Network sampling and estimation for hard-to-reach populations.

Sergiy Nesterko
G3
Harvard University

December 4, 2009

Plan

0. Motivation.
1. Background.
2. Simulations.
3. Conclusion.

Motivation

- Current. San Diego study – need to make educated estimation decisions.
- General. Develop better estimation and sampling techniques in the setting of hard-to-reach populations.

Background

- Respondent-driven sampling to sample from hard-to-reach populations
- Theoretical work on estimation done only by Heckathorn (2002, 2004)
- Performance tested by Goel and Salganik (2007), Gile and Handcock (2009), each with drawbacks

Heckathorn estimator

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

- Assume D_i known and fixed (and X and D independent).
- Assume uniform participant recruitment.
- Hopefully, stationarity quickly achieved.

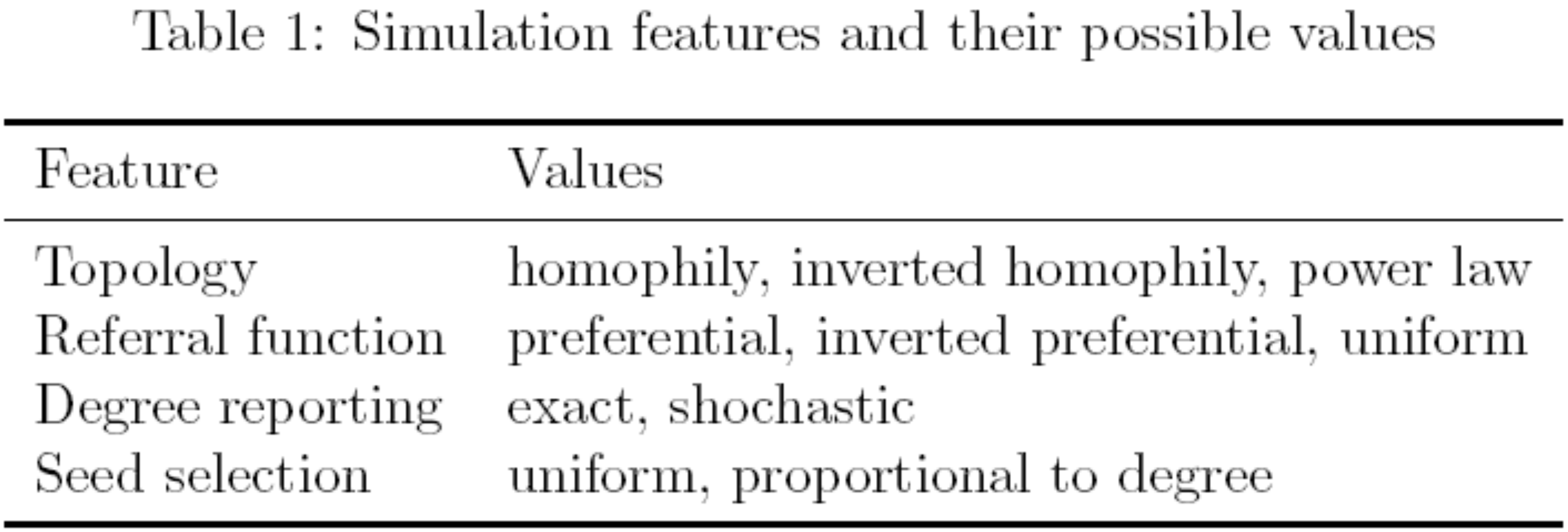


Table 1: Simulation features and their possible values

Feature	Values
Topology	homophily, inverted homophily, power law
Referral function	preferential, inverted preferential, uniform
Degree reporting	exact, shochastic
Seed selection	uniform, proportional to degree

- 36 possible combinations
- Each simulation consists of simulating 500 networks and 500 RDS processes on each.

Simulations setup details

0. Generate quantity of interest: 100 iid draws from Normal(170, 100), and assign them to vertices.
1. Create links based on differences, using one of three predefined functions.
2. Get 500 RDS samples using one of referral functions, calculate estimates.

- Repeat 0-2 500 times.

Particulars

- Topology functions:

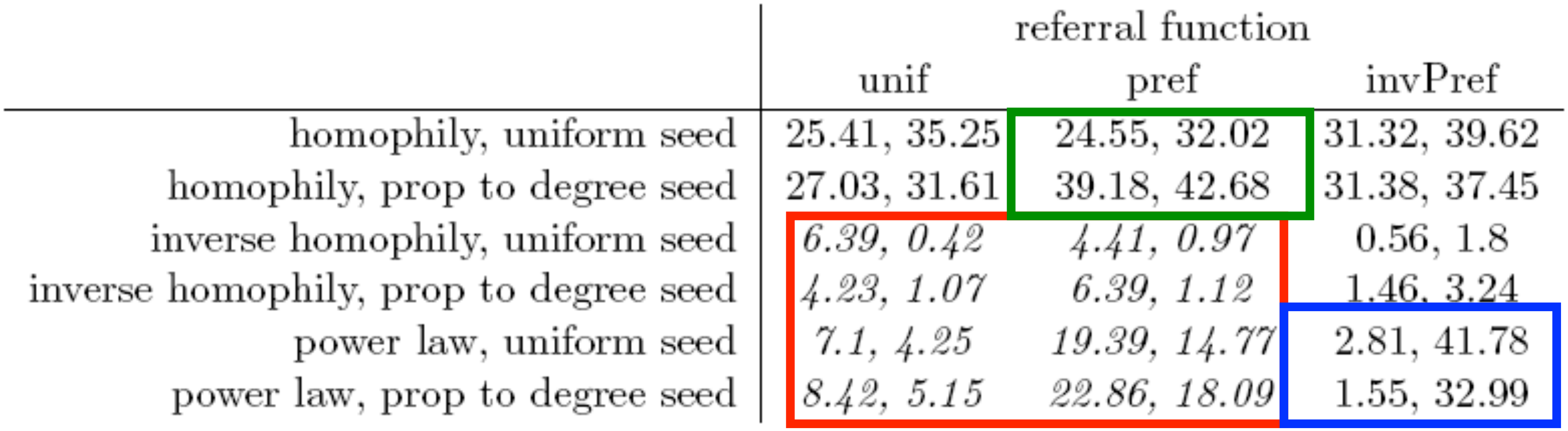
$$\begin{aligned}
\text{homophily:} \quad & P(l_{ij}=1) = invlogit\left(-d(x_i,x_j)\right), \\
\text{inverted homophily:} \quad & P(l_{ij}=1) = invlogit\left(-20+d(x_i,x_j)\right), \\
\text{power law:} \quad & P(l_{ij}=1) = .01 + \frac{.2}{|\chi|} rank\left(max(x_i,x_j)\right)
\end{aligned}$$

- Referral functions:

$$\begin{aligned}
\text{preferential:} \quad & P(X_{i+1}=x_{i+1}|X_i=x_i) = d(x_i,x_{i+1})^{-1.5} \Bigg/ \sum_j d(x_i,x_j)^{-1.5}, \\
\text{inverted preferential:} \quad & P(X_{i+1}=x_{i+1}|X_i=x_i) = e^{d(x_i,x_{i+1})} \Bigg/ \sum_j e^{d(x_i,x_j)}, \\
\text{uniform:} \quad & P(X_{i+1}=x_{i+1}|X_i=x_i) = I_{ij} \Bigg/ \sum_j I_{ij},
\end{aligned}$$

Simulation results

- Compare with plain mean.

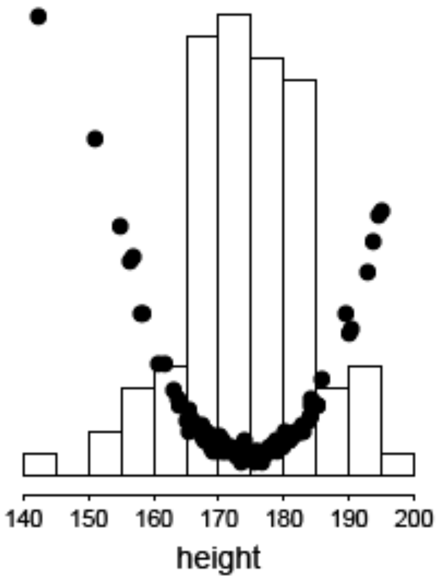


	referral function		
	unif	pref	invPref
homophily, uniform seed	25.41, 35.25	24.55, 32.02	31.32, 39.62
homophily, prop to degree seed	27.03, 31.61	39.18, 42.68	31.38, 37.45
inverse homophily, uniform seed	*6.39, 0.42*	*4.41, 0.97*	0.56, 1.8
inverse homophily, prop to degree seed	*4.23, 1.07*	*6.39, 1.12*	1.46, 3.24
power law, uniform seed	*7.1, 4.25*	*19.39, 14.77*	2.81, 41.78
power law, prop to degree seed	*8.42, 5.15*	*22.86, 18.09*	1.55, 32.99

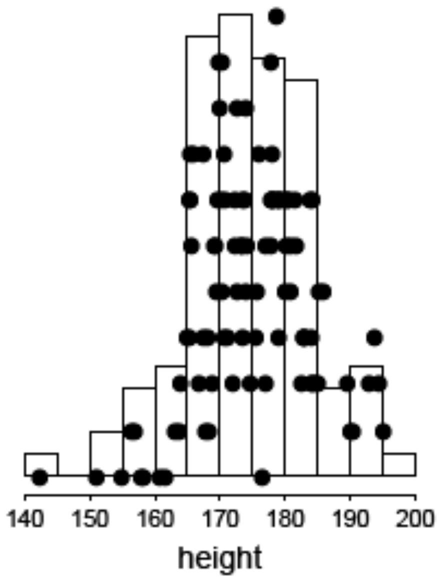
- Dirrefent relative performance in different settings.

Why is this happening?

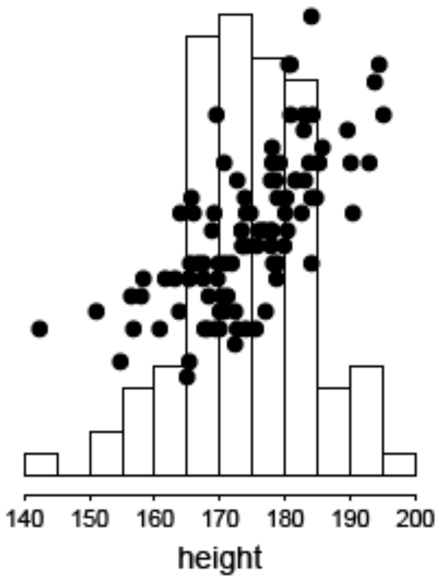
Inverted homophily



Homophily

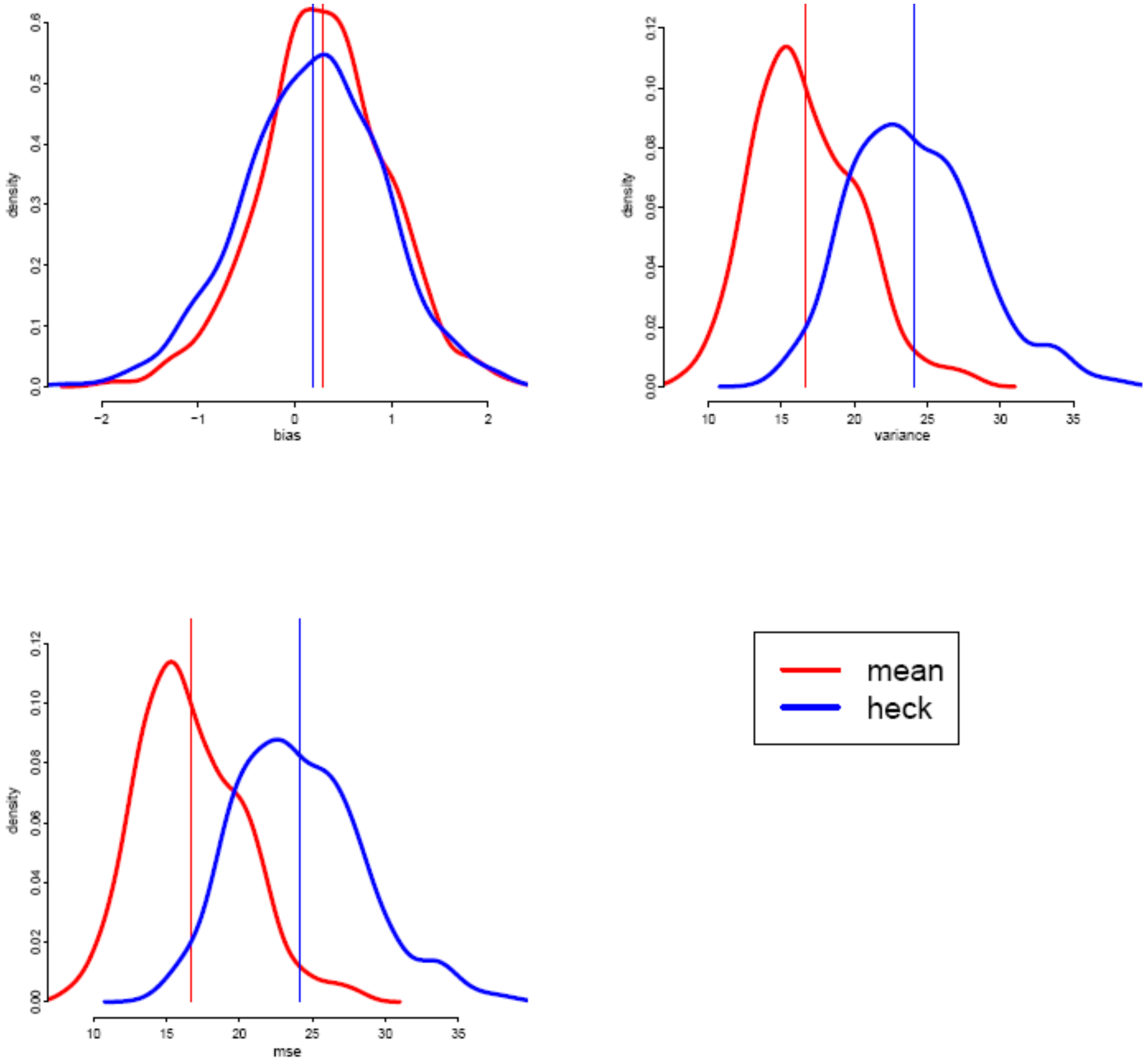


Power law

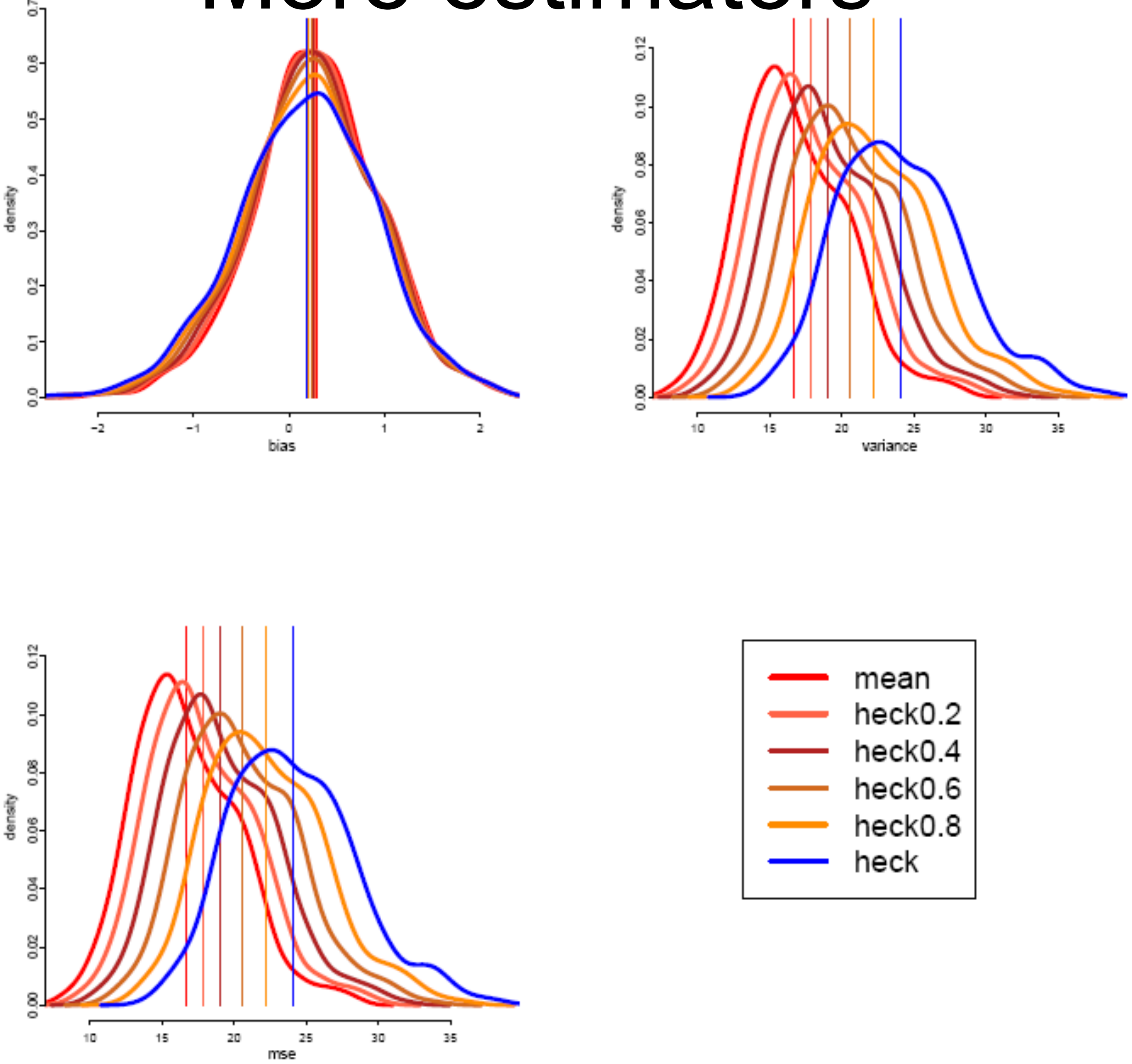


- Histograms are those of quantity measured
- Dots are normalized vertex counts

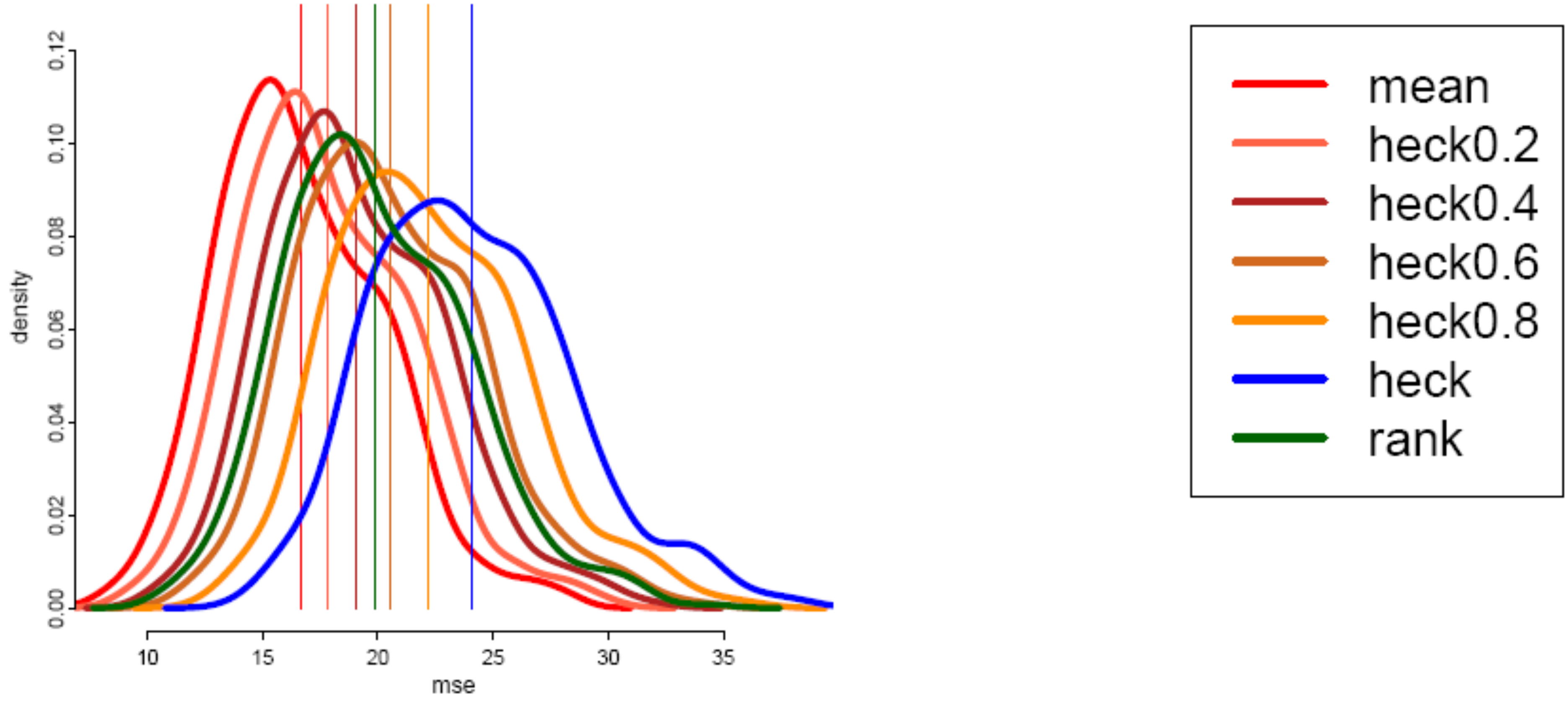
Graphical results



More estimators



One more



- Rank estimator usually between heck0.4 and heck0.6. Magic 0.5?

Conclusion

- Heckathorn estimator is not robust to violations of assumptions.
- Need to better understand what conditions we are in when performing estimation. For this, need better sampling design.
- Work with San Diego data.

- Thank you, Joe.