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, Leontine Alkema
University of Massachusetts Amherst

University of Massachusetts Amherst

Summary

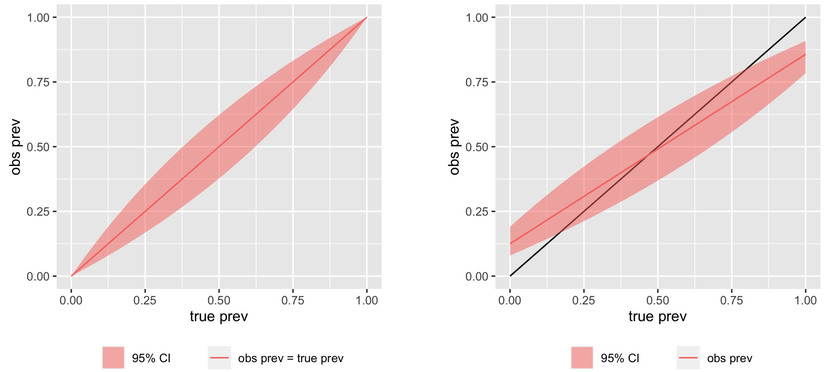
Challenge
Estimating of a population proportion using data subjects to unknown misclassification error. Observed misclassification data are too sparse to generalize to population-periods.
Application
Estimating contraceptive prevalence based on self-reported usage
Current vs new approaches
The current approach does not accurately account for increases in uncertainty and potential bias introduced by misclassification. We propose new approaches to better explain misclassification errors in the uncertainty assessment of the unknown population proportion. Simulation results show the new approaches outperform the current approach.

The Current Approach

Data sources aggregated self-reported contraception usage from surveys (i.e., Demographic and Health Surveys)
Likelihood function used currently (simplified form)

$$\text{logit}(y)|\theta, \sigma \sim N(\text{logit}(\theta), \sigma^2),\ \sigma^2 = logit.SE^2 + NSE^2,$$

$logit.SE^2$ logit-transformed sampling variance; NSE^2 non-sampling variance
When SE is negligible, NSE decreases as the propotion increases/decreases away from 0.5 (left: current data model based on NSE; right: findings from small studies)



The new approach

Data with misclassification error
Misreporting data from two post-survey studies show the existence of misclassification error in the survey data collection, but these findings are not generalizable.
We summarize the misclassification in terms of sensitivity se and specificity sp and allow for user to propose likely values for se and sp to inform the uncertainty associated with the estimated pop. proportion.
Relation between obs and true prevalence θ

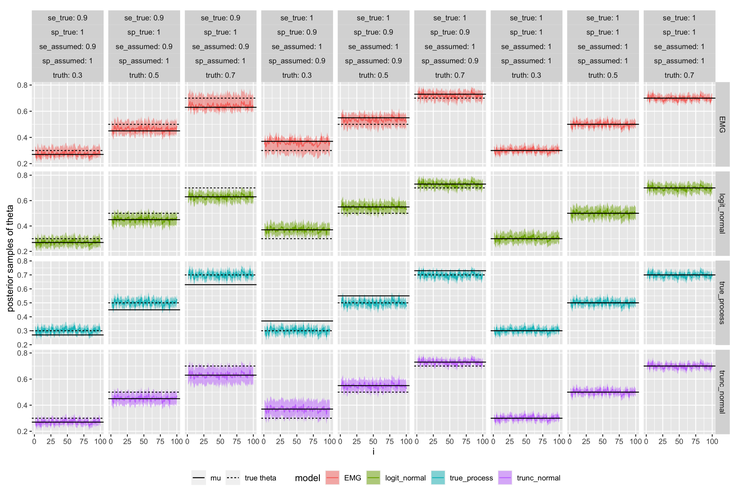
$$E(y) = se \cdot \theta + (1 - sp)(1 - \theta),$$

y observed proportion; θ true prevalence; se true sensitivity; sp true sensitivity
Bias-adjusted observations δ

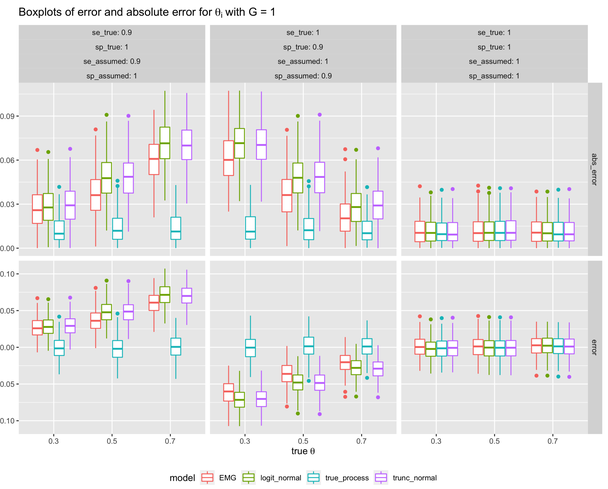
$$y = \delta \cdot se^a + (1 - \delta)(1 - sp^a),\ \delta = \max\left(0, \min\left(1, \frac{y-(1-sp^a)}{se^a-(1-sp^a)}\right)\right)$$

se^a, sp^a assumed sensitivity, specificity;
Alternative approaches
Truncated normal:
$y|\theta \sim N_{[0,1]}(\theta, SE^2 + NSE^2)$. The variable $SE^2 + NSE^2$ are set to include the bias-adjusted observation in the 95% CI for θ
Exponentially modified Gaussian (EMG):
$p(y|\theta) \sim EMG(\mu, \sigma, \lambda)$. EMG parameters μ, σ, λ are set such that bias-adjusted observation is contained in 95% CI for θ

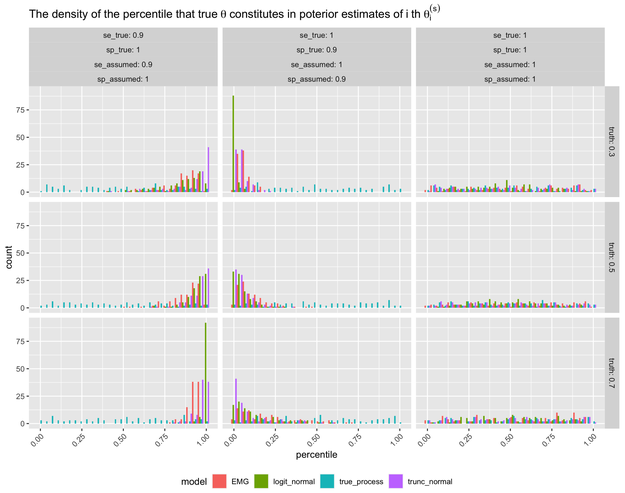
Simulation Results



(1) 95% posterior CI for θ_i of all datasets.



(2) The error and absolute error of posterior samples.



(3) The percentile that true θ substitutes in the posterior samples.

Simulation

Generate data with dataset $i, i = 1 \cdots n$
$\theta_i \sim N_{[0,1]}(\theta^{\text{true}}, 1)$, $y_i \cdot n_i \sim bin(n_i, \theta_i)$
Four models for ith datasets:
1. True process: $y_i \cdot n_i \sim bin(n_i, \theta_i)$
2. Logit normal:
$$logit(y_i) \sim N(logit(\theta_i), logit.SE_i^2 + NSE_i^2)$$
3. Truncated normal:
$$y_i \sim N_{[0,1]}(\theta_i, SE_i^2 + NSE_i^2)$$
4. EMG:
$$\log L(y_i|\theta_i) \propto EMG(\mu_i, \sigma_i, \lambda_i)$$

Prior $\theta_i \sim U(0, 1)$
Settings for se **and** sp
(1) $se = sp = 1$; (2) $se < 1, sp = 1$ w/ $se = se^a$, $se \neq se^a$; (3) $se = 1, sp < 1$;
Measurement
Error of posterior samples: $\theta_i^{(s)} - \theta^{\text{true}}$
% θ^{true} in $\theta_i^{(s)}$: $\frac{1}{S}\sum_{s=1}^{S} \theta_i^{(s)} \leq \theta^{\text{true}}$

References

[1] S. Purushothaman et al. “Hyper-EMG: A new probability distribution function composed of Exponentially Modified Gaussian distributions to analyze asymmetric peak shapes in high-resolution time-of-flight mass spectrometry”. In: *International Journal of Mass Spectrometry* 421 (Oct. 1, 2017), pp. 245–254. ISSN: 1387-3806. DOI: 10.1016/j.ijms.2017.07.014.

[2] Sarah Staveteig. “Fear, Opposition, Ambivalence, and Omission: Results from a Follow-up Study on Unmet Need for Family Planning in Ghana”. en. In: *PLOS ONE* 12.7 (July 2017), e0182076. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0182076.

[3] Sarah Staveteig et al. “Barriers to Family Planning Use in Eastern Nepal: Results from a Mixed Methods Study”. In: *DHS Qualitative Research Studies No. 21* (2018).