How to estimate a population proportion if data are possibly subject to misclassification error? The case of estimating contraceptive prevalence based on self-reported usage.

Chuchu Wei and Leontine Alkema [1]

May 6, 2021

[1]The work was supported by the Bill & Melinda Gates Foundation. Contact: chuchuwei@umass.edu and lalkema@umass.edu.

Background

- Goal: estimate global contraceptive prevalence via a Bayesian hierarchical model with survey data collection
- The model: family planning estimation model (FPEM) (Mark Wheldon et al., 2020)
  - A complex model that conducts national/regional/global levels estimates and forecasts
- Data input: self-reported summary prevalence (e.g. prevalence of modern contracption) from various national surveys
  - Survey sources: DHS, PMA, etc.

Example of the country-level estimation from FPEM

- Estimated trends and forecasts in Ghana/Nepal with FPEM
- (plots of Ghana/Nepal tba)

How is the data used in FPEM?

- FPEM: **data model** + process model
- Data is introduced to FPEM via the data model - the relationship between observed and true prev per country.
- No truth can we observe - an assumption being made for this relationship
- Estimates are informed by the data based on assumptions on the relation between the data and true prevalence

The current data model in FPEM

- FPEM simplified version

  $$logit(y) \sim N(logit(\theta), logit.s^2 + NSE^2)$$

  y observed modern prevalence, θ true modern prevalence, $logit.s$ the logit-transformed sampling error, NSE the non-sampling error

- This data model is based on the fact that obs are obtained in a survey and thus subject to sampling error, and the assumption that typically, data are subject to additional errors called non-sampling error

- No info of non-sampling error available at the time FPEM was developed, hence NSE was estimated for modern use ($NSE \sim 0.1$)

The visualization of the current model assumption

- The relationship between obs and true prev in the domain of (0, 1) along with the 95% CI
  - Assume no sampling error and $NSE = 0.1$ on the logit scale.
- With the current assumption, the uncertainty reaches the maximum at prev $= 0.5$, and decreases as prev move towards $0/1$
- (the plot to be added)

What if we observed some non-sampling error?

- Recap: no paticular assumption of the non-sampling error
- BUT what if there is evidence of non-sampling error? e.g. observations of non-sampling error in the form of misclassification
- Two post-survey studies of DHS (Staveteig, 2017; Staveteig et al., 2018) provides such evidence summarising into sensitivity *se* and specificity *sp*
  - numbers to be added here

The visualization of assumption based on misclassification evidence

- The relationship between obs prev y and true prev θ
  $y = se \cdot \theta + (1 - sp)(1 - \theta)$
- If this is true, the modified data model is
  $y \sim N(se \cdot \theta + (1 - sp)(1 - \theta), s^2)\, T[0, 1]$
  - Assume $s = 0.01, se = 0.8, sp = 1$
  - With such assumption, the current data model increasingly underestimates the true prev if misclassification exists
- (plot tba)

How to do a better estimation

- Recap: current uncertainty assumption in FPEM is not OK with the potential existence of misclassification
- Question: **How to better estimate the uncertainty due to misclassificaiton?**
- Motivation
  - Simple assumption of misclassification in the current assumption
  - Limited evidence of the misclassification from two post-survey studies

Our proposal: a new data model

- We propose a new data model to accomplish the mission
- The aims of the new model
  - $y = \theta$
  - If no misclassification: 95% CI determined by s
  - If misclassification: 95% CI determined by s, se^a, sp^a
- The form of the new model: a standardized likelihood function based on Normal-Laplace distribution (Reed, 2006)
  - $NL(\mu, \sigma, \alpha, \beta)$: the convolution of normal and Laplace distribution
  - The parameters of NL density is based on sampling error s, assumed misclassification se^a, sp^a
  - With a prior of $\theta \sim U(0, 1)$, the standardized likelihood is the posterior of θ given observed info

The visualization of the new data model

- We specify a parameterization routine based on the new model aims to pass the observed y, s and misclassification se^a, sp^a to the new data model
- Assume $s = 0.016, se^a = 0.9, sp^a = 1$
- Unchanged estimates, increased upper bound of 95% CI for estimates with increase of y
- plots to be added

Simulation of the new data model

- We fixed $\theta^{\text{true}} = 0.3, 0.5, 0.7$ and set various groups of true misclassification se, sp and assumed misclassification se^a, sp^a
- True data generation process with sample size n
  true process formula tbd
- Compare posterior estimates from
  - true process model
  - Current assumption in FPEM (logit-normal)
  - The new proposed model (NL likelihood)
- The results shows that (1) the estimates is unchanged in the new NL model; (2) If misclassification exists: the new model did a better job than the old logit-normal model

New data model in FPEM

- New proposal vs the current data model in FPEM
- New model apply to DHS data points only as the evidence is from studies of DHS
- Allow user to define the misclassification $se^a = 0.9, sp^a = 1$
- Observe changes in the estimated trends and projection across countries
    - Red line/shades - new data model; Blue line/shades - old data model; Shapes: data points
- plots to be added

Conclusion

- We propose a new data model to account for increased asymmetric uncertainty associated with potential misclassification errors
- Coverage improvement in the simulation study when data are subject to misclassification
- Changes observed in FPEM implementation