Discussion: Data Privacy

Kristen Olson
University of Nebraska-Lincoln

FESAC Meeting

June 9, 2023

Data Protection is Important

- Maintaining confidentiality of data is critical to integrity of official statistics
  - Information is shared with the expectation that individuals will not be identified
- In household surveys, any individual person has the risk of being identified, but risks are generally not large from tabular data
  - Or at least not greater than the risk of everyday life
- In business surveys, any individual establishment / enterprise has the risk, including in tabular data
  - Especially for small cells (e.g., industry * geography)
  - Even fact of filing must be protected, without detailed financial or other detailed information about an establishment being reported

Data protection vs. data utility

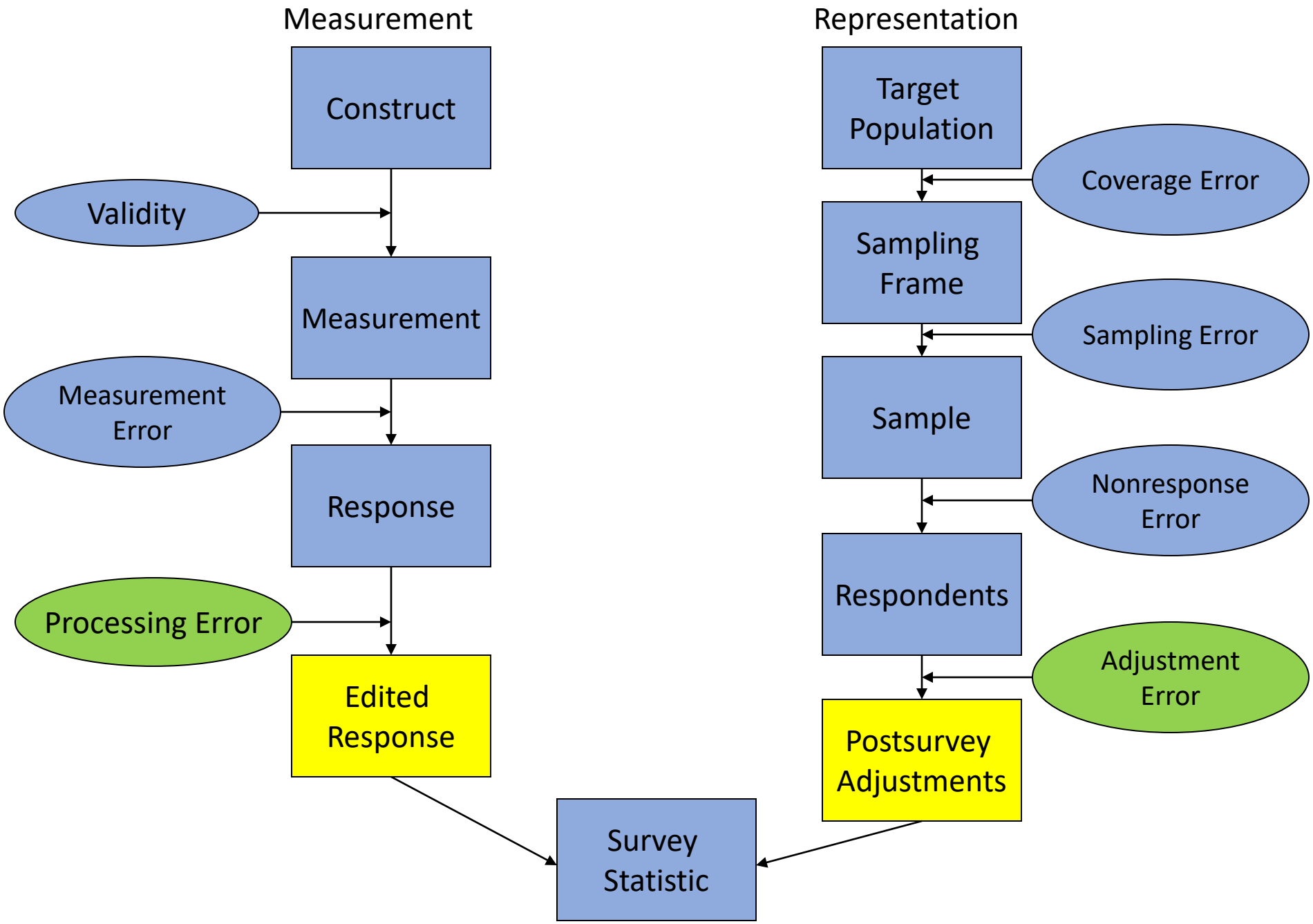
- It is well-known that data utility is maximized with no protective steps taken (release all data) and that data is most strongly protected when not released (protect all data)
  - Clearly, a middle road is sought

- Current approach: Release some data, suppress some data; assume that sampling and aggregation protects the released data
  - May result in a high (e.g., 1:3) risky:nonrisky suppression ratio

- Proposed approaches
  - Release all data, infuse noise across all/most data to protect data across tabulations, perhaps adjust (rake) the noisy data post-noise induction
  - Release all data, infuse noise in some data, do not infuse noise in other data

Noisy versus suppressed data?

All data are noisy, but some data are useful.



Groves, et al. 2004, *Survey Methodology* Figure 2.5

Is noise infusion better than suppression?

- Sometimes, yes.
  - Unlike most measurement errors, noise infusion for privacy purposes is a measurement error with a known distribution.
  - Likely fewer “good” cells suppressed because of “risky” cells.
  - Some data is better than no data.

- Sometimes, no.
  - How important is the decision being made off of noisy data?
  - How influential are the data in making that decision?
  - Does the revealed, but noisy, data yield information that would change a decision made off of the data compared to no information at all?

Other items that would be useful to explore

- How does the new noise infusion affect trend estimates?
  - In what ways do the induced deviations increase/decrease estimates of monthly, quarterly, or annual change? Overall and for what subcategories?
- How does the new noise infusion affect domain estimates?
  - What domains are no longer usable for a given purpose because of the size of the noise?
  - What about shares within a particular domain (noisy denominators and noisy numerators)?
- How sensitive are different estimands? Are individual establishments subject to different privacy constraints for different Y variables?
  - How attenuated are important measures of association?
- How do revisions to estimates inform / affect noise infusion practices and vice versa?
- What do per-record noise infusion processes mean for the timeliness of estimates / publication dates?
- What additional post-survey processing steps – beyond the noise addition itself – need to be added and communicated to users?
  - Additional data quality metrics? Additional table notes?

What level of distortion is too much?

Too much for what purposes?

- The question of too much noise assumes a use case. Different use cases may have different limits to how noisy a data point may be to be useful.
  - Levels vs. change over time vs. relative distributions/shares
  - Point estimates vs. interval estimates
  - Presence vs. absence of industry in an area
  - General trends vs. relative ordering of industries/other categories
  - Statistical tests across groups vs. “eyeball” tests of trends
  - Input into other models (e.g., small area estimation; benchmark totals for other purposes)
- Small denominators will have the largest percent change.
- Raking to marginal totals may “fix” the margins, but obscure important within-table distributional differences.
  - Raking assumes a log-linear model with no interaction effects. The margins may be important.
  - This may be ok, especially when the marginal distributions are critically important to provide to users without distortion. But if the interior cells are important too, raking may make estimates smoother than the real data would suggest.

Insights to provide to users?

Tradeoff between “undoing” the privacy mechanism and transparency

- Users need to feel comfortable that the conclusions or decisions that they are making from the data are “relatively” consistent with what they would make without the noise infusion.

- Technically sophisticated users need to understand enough to be able to conceptually reproduce the work on their own to understand the possible impact (e.g., through simulation) or to explain for their own audience
  - For this group, the more information about the approximate distributions used or bounds on the distribution seems particularly important.
  - Generally, more information is better for a sophisticated user.

- Users need to feel that the information is consistent with their perceptions of the world “on the ground”
  - And they need to be able to understand why the discrepancies exist

EDITOR'S PICK TOPICAL ALERT TOP STORY

Second resident of Nebraska's one-person town just a figment of Census Bureau's imagination

Peter Salter Aug 22, 2021 Updated May 24, 2023 0

Elsie Eiler has been the mayor and only resident of Monowi since her husband, Rudy, died in 2004. She has run the Monowi Tavern for 50 years.
Journal Star file photo

"The U.S. Census Bureau was reporting Monowi's population had exploded by 100% and was now home to two people, according to 2020 results it recently released.

"Well, then someone's been hiding from me, and there's nowhere to live but my house," Elise Eiler said Wednesday. "But if you find out who he is, let me know?"

His name is Noise, and he was created by an algorithm to try to protect Eiler's personal information. Monowi didn't add another resident to its population, but the Census Bureau did."

https://journalstar.com/news/state-and-regional/nebraska/second-resident-of-nebraska-s-one-person-town-just-a-figment-of-census-bureau-s/article_3dade4b8-8736-57af-9270-3a65973f5464.html#tncms-source=login

Assume a wide variety of users require a wide variety of levels of detail

- Communicate information across the widest range of users
  - Flag results by level of confidence/precision of estimates, either because of sampling/ imputation/nonresponse errors and/or because of noise infusion
  - Provide interval estimates accounting for record-level noise infusion
  - Anticipate communication about influence on rates of change and on shares
- If the goal is to release more data in tables, but the noise infusion results in more cells being flagged as problematic (or even not released for quality issues) than without record-level disclosure protection, users may question the goals of the new approaches relative to the old.
- Provide as much information as possible about how and why noise is added to the estimates
  - Not all users will understand all of the details, but some will
  - Some advanced users may be able to use (rough) estimates of the variance of the noise parameters for more sophisticated analyses (e.g., models accounting for measurement error)
- Take advantage of the usability and cognitive testing staff at BLS and Census for pretesting the technical documentation materials for a wide range of users

Fitness for use

- All survey decisions face tradeoffs between accuracy, timeliness, and availability
- Clearly decreasing accuracy for (some) availability
- Perhaps also trading timeliness for (some) availability
- Different users will value accuracy, timeliness, and availability differently for different estimands.

Thank you!

kolson5@unl.edu