Let $X_1, X_2, ..., X_n$ be all collected samples during RDS. Then RDS estimate μ_f of the population mean of f is defined [1] as

$$\mu_f = \frac{1}{\sum_{i=1}^{n} 1/degree(X_i)} \sum_{i=1}^{n} \frac{f(X_i)}{degree(X_i)}$$

Our estimator (sample mean):

$$\mu_{f2} = \sum_{i=1}^{n} \frac{f(X_i)}{n}$$

Estimator μ_f indeed performs better than estimator μ_{f2} when values of the nodes depend on the degree of the node.

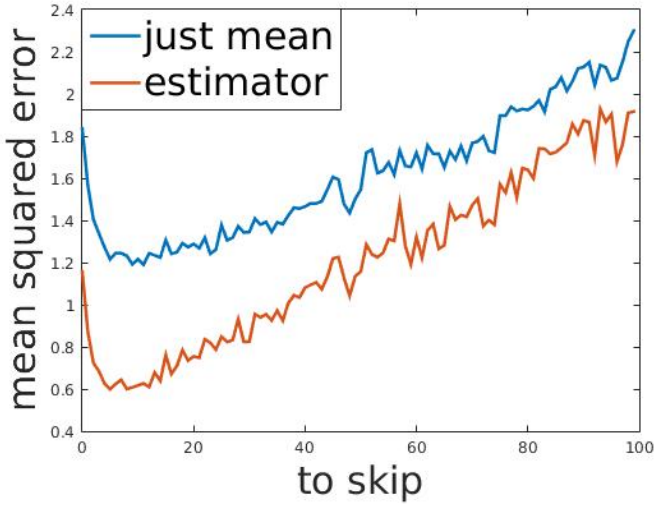

Figure 1: RGG(200, 0.13), measuring degree

As stated also [1] the advantage to use estimtor1 appears only when the needed for estimation trait depends on the degree of the node. To support this statement they compare particularly the standard error of the sample mean and RDS estimate on the data sets from Project 90 and Add Health. The results are presented on the figure [put pictures].

But why it so much better? Why it is so suspicious? Possible explanation: degree weights are used to correct the fact that we see high degree nodes more

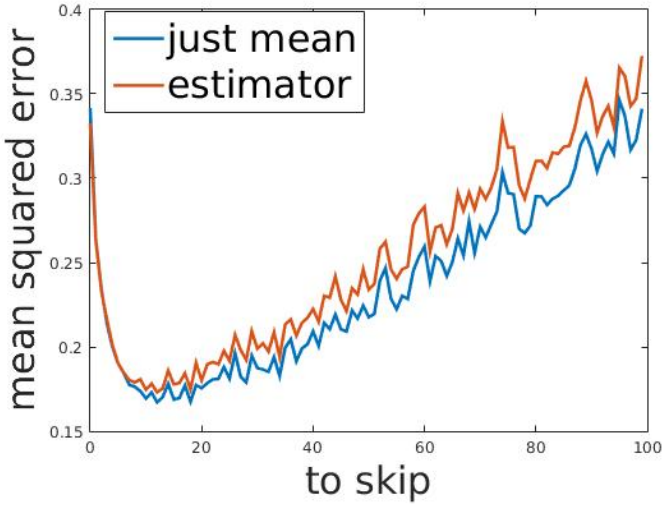

Figure 2: RGG(200, 0.13), measuring values

times. Then for sure, if the values are related to the degree of the node it is useful. But if not it may do worth? but... it should not

Another explanation: so we give less weight to the nodes with high degree. But If they are more representative (like in graph with Gibbs field)....?

try: separate bias and variance

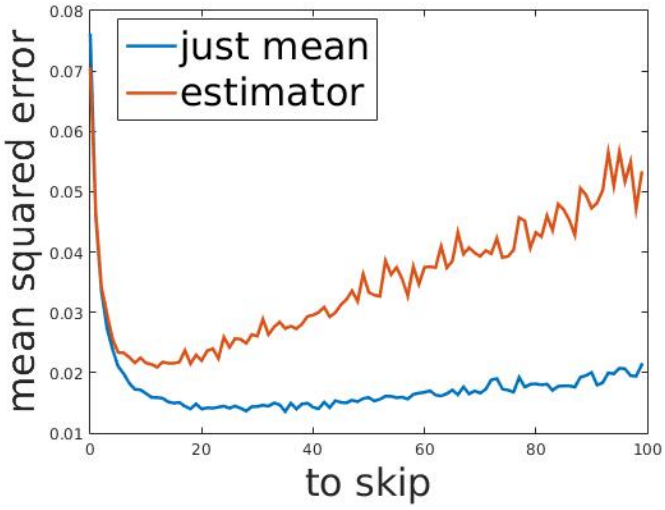

Figure 3: Project 90, measuring ??? race

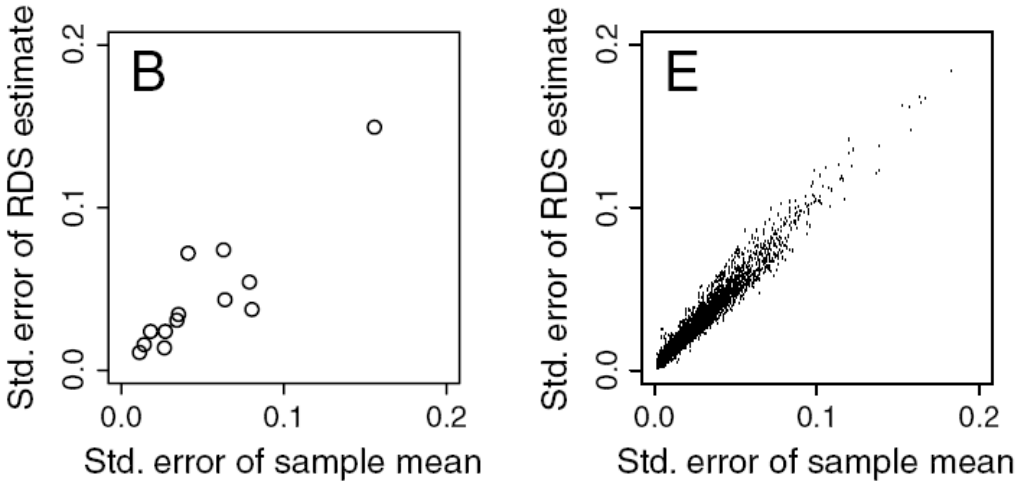

Figure 4: Comparison of standard error of RDS estimator and sample mean estimator on Project 90 data (left) and Add Health data (right) [1]

Bibliography

[1] Sharad Goel and Matthew J Salganik. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15):6743–6747, 2010.