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
University of Massachusetts Amherst[1]

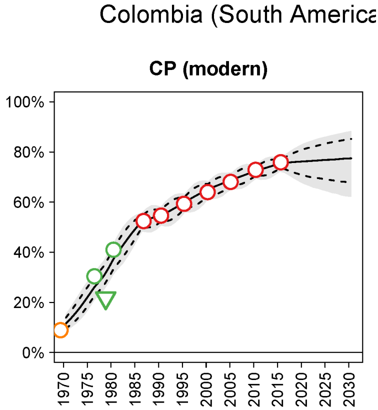
May 6, 2021

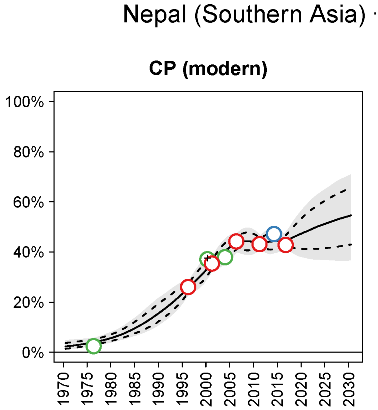
[1] The work was supported by the Bill & Melinda Gates Foundation. Contact: chuchuwei@umass.edu and lalkema@umass.edu.

Introduction

- Motivating question:
  How to estimate contraceptive prevalence using self-reported data collection, i.e. demographic health surveys (DHS)

- Approach: Family Planning Estimation Model (FPEM, Cahill et al., 2018)





How is the data used in FPEM?

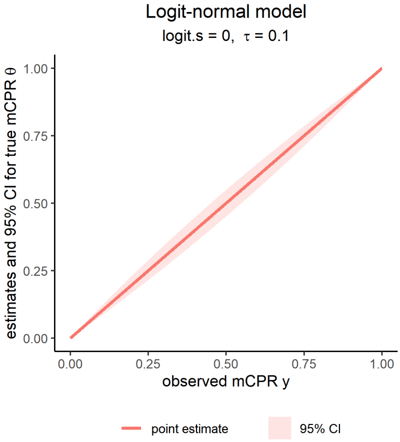
- FPEM: **data model** + process model
  - **data model**: describe how observed data relate to true modern contraceptive prevalence (mCPR)

$$\text{logit}(y) \sim N(\text{logit}(\theta), \text{logit}.s^2 + \tau^2)$$

- Notation
  - y observed mCPR,
  - θ true mCPR,
  - logit.s sampling error,
  - τ non-sampling error
- Visualization: posterior median and 95% CIs based on the posterior:

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

  - prior $\theta \sim U(0, 1)$
  - $p(y|\theta)$ given by logit-normal data model



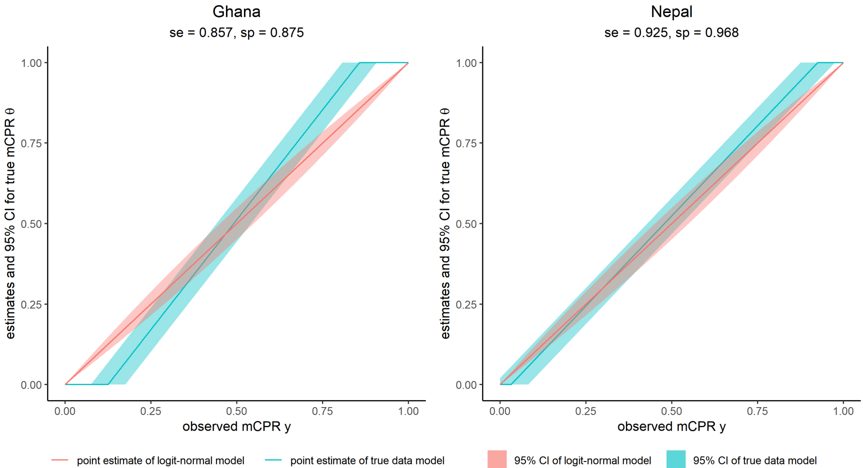
Data on non-sampling error in self-reported modern contraceptive use

- Two post-survey studies of DHS provides data on non-sampling errors in the form of misclassification
  - sensitivity (se) = proportion of modern users who reported themselves as such
  - specificity (sp) = proportion of non-modern users who reported themselves as such
- Findings:
  - 2014 Ghana DHS (Staveteig, 2017) (sample size = 48) $se = 0.857 \pm 0.106, sp = 0.875 \pm 0.094$
  - 2016 Nepal DHS (Staveteig et al., 2018) (sample size = 194) $se = 0.925 \pm 0.037, sp = 0.968 \pm 0.025$

Visualization of the relation between true prevalence θ and self-reported use y in presence of misclassification

True data model when data are subject to misclassification (simplified, s refers to sampling error):

$$y \sim N(se \cdot \theta + (1 - sp)(1 - \theta), s^2)\, T[0, 1],$$



How to estimate a population proportion if data are possibly subject to misclassification error?

- Conclusion so far:
  - Two small post-DHS studies suggest that self-reported mCPR is subject to misclassification
  - The additional uncertainty related to non-sampling error in FPEM does not capture the relationship implied by the studies
- Generalizability problem: only two studies in specific settings
  $\Rightarrow$ do not apply bias-adjustments to self-reported use for all DHS data points based on two studies only
- What we can do:
  Update the data model to better reflect uncertainty associated with potential misclassification errors

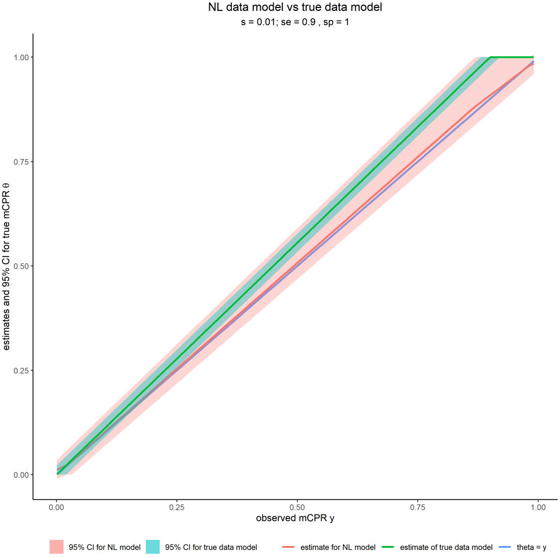
Proposed new data model based on a Normal-Laplace distribution

- Goal: data model to better reflect uncertainty associated with potential misclassification errors
- Aims for the posterior associated with the new data model with assumed sensitivity se^a and specificity sp^a (and $U(0,1)$ prior):
  - $\hat{\theta} = y$
  - If no misclassification: 95% CI determined by sampling error s
  - If misclassification: 95% CI determined by s, se^a, sp^a
- Model specification:
  - based on Normal-Laplace (NL) distribution (Reed, 2006) for the normalized likelihood,
  - posterior under U(0,1) prior given by

    $p(\theta|y) \sim NL(\mu, \sigma, \alpha, \beta)T(0,1)$
  - Parameters fixed to meet aims

Illustration of the new data model

- Visualization: true data model and NL data model
- True relation: $s = 0.01, se = 0.9, sp = 1$
- Settings NL: $s = 0.01, se^a = 0.9, sp^a = 1$, such that $\hat{\theta} = y$ and upper bound of 95% CI increases with y



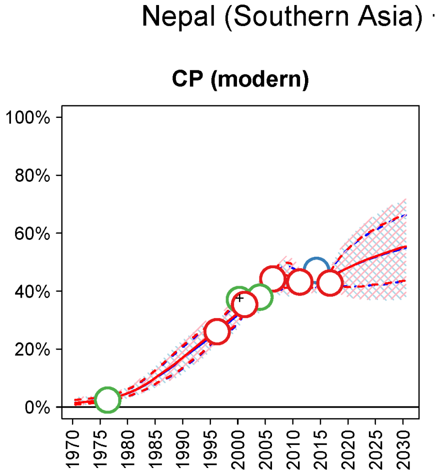
Simulation study

- We fixed $\theta^{\text{true}} = 0.3, 0.5, 0.7$, use different combinations of true misclassification se, sp and assumed misclassification se^a, sp^a, and generate 100 data sets per setting.
- Compare posterior estimates from FPEM logit-normal data model vs proposed NL model
- Findings (as expected):
  1. (bias in) point estimates are comparable between the logit-normal and NL model;
  2. 95% CIs are conservative with coverage exceeding 95%, when assumed misclassification $>$ true misclassification;
  3. NL model improves upon logit-normal model in terms of coverage of 95% CIs when misclassification is present and accounted for.

New data model in FPEM

- NL model used in FPEM, applied to DHS data only
- Two settings:
  - $se^a = 0.9, sp^a = 1$ (Nepal)
  - $se^a = sp^a = 0.9$ (Ghana)
- Findings:

(1) Differences in FPEM fit are more variable as compared to the simulations



Blue = logit-normal,
Red = NL with $se^a = 0.9, sp^a = 1$

New data model in FPEM (ctd)

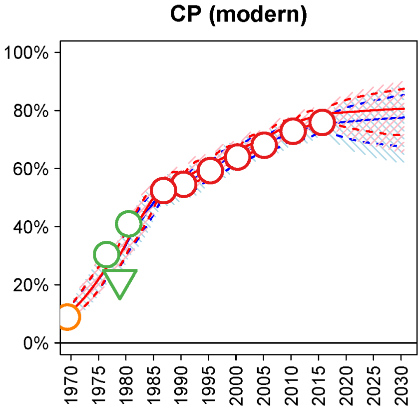
- Findings (ctd):

(1) Differences in FPEM fits are more variable as compared to the simulations

(2) When there are differences,
(i) estimated bounds vary in expected direction based on se^a, sp^a;
(ii) some updating of point estimates as well

- Example: Colombia with $se^a = 0.9, sp^a = 1$

Colombia (South America



Blue = logit-normal,
Red = NL with $se^a = 0.9, sp^a = 1$

Summary

- We investigated how to estimate a population proportion if data are possibly subject to misclassification error, motivated by reported evidence on misclassification in self-reported modern contraceptive use.
- We propose a new normal-laplace data model to account for increased asymmetric uncertainty associated with potential misclassification errors.
- Simulation study shows improvement in coverage of credible intervals when data are subject to misclassification.
- Ongoing: assessing the effect a change in data model from logit-normal to the normal-laplace in FPEM;
  findings so far: differences are typically smaller but more variable as compared to what we expect, and include changes in point estimates.

Thanks!

Contact: Chuchu Wei (chuchuwei@umass.edu) and
Leontine Alkema (lalkema@umass.edu)