Surveys, Platforms, and Crowdsourcing

Jared Edgerton

Conceptual Questions

Please write three to ten sentence explanations for each of the following questions. **You are only required to answer ONE of the two questions below.**

1. Platform-based surveys (e.g., Prolific, MTurk, convenience pools) can generate measurement error in ways that differ from traditional survey modes. Identify two plausible sources of measurement error or bias in platform data and explain how you would detect each issue using diagnostics in the survey export (e.g., duration, attention checks, missingness patterns, straightlining, duplicate IDs).
2. Missing data is not the same thing as "random noise." Explain the difference between missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) in the context of survey platforms. Give one concrete example of how platform incentives or survey design could create MNAR missingness, and explain one implication for analysis.

Applied Exercises

Use the code in the week's code tutorial and the lecture slides to answer the following questions.

3. **Survey export: load + inspect.** Using the week's survey tutorial script, load a survey export (your own export or the provided toy example).
   - Print the dataset and use `glimpse()` to inspect variable types.
   - Identify at least **five** variables that typically appear in platform exports (e.g., respondent ID, start/end time, duration, platform ID, consent, attention check).
   - Briefly note (1–3 sentences) one potential cleaning issue you see immediately (e.g., missing-value strings, type mismatches, extra header rows).
4. **Cleaning: names, types, and missing values.** Extend the tutorial cleaning steps:
   - Standardize column names (e.g., `clean_names()`).
   - Parse start/end timestamps into a time class (e.g., `ymd_hms()`).
   - Convert at least **two** variables from strings to appropriate types (e.g., age to numeric; duration to minutes).
   - Recode at least **two** platform-style missing-value strings into `NA` (e.g., `""`, `"Prefer not to say"`).
5. **Codebook: document your data.** Create a codebook table (e.g., a `tibble`) with at least **10** rows, including columns:

   `variable` `description` `notes`

   - Include both survey items and platform metadata variables.

- Export the codebook to `outputs/week_codebook.csv`.

6. **Labeling: variable labels + value labels.** Using the `labelled` package:
   - Add variable labels to at least **five** variables (e.g., `var_label()`).
   - Create at least **one** numeric-coded variable with value labels (e.g., party ID coded as 1/2/3 with labels).
   - Print a small excerpt showing the labeled variable(s) so it is clear that labels are attached.
7. **Quality checks: flags + summary.** Implement the following checks (hard-coded thresholds are fine):
   - **Speeding:** flag respondents below a chosen duration threshold (e.g., < 120 seconds).
   - **Attention check:** flag respondents who fail an attention-check item.
   - **Missingness:** compute each respondent's share missing across a set of key variables and flag high-missingness cases.
   - **Straightlining:** flag respondents who give identical responses across multiple grid items.

   Then:
   - Create a one-row summary table of how many respondents are flagged by each rule.
   - Print a table that shows `response_id` (or equivalent) and all flags for every respondent.
8. **Analysis-ready dataset: filter + save + visualize.** Using your quality flags:
   - Create a filtered dataset that excludes at least **non-consent**, **speeders**, **attention-failures**, and **high-missingness** cases.
   - Print the row count **before** and **after** filtering.
   - Save the cleaned dataset to `data_processed/week_survey_clean.csv`.
   - Save at least **two** diagnostic plots (examples: duration histogram; missingness histogram; share flagged by platform) to the `figures/` folder.
9. **Challenge Question (Optional — if you finish early):** Run a simple sensitivity analysis to show how results depend on quality-screening choices.
   - Choose **one** screening threshold to vary (e.g., speeding threshold or missingness threshold).
   - Recompute how many respondents are retained at **three** different threshold values.
   - Make a small table (and optionally a plot) summarizing how the retained sample changes.
   - In 3–6 sentences, explain what this implies about the credibility and robustness of platform survey inferences.