Bias-variance and breadth-depth tradeoff in Respondent-Driven Sampling

Sergiy Nesterko

Harvard University

NESS 2011

What is RDS?

- Used to collect information (income, illness status) from hard-to-reach populations (injection drug users, at high risk of HIV populations)
- Infer population mean
- Over 150 studies worldwide using RDS in 2003-2007
- How does it work?

VH estimator

- Current state of the art estimator, Volz-Heckathorn (2008):

$$\hat{\theta} = \frac{1}{\sum D_i^{-1}} \sum D_i^{-1} X_i$$

Motivation and goals

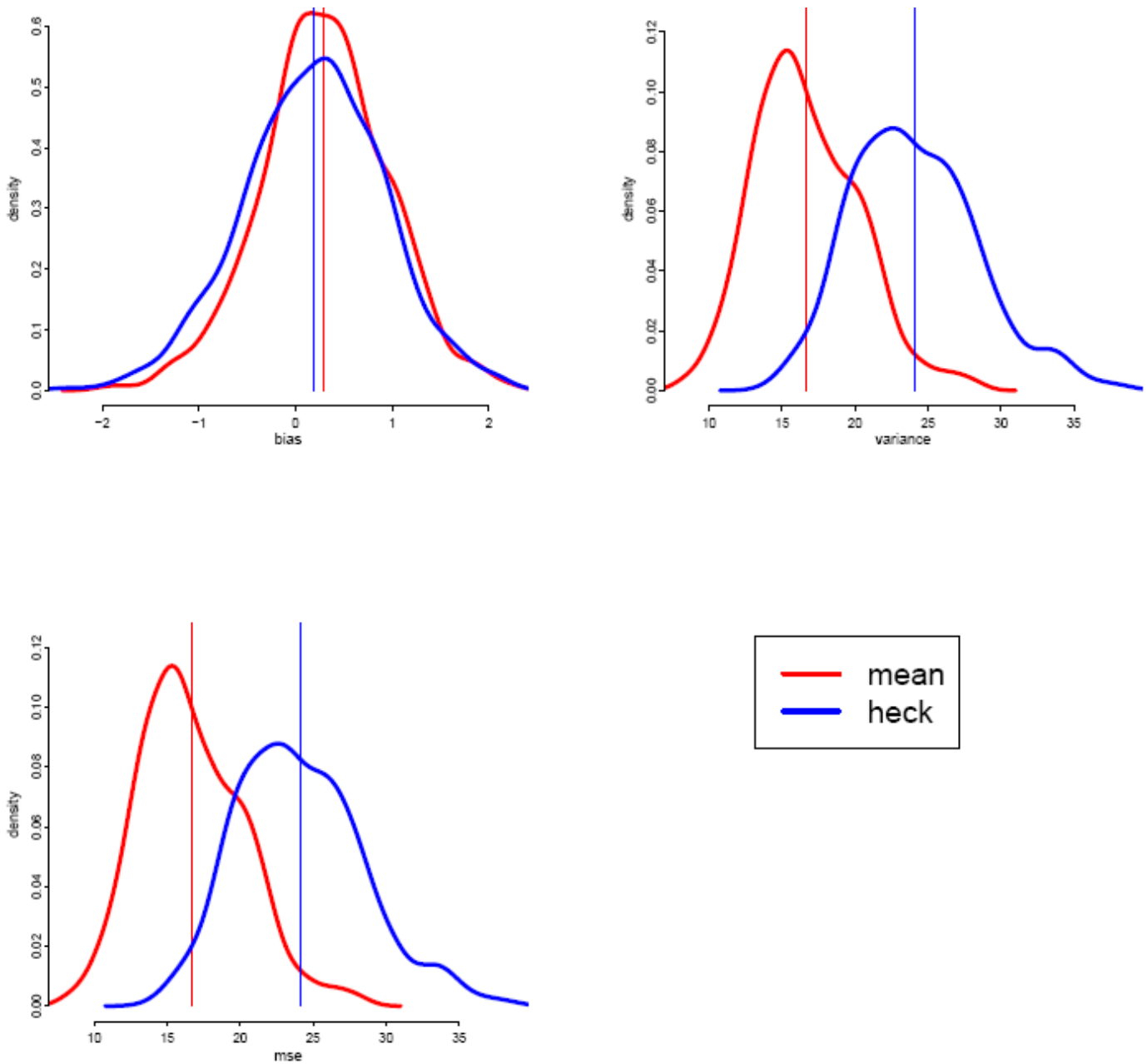
- Conduct a simulation study
- Check relative performance of the VH and plain mean estimator in terms of bias and variance, and derive intuition as to what may drive the behavior
- Use the intuition to better inform design of studies, and theoretical work on estimation

How the simulations work

- Simulate RDS process with arguably more realistic features, for example
  - homophily
  - rich-gets-richer
- Essentially explore different types of dependence of participant referral/network structure on the quantity of interest

An example of a finding

- The plain mean outperforms the VH estimator on networks with homophily


density
bias
variance
mse
mean
heck

Why so

Other findings

- The VH estimator generally outperforms the plain mean under rich-gets-richer scenarios
- A number of coupons larger than 2 is necessary to keep the chains going
- The gap in performance between the estimators is increasing with population size
- Personal network size estimation may be detrimental to the performance of the VH estimator

Conclusion

- Valuable intuition for RDS process and related estimation may be obtained via simulation
- The plain mean may be more appropriate than the VH estimator in some settings
- Development of new estimation techniques is necessary (work in progress)

Thank you

- The audience
- My advisor Joe Blitzstein

Sample meta-visualization