# 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]

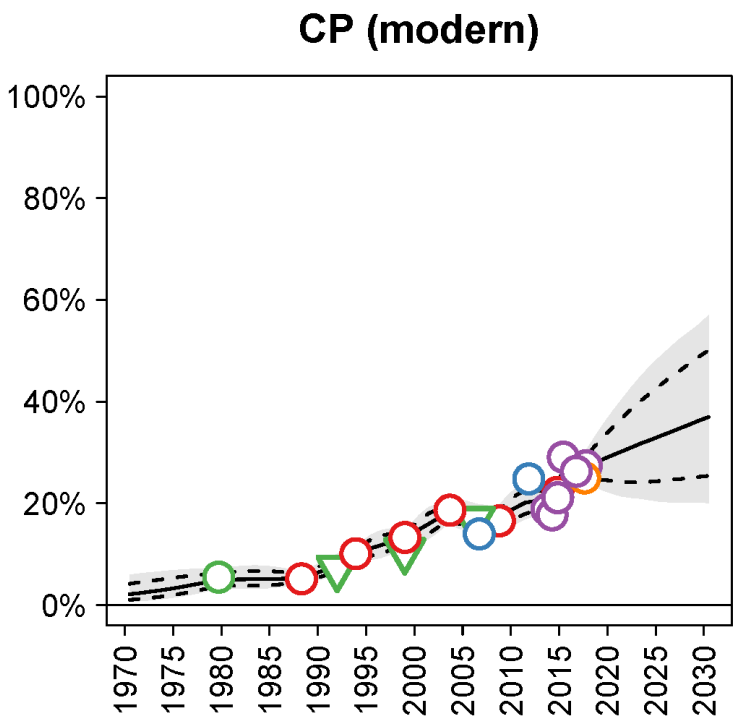
May 6, 2021

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

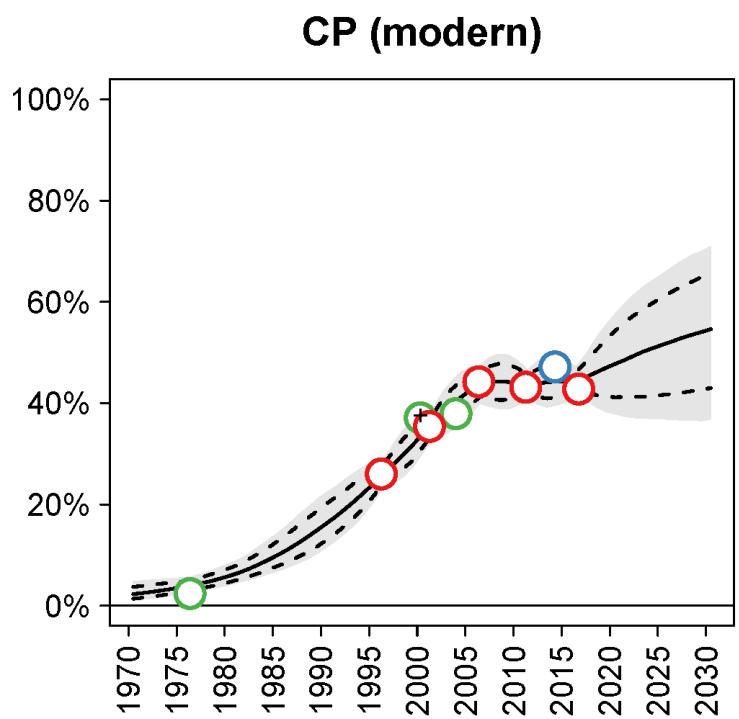
# Introduction

- 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)

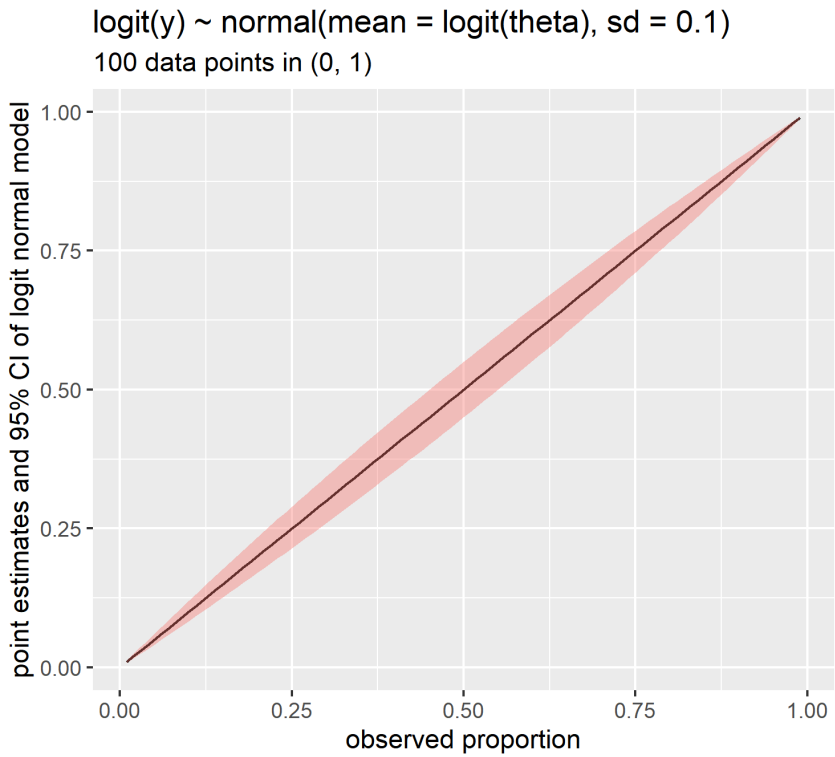


# How is the data used in FPEM?

- 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,
  - $\theta$ true mCPR,
  - logit.$s$ sampling error on logit-scale,
  - $\tau$ non-sampling error
- The visualization of data model assumption with $\text{logit}.s = 0, \tau = 0.1$

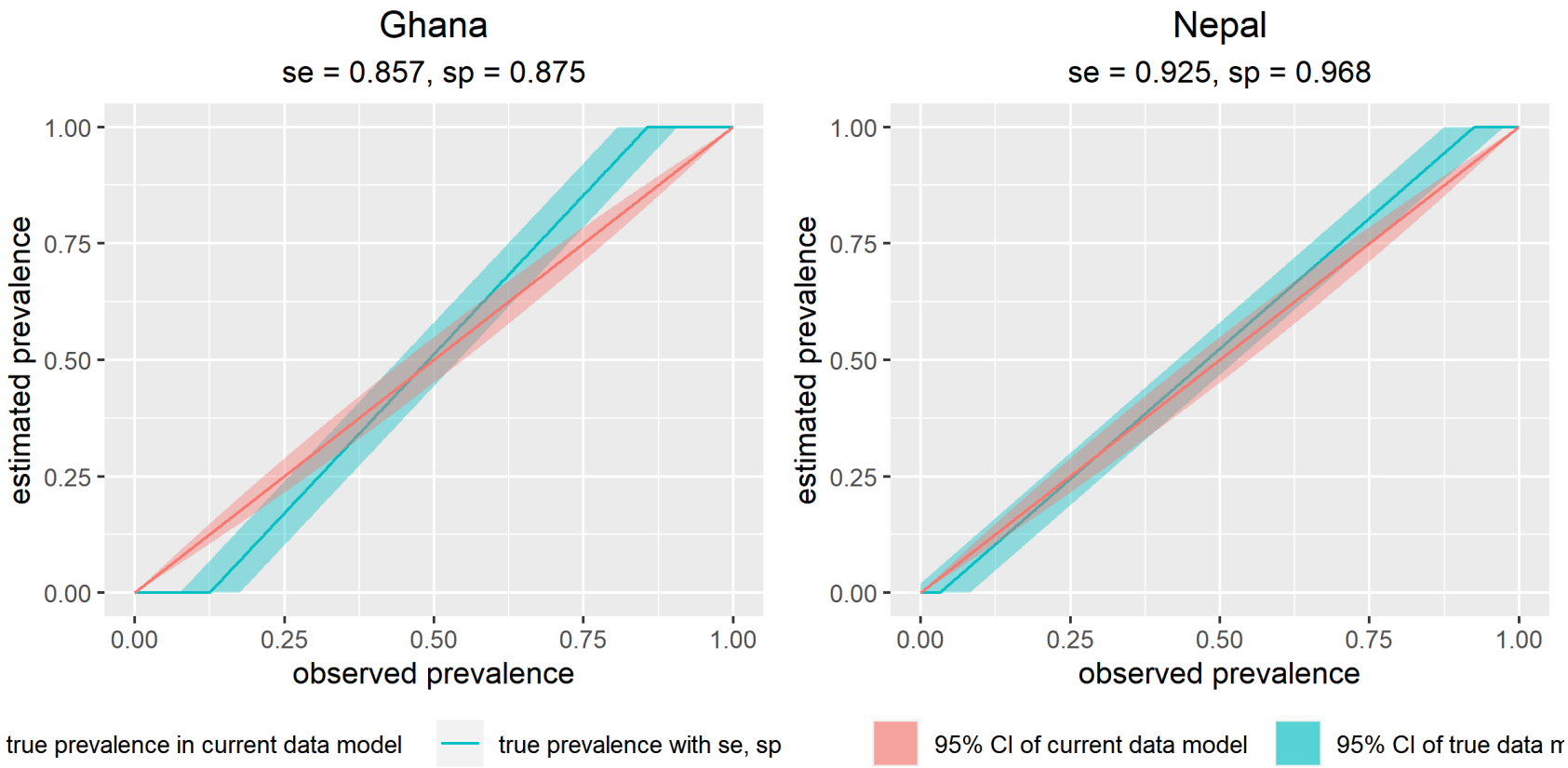


# What if we observed some non-sampling error?

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



# 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 $\theta$ 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