents may refuse to answer, skip questions, not know the answer, or answer incorrectly or untruthfully.
  - Not understanding the question or not reading carefully (self-administered surveys)

- Confidence is higher for analytic results when key data points are less noisy.

- Data cleaning processes examine the raw response data and then recode fields to produce a clean file for analyses.

Traditional Deterministic Methods

- Traditional data cleaning follows a deterministic path of the form:

  If <some set of criteria> then <recode to predefined values>

- Appropriate usage
  - Range checks
  - Valid value checks
  - Single-field recoding, such as missing value codes
  - Multiple-field recoding where the criteria have a low level of missing data
  - Aggregation and predefined categorization
- Benefit
  - Very predictable results
  - Straightforward to apply

Probabilistic Methods

- Probabilistic data cleaning follows a different form
  - Step 1: Use a set of raw or cleaned values to compute a score
  - Step 2: Select cut-point(s) based on the score distribution
  - Step 3: Recode the raw data based on the cut-point(s)
- Appropriate usage
  - Inconsistent or noisy data for key field(s)
  - Moderate to high levels of missingness in the criteria fields
  - Availability of indirect information, such as related questionnaire items
- Benefit
  - Recovers cleaner information from noisy data
  - Increases confidence in results
  - May support subsequent activities such as drawing future samples

Putting Theory into Practice

The information in this presentation comes from

- A health survey of individuals
- Conducted in the US
- Self-administered in multiple languages
- Web and paper options
- Just under 20,000 respondents in each mode
- Respondent may be a proxy, not the subject him/herself
- Key analytic question: Is the subject alive or deceased?

Survey name, location and sponsor must remain confidential.

Key Analytic Item: Was the Subject Alive or Deceased?

- Survey questions seem straightforward
    1. Are you the subject?* (yes, no)
    2. If not, why?* (pick one from a list)
        - Several response options indicating that a living subject (e.g. language barrier)
        - The subject is deceased
        - Other? Specify (free text field)
- Response data fall into three categories: clearly alive, clearly deceased and uncertain
- Deterministic coding leaves over 5% uncertain
- Information is needed for analysis and for drawing samples for future surveys

** Questions were rephrased for simplicity and confidentiality.*

Responses on the Paper Form

- Are you the subject? (yes, no)
  - Yes: 96%
  - No: 1%
  - Missing: 3%
- If not, why? (pick list)
  - Response indicates the subject is living (e.g. “language barrier”): 1% (n=243)
  - Deceased: 0.1% (n=13)
  - Other-specify: 2% (n=383)
  - Missing: 97%
- If the subject is deceased
  - Date of death (provided a year, at least): 0.2% (n=32)
  - Location of death (excluding “NA”, “not applicable”, etc.): 0.3% (n=56)
- If deceased, skip the remainder of the survey

Responses on the Web Form

- Are you the subject? (yes, no)
  - Yes: 97%
  - No: 1%
  - Missing: 2%
- If not, why? (pick list shown only to those who said "No, not the subject")
  - Response indicates the subject is living: 0.2% (n=33)
  - Deceased: 0.02% (n=4)
  - Other, specify: n= 0.3% (n=75)
  - Missing (mostly logical skips): 99%
- If the subject is deceased (shown only to those who said "deceased")
  - Year of death: provided for all 4
  - Location of death: provided for 1 of 4
- If deceased, skip logic terminated the survey

What’s the Problem? Data from Paper Forms

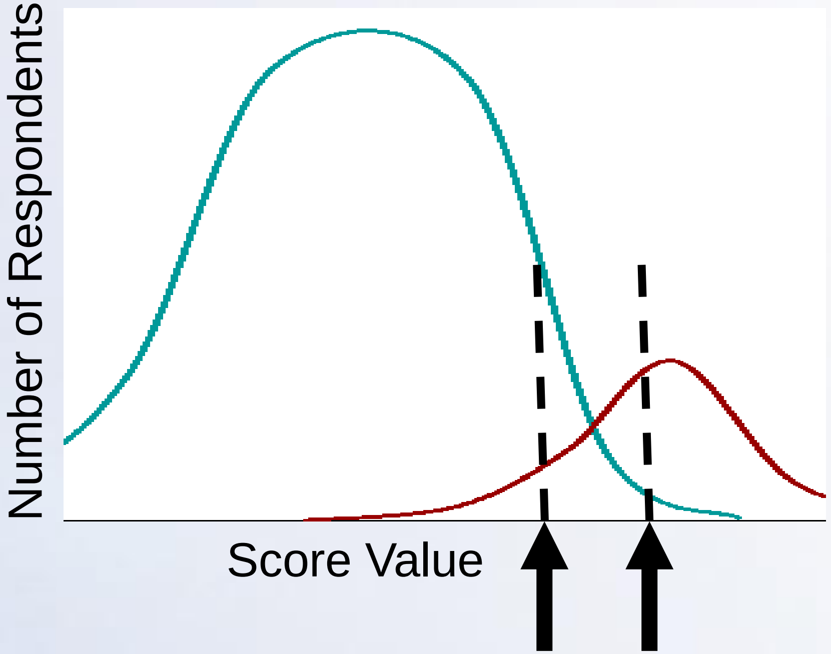
- High levels of inconsistency between the first two questions
  - More death dates and death locations than responses of “deceased”
  - Are you the subject? “Yes”. If not, why? “My boyfriend” “Already done” “He is deceased”
  - Are you the subject? “No”. If not, why? “Self”
- Poor compliance with skip rules on paper forms
  - Presence of data in follow-up questions did not match responses to gate questions well
- Issues with year of death provided by respondents
  - Prior to or same as subject’s birth date
  - Same as or later than the survey completion date
  - Impossible numbers, partial or missing digits
  - Date prior to drawing survey sample

Probabilistic Recoding Steps 1 and 2

Step 1: Choose a scoring method.

Step 2: Look at the distribution of scores to choose cut-points.

- In the area of overlap, you will always be uncertain of the "truth"
- Adjust cut-points for the degree of certainty desired
  - By inspection
  - Mathematically



This illustration is conceptual only, not drawn from actual data.

Is the Subject Alive? 48 Ways to Say Yes, No or Maybe

- Are you the subject?
  - Three possibilities: Yes, no, or missing
- If not the subject, why?
  - Four possibilities: “Living” reason, “Deceased”, Other-specify, or missing
- Year of death
  - Two possibilities: Valid (Not the survey date, not earlier than the birth date, not in the future or impossible) or invalid
- The presence of responses to later subjective personal questions
  - Two possibilities: Responses present or absent
  - Absent might reflect a breakoff, and answers could be given by a proxy.
- Possible combinations: 3 x 4 x 2 x 2 = 48

The Chosen Probabilistic Scoring Algorithm

- Living Score: Indications that the subject is alive

  +2 points for answering “Yes” to “Are you the subject?”

  +2 points for choosing the “living” options to “If not, why?”

  +2 points for the presence of responses to subsequent highly personal items

- Deceased Score: Indications that the subject is deceased

  +2 points for selecting “deceased” in response to “If not, why?”

  +2 points for a valid death date

  +1 point for lack of any subsequent responses in the survey

- Overall Score: Living – Deceased

- Why not give -2 points for terminating after the deceased questions?
  - It may have just been a breakoff.

Probabilistic Approach: Programming

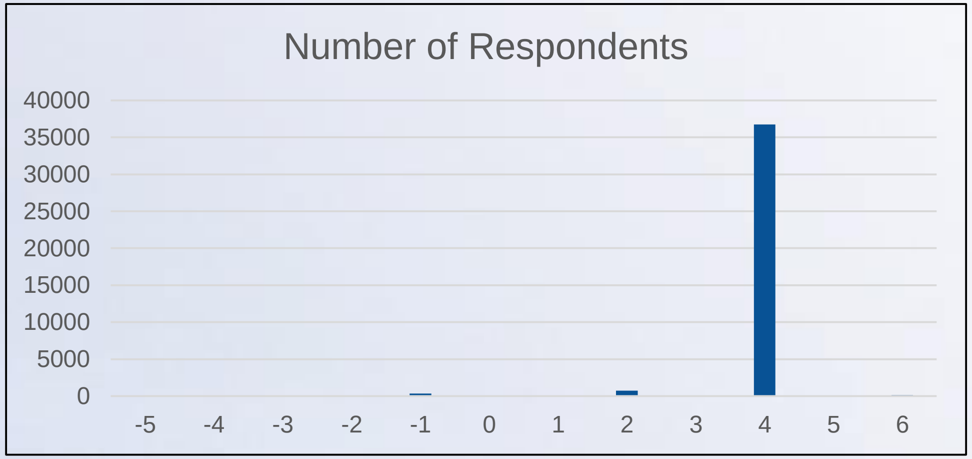
- Step 1. Compute the score

- Step 3. Compare score with the cut-points
  - If <below the lower cut-point> then <follow rule for low scores>
  - Else if <between upper and lower cut-points>
    then <follow rule for mid-range scores>
  - Else if <above upper cut-point> then <follow rule for high scores>

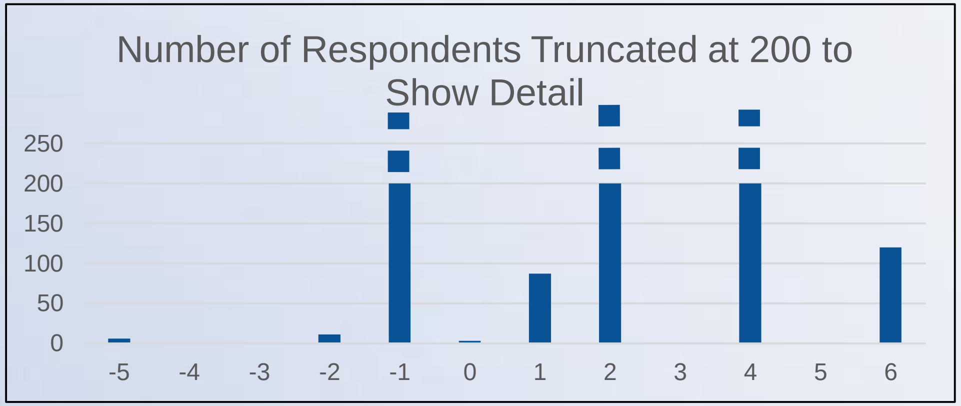
- Step 2, choosing the cut-points, is only done once for the entire dataset.

Score = Living - Deceased

- Subjects were mostly living
  - Over 37,000 alive
  - Hundreds potentially deceased
- Scores computed as integers
  - Range from -5 to +6
  - Scores of 3 and 5 were not possible
- Cut-points chosen:
  - 2 or higher, assumed alive
  - -2 or lower, assumed deceased
  - -1 to +1, uncertain



-5 -4 -3 -2 -1 0 1 2 3 4 5 6



Recoding Results

- Deterministic recoding “cleaned up” death dates and locations
- Probabilistic recoding reduced uncertainty among those who skipped one or more of the identity, proxy and living/deceased questions
- Probabilistic method reduced uncertainty by 4-5%
- Less effort needed to verify deaths or confirm that subjects are alive before drawing other samples

Paper Surveys Only	Percent Alive	Percent Unknown	Percent Deceased
Deterministic	95.7	5.17	0.07
Probabilistic	99.9	0.04	0.07

All Surveys	Percent Alive	Percent Unknown	Percent Deceased
Deterministic	96.6	5.4	0.04
Probabilistic	98.7	1.2	0.04

Confirmation of Approach

- For paper surveys, all of the “uncertain” cases were reviewed
  - Two independent reviewers, one at 80% of data collection, one at 100%
  - Examination of survey pages for marginal notes or other information
  - 100% agreement between human and probabilistic recoding
- Example: This subject was recoded from “uncertain” to living by algorithm
  - Are you the subject? Yes
  - If not, why? “Repetitive and lengthy”
  - Year of death? 2015
  - Location of death? Washington, DC
  - Other responses present in survey? Yes
  - Score = 4 (Living)
- Conclusion: The approach is helpful and appears trustworthy

Thank you!

Thank you for listening!

Questions?

Comments?

For additional information, please contact

Rita Thissen
RTI International, Research Triangle Park, NC 27709 USA
919-485-7728
rthissen@rti.org

Many thanks to my unnamed colleagues on this confidential survey, especially to my data-quality counterpart!