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]

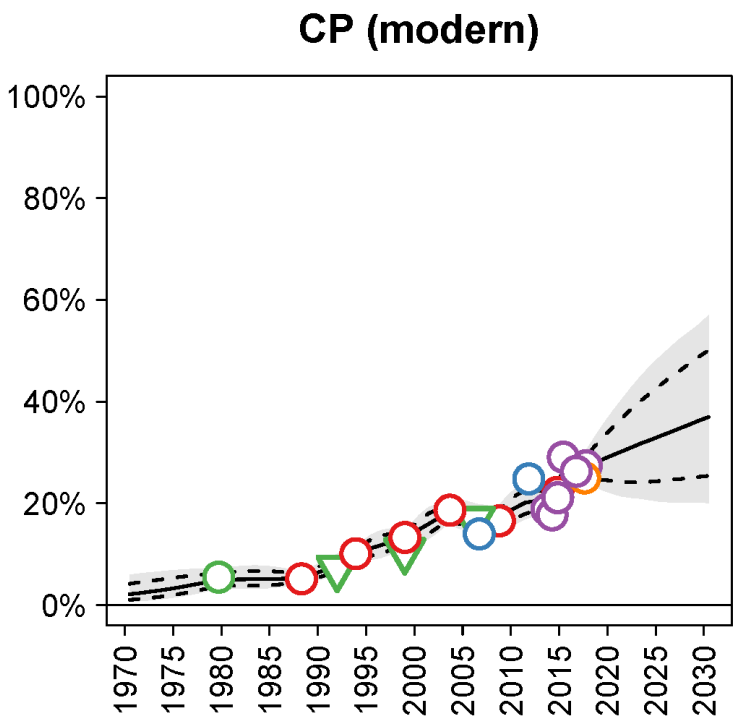
May 6, 2021

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

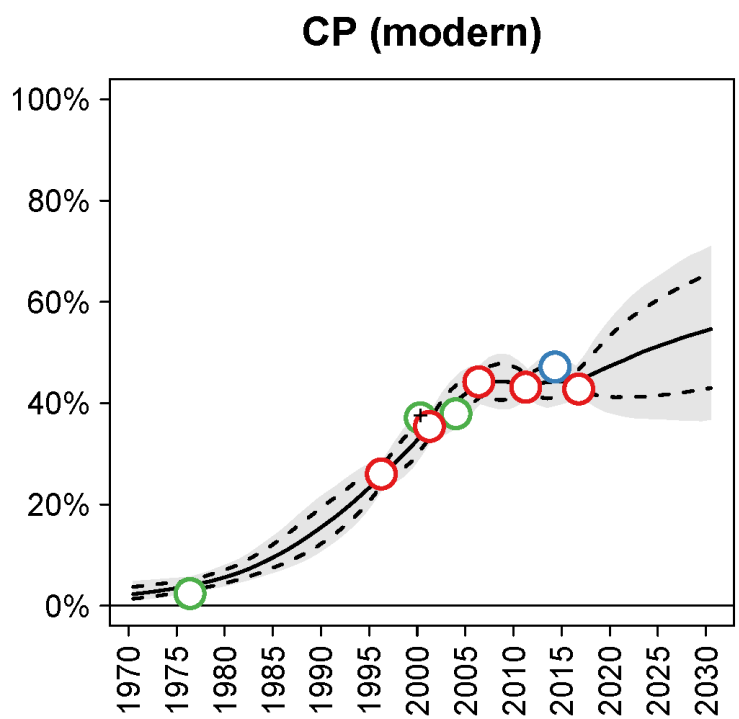
Introduction

- Motivating question:
  How to estimate contraceptive prevalence using self-reported data collection, i.e. demographic health surveys (DHS)
- The current approach: Family Planning Estimation Model (FPEM, Cahill et al., 2018)

Ghana (Western Africa)



Nepal (Southern Asia)

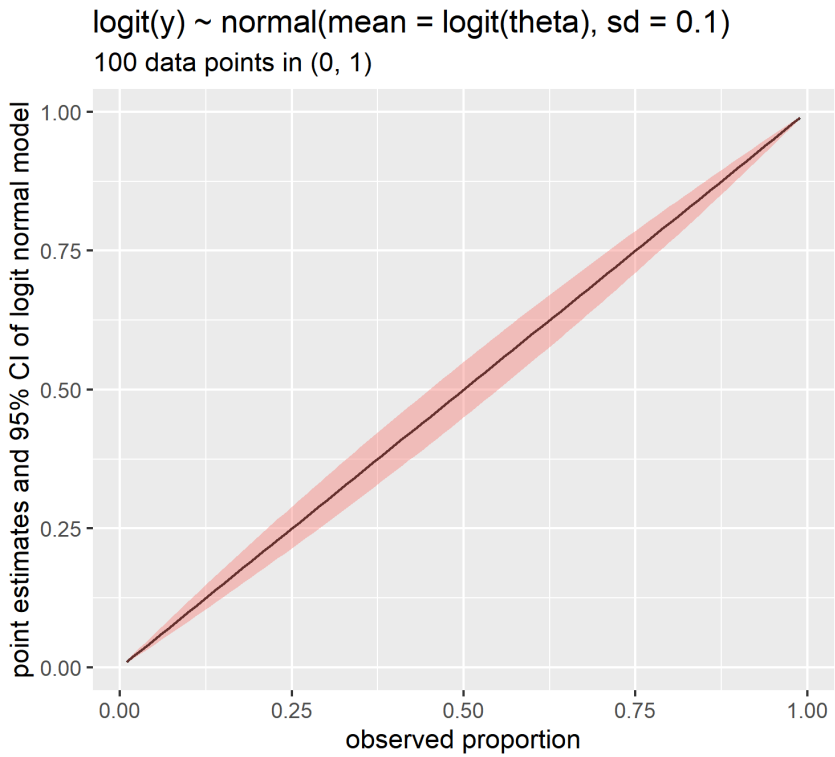


How is the data used in FPEM?

- FPEM: **data model** + process model
  - **data model**: describe how observed data relate to true prevalence

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

- Data model for modern contraceptive prevalence rate (mCPR)
  - y observed mCPR,
  - θ true mCPR,
  - logit.s sampling error on logit-scale,
  - τ non-sampling error
- The visualization of data model assumption with $\text{logit}.s = 0, \tau = 0.1$

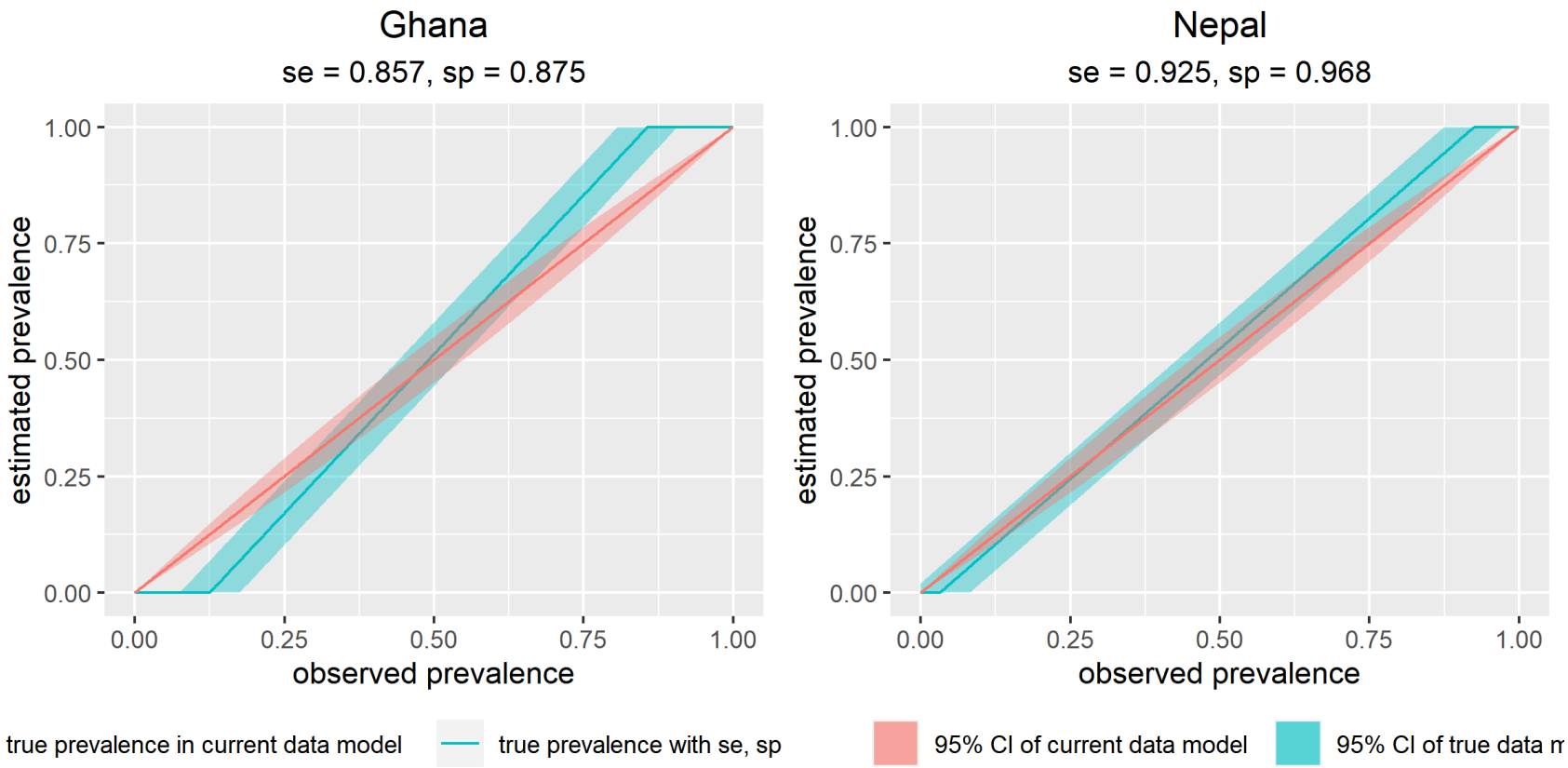


What if we observed some non-sampling error?

- 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 provides such evidence summarising into sensitivity *se* and specificity *sp*
  - 2014 Ghana DHS (Staveteig, 2017) (sample size = 48) $se = 0.857, sp = 0.875$
  - 2016 Nepal DHS (Staveteig et al., 2018) (sample size = 194) $se = 0.925, sp = 0.968$

The visualization of assumption based on misclassification

- The relationship between obs prev y and true prev θ
  $y = se \cdot \theta + (1 - sp)(1 - \theta)$
- The true data model when data subject to misclassification
  $y \sim N(se \cdot \theta + (1 - sp)(1 - \theta), s^2)\, T[0, 1]$



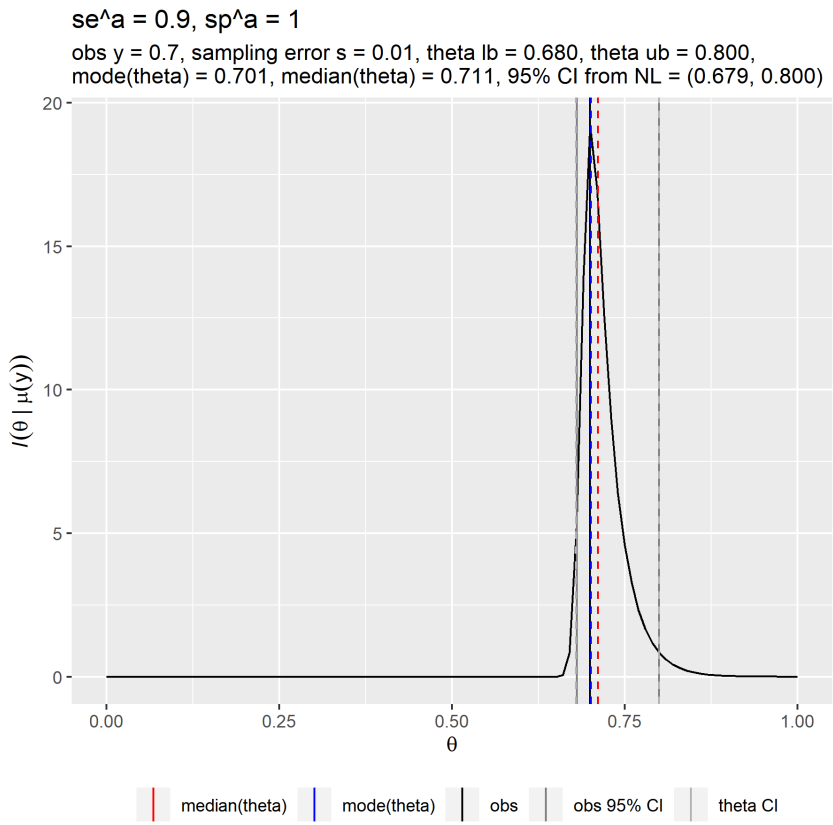
How to do a better estimation?

- 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
- Not be able to apply bias-adjustments to self-reported use for DHS data points
- What we can do:
  Update the data model to better reflect uncertainty associated with potential misclassification errors

Our proposal: a new data model

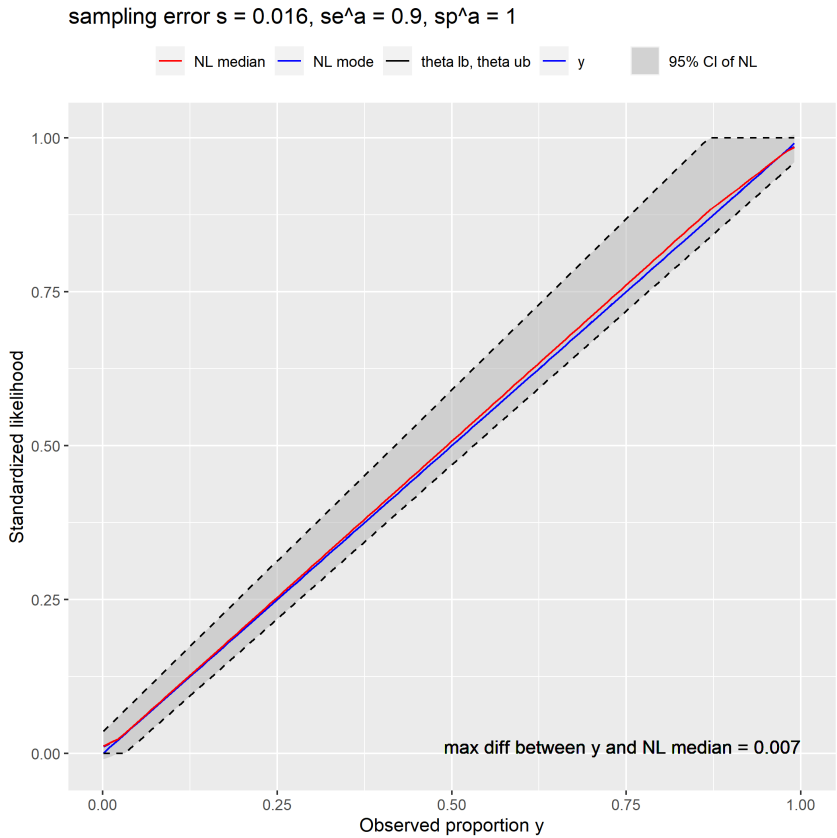
- Aims of the new model with assumed sensitivity se^a and specificity sp^a
  - $\hat{\theta} = y$
  - If no misclassification: 95% CI determined by s
  - If misclassification: 95% CI determined by s, se^a, sp^a
- Model specification: based on Normal-Laplace (NL) distribution (Reed, 2006)
  - Parameter fixed to meet aims

$$q(\theta|y) \sim NL(\mu, \sigma, \alpha, \beta)T(0,1)$$



The visualization of the new data model

- The NL model allows asymmetric CI, thus cover the potential biased prev in the uncertainty
- The visualization of the new NL model in $(0, 1)$
  - 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



Simulation of the new data model

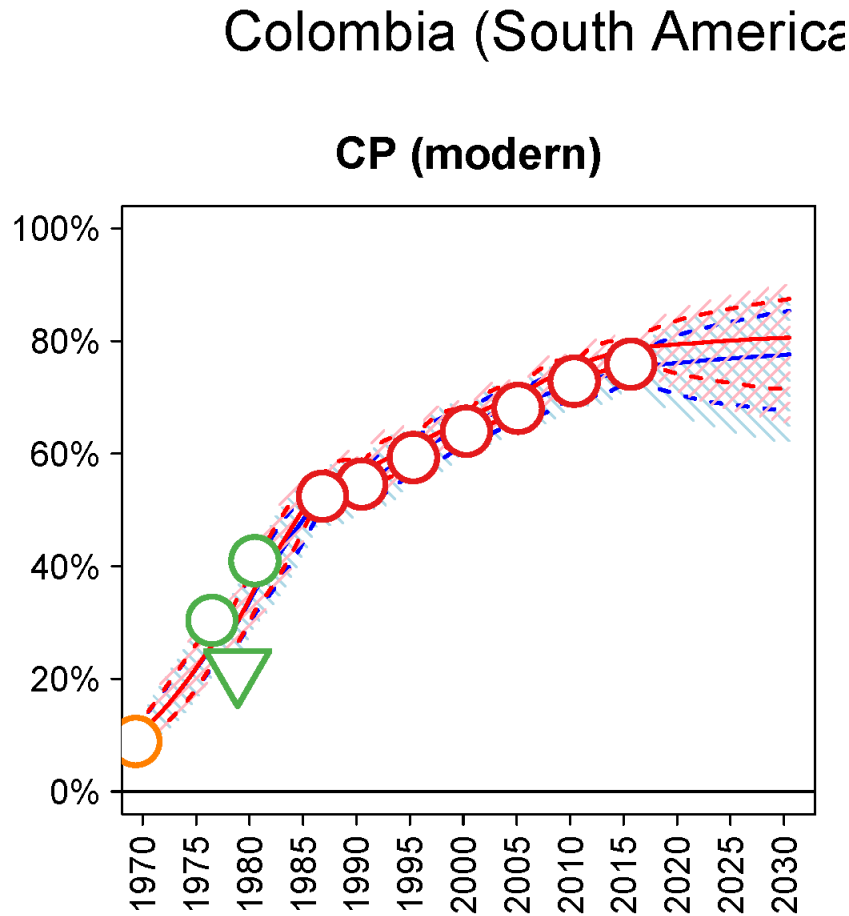
- We fixed $\theta^{\text{true}} = 0.3, 0.5, 0.7$ and set various groups of true misclassification se, sp and assumed misclassification se^a, sp^a
- True data generation process with sample size n

$$n \cdot y \sim Bin(n, se \cdot \theta^{\text{true}} + (1 - sp)(1 - \theta^{\text{true}}))$$

- Compare posterior estimates from
  - True process model
  - Current assumption in FPEM (logit-normal)
  - The new proposed model (NL likelihood)
- Simulation summary (1) the estimates is unchanged in the new NL model regardless of the misclassification (2) If misclassification exists: the new model did a better job of including θ^{true} in 95% credible interval compare to the old logit-normal model

New data model in FPEM

- New proposal vs the current data model in FPEM
- Apply to DHS data points only
- Allow user to define the misclassification, e.g. $se^a = 0.9, sp^a = 1$
- Observe changes in the estimated trends and projection across countries



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