The New England Journal of Statistics in Data Science Volume 4, 64–76 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS100

Inverse Probability Weighting: From Survey Sampling to Evidence Estimation☆

Jyotishka DATTA* and Nicholas POLSON

Abstract

We consider the class of inverse probability weight (IPW) estimators, including the popular Horvitz–Thompson and Hájek estimators used routinely in survey sampling, causal inference and for Bayesian computation. We focus on the 'weak paradoxes' for these estimators due to two counterexamples by Basu (1988) and Wasserman (2004) and investigate the two natural Bayesian answers to this problem: one based on binning and smoothing: a 'Bayesian sieve' and the other based on a conjugate hierarchical model that allows borrowing information via exchangeability. We compare the mean squared errors for the two Bayesian estimators with the IPW estimators for Wasserman's example via simulation studies on a broad range of parameter configurations. We also prove posterior consistency for the Bayes estimators under missing-completely-at-random assumption and show that it requires fewer assumptions on the inclusion probabilities. We also revisit the connection between the different problems where improved or adaptive IPW estimators will be useful, including survey sampling, evidence estimation strategies such as Conditional Monte Carlo, Riemannian sum, Trapezoidal rules and vertical likelihood, as well as average treatment effect estimation in causal inference.

KEYWORDS AND PHRASES: Inverse probability weighting, Horvitz–Thompson, Hájek, Importance sampling, Stein phenomenon, Bias-variance trade-off, Evidence Estimation.

1. INTRODUCTION

Inverse probability weight (IPW) estimators have been used across statistical literature in diverse forms: in survey sampling, in designing importance sampling in Monte Carlo techniques and in the context of average treatment effect estimation in causal inference. In survey sampling, the goal is often to estimate population mean ψ from a finite sample $(y_1, \ldots, y_n)$, and a common approach is to weigh each observation y_i by a weight w_i inversely related to their probability of inclusion (probability proportional to selection, or PPS). For example, common PPS estimators admit the form $\hat{\psi} = \sum_{i\in s} w_i y_i/n$ or a 'ratio' estimator: $\hat{\psi} = \sum_{i\in s} w_i y_i / \sum_{i\in s} w_i$, where s denotes the sample, and w_i's denote the sampling weights. These weights or probabilities of inclusion might be known, or unknown depending on whether their source is sampling design or non-response. Two of the most popular examples of such estimators are the Horvitz–Thompson estimator (Horvitz and Thompson, 1952), which uses fixed weights to provide an unbiased estimate of the population mean, and the Hájek estimator (Hájek, 1971), which normalizes the weights to improve stability at the cost of introducing slight bias.

Similarly, in importance sampling, the goal is to estimate the evidence $\psi = \mathbb{E}_F\{l(X)\} = \int l(x)dF(x)$ where F might be difficult to sample from and a common technique is to estimate ψ by drawing samples from a candidate distribution G (with density g) and calculate:

$$\hat{\psi} = n^{-1}\sum_{i=1}^{n} l(x_i)f(x_i)/g(x_i) \doteq n^{-1}\sum_{i=1}^{n} l(x_i)w(x_i).$$

The candidate density $g(\cdot)$ is chosen such that the 'weights' $w(x_i)$ is nearly constant (Firth, 2011). The 'ratio' estimator analog in importance sampling would correspond to $\sum_{i=1}^{n} l(x_i)w(x_i)/\sum_{i=1}^{n} w(x_i)$ as developed by Hesterberg (1988, 1995), and further developed in (Firth, 2011). The unattainable 'optimal' choice of $w(\cdot)$ is of course $f(\cdot)$ itself, and a key insight in producing more accurate estimation is that self-normalization or biasing the sampler towards low probability regions can help. Methods such as nested sampling or vertical likelihood use a Lorenz curve re-ordering of summation to achieve this goal (Skilling, 2006; Chopin and Robert, 2010; Polson and Scott, 2014; Datta and Polson, 2025).

A third important setting where inverse probability weighting plays a central role is causal inference. In the potential outcomes framework, IPW estimators form the backbone of average treatment effect (ATE) estimation by reweighting observed outcomes according to treatment assignment probabilities (Rubin, 1974; Imbens, 2004;

arXiv: 2204.14121
☆Corrected on 28 September 2026: in Section 3.1 (Example 1), "each of the remaining with 1/100" was changed to "each of the remaining with 1/4900".
*Corresponding author.

Cunningham, 2021). We briefly discuss the ATE problem in subsection 2.3. IPW estimators provide a common design-based principle whether adjusting for unequal inclusion probabilities in surveys, stabilizing evidence estimation in Bayesian computation, or addressing confounding in observational studies. The wide applicability of IPW motivates our reexamination of their weaknesses and possible Bayesian resolutions.

In this paper, we first contrast and compare the popular inverse probability weight estimators—Horvitz–Thompson and Hájek—and discuss their relative merits and demerits in light of two weak paradoxes, viz., Basu's circus example (Basu, 1988) and the Robins–Ritov–Wasserman example (Robins and Ritov, 1997; Wasserman, 2004). We then discuss a binning-and-smoothing estimator by Ghosh (2015) and a hierarchical Bayes estimator by Li (2010), in the light of the Wasserman (2004)'s example, and show how they can lead to a possible resolution. We also build connections between IPW estimators and evidence estimation techniques and argue that these innovative ideas from Monte Carlo methods can be exploited in designing IPW estimators to achieve a lower variance and higher stability. We argue that the issue of choice of weights – an optimal proposal $g(\cdot)$ in Monte Carlo or sampling weights – provides new insights on a long-standing controversy about a 'weakness' in Bayesian paradigm (e.g. Wasserman, 2004; Sims, 2010; Ghosh, 2015; Li, 2010).

There is a large, influential literature on statistical methods for improving survey-weighted estimates, particularly, in the context of complex survey designs. Our goal is not to attempt a review of these methodological advances in survey weight regularization or calibration, and we point the interested reader to the comprehensive review by Chen et al. (2017) and the references therein, e.g., Haziza and Beaumont (2017). Haziza and Beaumont (2017) provide a comprehensive review of methods for constructing survey weights, focusing on techniques to adjust for unequal probabilities of selection, nonresponse, and post-stratification to improve the representativeness and accuracy of survey estimates. Chen et al. (2017) discuss the limitations of basic design-based weights, derived from the inverse of inclusion probabilities, and propose modifications such as weight trimming, weight modeling, and incorporating weights into statistical models.

The structure of this article is as follows. In §2, we define the popular IPW estimators and recent developments, and their connections with importance sampling and ATE estimation. In §3, we discuss two popular examples of 'weak paradoxes' due to Basu, and Robins-Ritov-Wasserman. For the latter, we derive asymptotic properties of a Bayes estimator due to Li (2010) in §3.4. We compare different survey sampling estimators through a range of simulation studies in §4, to illustrate their relative merits and demerits. Finally, in §5, we discuss other areas of connection such as average treatment effect estimation in potential outcomes framework, and suggest a few possible future directions.

2. INVERSE PROBABILITY WEIGHT ESTIMATORS

Horvitz–Thompson and Hájek Estimators

We begin by defining two widely used estimators in survey sampling: the Horvitz–Thompson (HT) estimator (Horvitz and Thompson, 1952)[1] and the Hájek estimator (Hájek, 1971) briefly. Consider a finite population $U = \{1, 2, \ldots, N\}$ with values Y_k for each unit $k \in U$. A sample $s \subset U$ is drawn according to a sampling design with known inclusion probabilities $p_k = P(k \in s)$.[2] The Horvitz–Thompson estimator for the population mean $\psi = N^{-1} \sum_{k \in U} Y_k$ is then

$$\hat{\psi}_{HT} = \frac{1}{N} \sum_{k \in s} \frac{Y_k}{p_k}. \quad \text{(Horvitz–Thompson)} \tag{2.2}$$

The Hájek estimator is an alternative procedure that estimates the population sum as well as N, i.e., it normalizes the estimated sum by the estimated total:

$$\hat{\psi}_{\text{Hájek}} = \frac{\sum_{k \in s} Y_k / p_k}{\sum_{k \in s} 1/p_k}. \quad \text{(Hájek)} \tag{2.3}$$

The HT estimator (2.2) is an unbiased estimator of ψ, while the Hájek estimator (2.3) is non-linear, and approximately unbiased estimator that does not require the knowledge of population size N, irrespective of whether N is known.

Missing Data Framework and Adaptive Normalization

An equivalent framework arises in the context of missing data, *i.e.,* when subjects are observed with nonuniform probabilities. We follow the notations of Khan and Ugander (2023) to describe it here. Here our goal is to estimate the mean ψ of $Y_1, \ldots, Y_n$, with the observations missing at random. The Bernoulli indicators $R_k, k = 1, \ldots, n$ indicate whether $Y_k, k = 1, \ldots, n$ were observed or not. We assume $R_k \overset{\text{ind}}{\sim} \text{Bernoulli}(p_k)$ for $k = 1, \ldots, n$, which reflects non-response bias in sample surveys. Then, the Horvitz–Thompson and Hájek estimator of $\psi = n^{-1} \sum_{k=1}^{n} Y_k$ are:

$$\hat{S} = \sum_{k=1}^{n} \frac{Y_k R_k}{p_k}, \quad \hat{n} = \sum_{k=1}^{n} \frac{R_k}{p_k} \tag{2.4}$$

$$\hat{\psi}_{HT} = \frac{\hat{S}}{n}, \quad \hat{\psi}_{\text{Hájek}} = \frac{\hat{S}}{\hat{n}}. \tag{2.5}$$

A notable generalization of the aforementioned estimators is the Trotter–Tukey estimator (Trotter and Tukey,

[1] Also called Narain-Horvitz-Thompson estimator after Narain (1951) by Rao et al. (1999); Chauvet (2014).

[2] In a probability proportional to size (PPS) design, where the inclusion probabilities are proportional to an auxiliary variable x_k, they take the form

$$p_k = \frac{n x_k}{\sum_{i=1}^{N} x_i}. \tag{2.1}$$

1956). Recently, Khan and Ugander (2023) rediscovered the Trotter–Tukey estimator as the adaptive normalization (AN) idea by letting the data choose the tuning parameter λ.

$$\hat{\psi}_{TT} = \frac{\hat{S}}{(1-\lambda)n + \lambda\hat{n}},\ \lambda \in \mathbb{R} \quad \text{(Trotter–Tukey)} \tag{2.6}$$

$$\hat{\psi}_{AN} = \frac{\hat{S}}{(1-\hat{\lambda})n + \hat{\lambda}\hat{n}}. \quad \text{(Adaptive Normalization)} \tag{2.7}$$

Furthermore, Khan and Ugander (2023) show that the 'adaptive normalization' idea dating back to Trotter and Tukey (1956) leads to a lower asymptotic variance than both while generalizing these estimators, in particular, $\mathbb{V}(\hat{\psi}_{AN})$ is lower than both $\mathbb{V}(\hat{\psi}_{HT})$ and $\mathbb{V}(\hat{\psi}_{\text{Hájek}})$. Khan and Ugander (2023) also provide empirical evidence that their adaptive normalization scheme leads to a lower mean squared error of IPW estimators in different application areas in average treatment effect estimation and policy learning.

2.1 Horvitz-Thompson vs. Hájek Estimator

Horvitz–Thompson estimators possess many desirable properties: they are unbiased, admissible and consistent. Ramakrishnan (1973) provides a simple proof that the HT estimator is admissible in a class of all unbiased estimators of a finite population total. Isaki and Fuller (1982) proves that they achieve consistency under suitable conditions, such as, specifically requiring boundedness of population and weight sequences, bounded product of inclusion probabilities and second moments, and a variance that is $O(n_t^{-1})$. Delevoye and Sävje (2020) provides conditions such as boundedness of moments for the outcome and inclusion probabilities and weak-design dependence. Delevoye and Sävje (2020) also point out that these conditions might be violated in ill-behaved setting with heavy-tailed outcomes or skewed sampling designs, common in natural experiments, where the practitioner has little control over the design.

It is worth noting here that IPW estimators, in particular the HT estimator, can be looked at as a weighted estimator, where the weights are not model-based, but design-based, as pointed out by several authors from Neyman (1934) to Little (2008). For example, in the HT estimator (2.2), the kth unit is assigned a weight proportional to the inverse of the selection probability as it 'represents the $1/p_k$ units of the population'.

2.1.1 When Is the H'ajek Estimator Preferable?

We follow Särndal et al. (2003, ch. 5) for this discussion. First, note that the two estimators, HT and Hájek, can produce identical results under certain designs. However, if the population total N is unknown, only the Hájek estimator (2.3) can be used. On the other hand, if N is known—as is typically the case in the finite population estimation problem—the Hájek estimator "*is usually the better estimator, despite estimation of an a priori known quantity*" (Särndal et al., 2003). Särndal et al. (2003) supports this by identifying three situations where the weighted mean estimator (Hájek) outperforms the simple unbiased estimator (HT). In particular, Särndal et al. (2003, p. 183) shows that the $\hat{\psi}_{\text{Hájek}}$ estimator achieves lower variance than $\hat{\psi}_{\text{HT}}$ in each of these cases:

(i) When the values of y_k are closer to the population mean ψ, a special case being fixed $y_k = c$ for $k = 1, \ldots, N$.
(ii) When the sample size n_s is variable with equal inclusion probabilities, e.g. the case where $y_k = c$ and $p_k = p_0$ for $k = 1, \ldots, N$, but the sample sizes vary.
(iii) Finally, when the inclusion probabilities p_k are negatively correlated with the y_k values, with large y_k values corresponding to small p_k values and vice versa. It is easy to see that the Hájek estimator is adaptable to high fluctuations in y_k/p_k values that HT will suffer from, leading to high variability. Basu (1988) pointed this out in his famous circus example, recounted in §3.1.

Despite being popular choices, both the Horvitz–Thompson (HT) (Horvitz and Thompson, 1952), and the Hájek estimator (Hájek, 1971) have generated debates and controversies in the recent past. For example, the Horvitz–Thompson estimator where we encounter randomly missing observations and a very high-dimensional parameter space is presented as evidence of weakness of Bayesian method in this problem, in an example due to Wasserman (2004), based on Robins and Ritov (1997). In response, a few authors (Sims, 2010; Ghosh, 2015; Li, 2010; Harmeling and Touissant, 2007) have tried to provide a Bayesian answer by constructing Bayesian estimators that achieve a lower variance than the HT estimator at the cost of admitting a small (and in some cases, vanishing) bias: another example of the bias-variance trade-off.

Stein's Paradox: Here, we argue that this phenomenon is another example of Stein's paradox (James and Stein, 1961; Efron and Morris, 1977; Stigler, 1990) observed in a different context. The original Stein's paradox showed the inadmissibility of the ordinary maximum likelihood estimator which is both unbiased and minimax for normal means in dimensions more than 3 by shrinking each component towards the origin (or a pre-fixed value), thereby borrowing strength from each other. The Stein's paradox and the James–Stein shrinkage estimator has been truly transformative in statistics in both establishing Empirical Bayes' success in high-dimensional inference and inspiring both frequentist shrinkage estimators and Bayesian shrinkage priors in such problems. We argue that the simple HT estimator can be improved upon by borrowing strength by moving from an independence framework to an exchangeable one, exhibiting Stein's shrinkage phenomenon and effects of regularization.

Before we discuss the weak paradoxes, we briefly discuss how the connection between IPW estimators and Monte Carlo sampling, in particular importance sampling, and various improvements such as Riemann sums (Philippe, 1997; Philippe and Robert, 2001) or vertical likelihood integration that applies 'binning and smoothing' using a score-function heurism to choose the weight function (Polson and Scott, 2014; Madrid-Padilla et al., 2018). We then show how the idea of binning and smoothing also improves the HT estimator (Ghosh, 2015) in the apparent weakness example due to Wasserman (2004). We discuss these connections briefly in the next subsection and point out the similarities between estimators employed in survey sampling and Monte Carlo integration and their connections with statistical mechanics.

2.2 Importance Sampling

One way to look at this connection between sampling strategies and integral approximation is to represent the former as a missing data problem. Suppose that our goal is to estimate:

$$\psi = \int_0^1 y(x)dx,$$

which can be thought of as a limiting value of $\psi_n = n^{-1}\sum_{i=1}^n y_i$ as $n \to \infty$, and can be approximated up to any degree of precision by $\psi_N = N^{-1}\sum_{i=1}^N y_i$ for a sufficiently large N. Given a random sample of size $n \ll N$, with sample probabilities π attached to y_i, we can view estimating θ as a problem of estimating the population quantity ψ_N by ψ_n. The usual importance sampling estimator in this case is akin to using $\mathbb{E}_\pi[\sum_{i=1}^n y_i/\pi_i] = \psi$, the usual HT estimator. Just like the HT estimator, the variance of importance sampling could blow up for poor choices of $g(\cdot)$, while the bias may shrink to zero. As we show below, these connections have been exploited by several authors to propose alternative importance sampling strategies.

Given a sample $(y_1, \ldots, y_n)$, from either the density f corresponding to F itself or a suitable proposal density g, the usual importance sampling estimates ψ by the empirical average:

$$\hat{\psi}_{IS} = \frac{1}{n}\sum_{i=1}^n \frac{l(y_i)f(y_i)}{g(y_i)}. \tag{2.8}$$

If the underlying probability measure is easy to sample from, *i.e.,* if we can afford $f = g$, the above will reduce to the empirical average $\hat{\psi} = n^{-1}\sum_{i=1}^n l(y_i)$, which by the Law of Large Numbers, converge to the true value of ψ at $O(n^{-1})$ rate. It is worth pointing out that replacing the empirical average for the naïve importance sampling estimator in (2.8) by a Riemann sum estimator provides a remarkable improvement in stability and convergence as demonstrated in (Philippe, 1997; Philippe and Robert, 2001) or (Yakowitz et al., 1978). Recently, Datta and Polson (2025) showed that combining Riemann sum estimators with nested sampling ideas can yield sharper convergence rates (up to $O(n^{-4})$) for marginal likelihood estimation in Bayesian problems. We refer the reader to Datta and Polson (2025) for a more detailed discussion of vertical likelihood, and how it connects importance sampling with nested inference. This framework also offers a unifying perspective on several sampling-based approaches that aim to improve the stability of likelihood-weighted estimators.

2.2.1 A Bayesian Perspective

We conclude the discussion of evidence estimation with a pertinent and profound point raised by Diaconis (1988) in support of Bayesian approaches for probabilistic numerical integration. Diaconis (1988) starts with a algebraic function:

$$f(x) = \exp\left\{\cosh\left(\frac{x + 2x^2 + \cos x}{3 + \sin x^3}\right)\right\}, \quad f : [0,1] \mapsto \mathbb{R},$$

for which we want $\int_0^1 f(x)dx$ and asks "*What does it mean to 'know' a function?.*" In such a situation, where we might know a few properties of f and not others, a Bayesian approach seems natural, where one starts with a prior on $\mathcal{C}[0,1]$, the space of continuous functions, and estimate the integral using Bayes' rule. Diaconis (1988) shows that Brownian motion, the 'easiest' prior on $\mathcal{C}[0,1]$, yields a linear interpolation leading to the trapezoid rule. Diaconis (1988) then shows that a host of well-known methods can be recovered as Bayesian estimates. This key question is revisited in (Owen, 2019) where he asks what does it mean to know the error $\Delta = |\hat{\psi} - \psi|$ and discusses advantages of Bayesian approaches over classical methods. Owen (2019) argues in favour of Bayes methods in difficult problems where extreme cost of function evaluation or skewness or heavy-tailed properties of $l(x)$ or unavailability of central limit theorem exposes the weakness of classical methods.

2.3 Average Treatment Effect Estimation

A natural application of IPW estimator is average treatment effect (ATE) estimation in potential outcomes causal inference framework as pointed out in (Delevoye and Sävje, 2020; Khan and Ugander, 2023). We refer the readers to the excellent references in (Cunningham, 2021; Rubin, 1974; Imbens, 2004) for in-depth coverage and historical backgrounds. Here we measure the difference in potential outcomes observed over time points $t_2 > t_1$, where only one of the two potential outcomes are observed for any unit. Using Khan and Ugander (2023)'s notation: we have the triplets $(Y_k(0), Y_k(1), p_k)$ for $k = 1, \ldots, n$, where we observe $Y_k(I_k)$ for $I_k \sim$ Bernoulli(p_k). and the parameter of interest is $ATE = \tau = \mathbb{E}[Y_k(1) - Y_k(0)]$ from the observations available to us. The IPW estimators are employed here to estimate the two population means $\psi_1 = \mathbb{E}[Y_k(1)]$ and $\psi_0 = \mathbb{E}[Y_k(0)]$ separately and estimating τ by $\hat{\tau} = \hat{\psi}_1 - \hat{\psi}_0$. Estimating the population means ψ_j, $j = 1, 2$ turns ATE into a survey sampling problem and one can use all the estimators: HT, Hájek or the various improvements such as

the adaptive IPW estimator by Khan and Ugander (2023) based on the Trotter–Tukey idea (Trotter and Tukey, 1956) here. Khan and Ugander (2023) further develop an adaptive estimator by minimizing the variance of between-group differences and show that both the separate AIPW and the joint AIPW estimators achieve lower mean squared errors compared to the usual HT and Hájek.

3. WEAK PARADOXES IN RATIO ESTIMATION PROBLEM

3.1 Basu (1988)'s Example

Example 1. Basu's circus example: The famous circus example, due to (Basu, 1988), shows that the HT estimator (2.2) could lead to absurd estimates in some situations. Here, we imagine a circus owner, trying to estimate the total weight of his 50 elephants (say Y), picks a representative elephant from his herd (*Sambo*) and multiply his weight by 50. But, then a circus statistician, appalled by this estimate, devices a plan where *Sambo* is picked with probability 99/100 and each of the remaining with 1/4900. Unfortunately, now the HT estimate is Sambo's weight $\times$ 100/99, a serious underestimate, and moreover, if the owner picks *Jumbo*, the biggest in the herd, the HT estimate is Jumbo's weight $\times$ 4900, an absurdity.

Discussing Basu (1988), Hájek (1971) points out that the HT estimator's *"usefulness is increased in connection with ratio estimation"* and proposed an estimator in presence of auxiliary information A_k, related to Y_k and with known total:

$$\hat{Y}_{\text{Hájek}} = \sum_{k=1}^{n} A_k \times \frac{\sum_{k=1}^{n} \frac{Y_k}{p_k}}{\sum_{k=1}^{n} \frac{A_k}{p_k}}, \text{ where } Y_k \propto A_k, k = 1, \ldots, n, \tag{3.1}$$

which would not be affected in this example like the standard HT estimator. The Hájek IPW estimate for the population mean in (2.3) can be derived from the special case $A_k \equiv 1, \forall k$.

Although the circus example was intended to be a pathological example, it provides at least two useful insights. First, weighted estimators can lead to nonsensical answers despite having nice large sample properties (Little, 2008). Second, problems of similar nature occur in importance sampling or Monte Carlo estimation of the marginal likelihood where empirical averages or unbiased estimator could have high or infinite variance (Li, 2010; Raftery et al., 2006). As pointed out in (Datta and Polson, 2025), strategies like Riemann sum (Philippe, 1997; Philippe and Robert, 2001), or careful choice of weight functions like vertical likelihood (Polson and Scott, 2014) or nested sampling (Skilling, 2006), or using ratio estimators (Firth, 2011), or adaptive normalization (Khan and Ugander, 2023) can resolve these issues. In particular, the same trick of ordering the draws and applying a Riemann-sum type approach is the key to avoid falling into examples like Basu's circus.

3.2 Wasserman (2004)'s Example

We first re-state the example which is itself a simplification of an example from (Robins and Ritov, 1997), in a similar spirit. We consider IID samples (Y_i, X_i, R_i), $i = 1, \ldots, B$ such that Y_i's are generated as a mixture of Bernoulli distributions with individual parameters indexed by the component label X_i, and the 'missingness' indicator R_i denoting whether Y_i was observed or not. Let the 'success' probabilities associated to R_i be known constants p_{X_i} satisfying:

$$0 < \delta \leq p_j \leq 1 - \delta < 1, \; j = 1, \ldots, B. \tag{3.2}$$

Note that this strong condition on p_j's in (3.2) ensures that the HT estimator will not lead to absurd answers like Basu's paradox stated earlier. The hierarchical model for each draw (Y_i, X_i, R_i), $i = 1, \ldots, n$, is:

$$X_i \sim \mathcal{U}(1, \ldots, B) \tag{3.3}$$

$$[R_i \mid X_i = x_i] \sim \text{Bernoulli}(p_{x_i}) \tag{3.4}$$

$$[Y_i \mid R_i, X_i = x_i] \sim \begin{cases} \text{unobserved if } R_i = 0 \\ \text{Bernoulli}(\theta_{x_i}) \text{ if } R_i = 1. \end{cases} \tag{3.5}$$

The parameter of interest is the average $\psi = (1/B) \sum_{b=1}^{B} \theta_b$. Wasserman (2004) argues that since the likelihood has little information on most θ_j's and the known constants B and p_j's drop from the likelihood, Bayes' estimates for ψ are going to be poor. On the other hand, the HT estimate:

$$\hat{\psi}_{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{p_{X_i}} \tag{3.6}$$

is easily seen to be unbiased, given $\mathbb{E}(R_i Y_i / p_{X_i}) = \mathbb{E}\{\mathbb{E}(R_i \mid Y_i, X_i) \cdot \mathbb{E}(Y_i \mid X_i)/p_{X_i}\} = \mathbb{E}(\mathbb{E}(Y_i \mid X_i)) = \psi$ since $\mathbb{E}(R_i \mid Y_i, X_i) = p_{X_i}$ by construction. The HT estimate will also satisfy $\mathbb{V}(\hat{\psi}_{HT}) \leq 1/n\delta^2$, using Hoeffding's inequality. We would like to note here that the *assumption* (3.2) that the selection probabilities are between δ and $1 - \delta$ is exploited explicitly in this asymptotic result, *i.e.,* the HT estimator might have infinite variance as δ goes to zero, making the assumptions crucial.

This example of apparent weakness in Bayesian paradigm and the concluding remarks by Wasserman (2004)[3] has since been a source of debate and elicited response from Bayesian community (Li, 2010; Sims, 2010; Harmeling and Touissant, 2007; Ghosh, 2015) which we review briefly below.

3.3 Bayesian Solutions for Wasserman (2004)'s Problem

We review the existing Bayesian resolutions for the Robins-Ritov-Wasserman problem using Bayesian ideas and argue that, in some special cases, the improvement in accuracy of Bayes' estimate is attained via exploiting the 'borrowing strength' phenomenon in Stein's shrinkage.

[3] "*Bayesians are slaves to the likelihood function. When the likelihood goes awry, so will Bayesian inference.*" (Wasserman, 2004, pp. 189)

3.3.1 Full Bayes Solution (Li, 2010)

Li (2010) provides a simple Bayes estimator by assuming that $\theta_1, \ldots, \theta_B$ are exchangeable, not independent (see Fig. 1). Li (2010) estimates ψ by augmenting (3.5) by with Beta priors for $\theta_1, \ldots, \theta_B$ and a further hyperprior on the mean of these Beta priors, as follows:

$$\theta_b \mid \alpha_T, \psi \overset{\text{ind}}{\sim} \text{Beta}(\alpha_T \psi, \alpha_T(1-\psi_0)) \tag{3.7}$$

$$(\psi, \alpha_T \mid \alpha_F) \sim \text{Beta}(\alpha_F, \alpha_F) \times \pi(\alpha_T), \tag{3.8}$$

where, ψ is the mean of θ, and the shape parameters α_T and α_F control the width of range of $\boldsymbol{\theta}$ and the concentration of ψ.

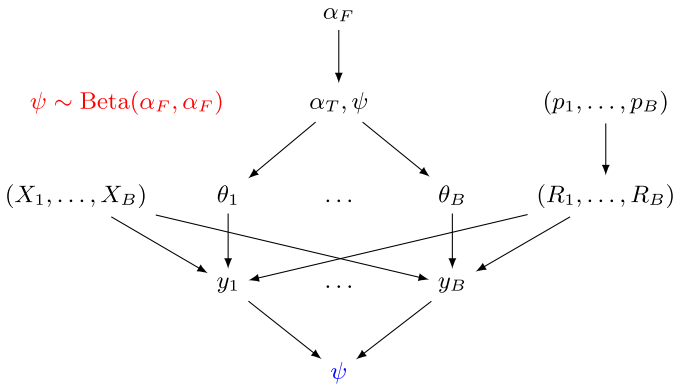


Figure 1: Hierarchical model for the Robins-Ritov-Wasserman problem.

Using this model, Li (2010) derives a posterior mean estimate of ψ as:

$$\hat{\psi}_{Li} = \frac{\sum_{i=1}^{n} R_i Y_i + \alpha_F}{\sum_{i=1}^{n} R_i + 2\alpha_F} \tag{3.9}$$

Through extensive numerical simulation, Li (2010) shows that this estimator achieves a smaller mean squared error compared to the HT estimator in Wasserman (2004). Li (2010) also argues that the variance of HT estimator, given by:

$$\mathbb{V}(\hat{\psi}_{HT}) = \frac{1}{n}\left\{\frac{1}{B}\sum_{b=1}^{B}\frac{\theta_b}{p_b} - \psi^2\right\}, \tag{3.10}$$

will lead to inflated variance for small p_b values, necessitating the bounds $\delta \leq p_j \leq 1-\delta$ in (3.2), but the variance for Li's Bayes estimator remains unaffected and does not need this extra restriction. On the other hand, Li's estimator (3.9) is prone to bias if the missingness mechanism p_X and the parameter for observed outcomes θ_X are correlated, *i.e.,* if the missingness is missing-at-random (MAR) and not missing-completely-at-random (MCAR). We shall formalize this via the derived approximation for the mean squared error of the Bayes estimator in our theoretical properties section. Linero (2024) argues that the 'seemingly innocuous' priors like the one used by Li (2010) encode what is effectively a-priori knowledge that the amount of selection bias is minimal.

3.3.2 Binning and Smoothing (Ghosh, 2015)

Ghosh (2015) provides another estimator by reducing the dimension of $(p_1, \ldots, p_B)$ by clubbing them into k ($k \ll B$) groups by utilizing the boundedness assumption (3.2). Based on this idea, we define the general binned-smoothed estimator as follows.

Definition 2. Given a fixed $\delta \in (0,1)$ satisfying (3.2), we divide the whole range $(\delta, 1-\delta) \ni p_b, b = 1, \ldots, B$ into k sub-intervals $\{\delta_0 = \delta, \delta_1, \ldots, \delta_k = 1-\delta\}$ such that $\delta_{i+1} - \delta_i = (1-2\delta)/(k-1)$. We order the observed p_{X_i}'s into increasing order and define $\tilde{p}_j$ to be the mean of the p_{X_i} values in the jth partition consisting of n_j points, i.e.

$$\tilde{p}_j = \begin{cases} \frac{1}{n_j}\sum_{i: p_{X_i} \in (\delta_j, \delta_{j+1})} p_{X_i}, & \text{if } n_j > 0 \\ (\delta_j + \delta_{j+1})/2 & \text{otherwise} \end{cases}$$

Then, the *binned-smoothed* estimator, based on (Ghosh, 2015) is:

$$\hat{\psi}_{BS:HT} = \sum_{j=1}^{k} \frac{n_j}{n}\frac{1}{n_j}\frac{(\sum_{i=1}^{n_j} R_i Y_i)}{\tilde{p}_j}. \tag{3.11}$$

where n_j is the number of p_{X_i}'s falling into the jth class. Note that, if the sample size $n \gg k$, the number of partitions, the number of p_{X_i}'s in any interval, i.e. n_j values would typically be non-zero, and in the unlikely case $n_j = 0$ for any $j = 1, \ldots, k$, the corresponding term will not contribute to the numerator in (3.11).

Ghosh (2015)'s estimator performs at least as well as the HT estimator in small simulation studies. The idea of 'binning and smoothing' can be applied to other inverse probability weighted estimators such as Hájek (2.3). The binned-smoothed Hájek estimator would take the form:

$$\hat{\psi}_{BS:Hajek} = \frac{\sum_{j=1}^{k}\{(\sum_{i=1}^{n_j} R_i Y_i)/\tilde{p}_j\}}{\sum_{j=1}^{k}\{(\sum_{i=1}^{n_j} R_i)/\tilde{p}_j\}}, \tag{3.12}$$

where n_j and $\tilde{p}_j$'s are as used in (3.11) before.

3.3.3 Bayesian Sieve (Sims, 2012)

We note that the idea of grouping the probabilities attached to the missingness indicators R_i's, *i.e.,* clubbing p_{X_i}'s into p_j's was also proposed in (Sims, 2012). Sims (2012) argues that estimators better than the HT estimator (3.6) can be constructed using Bayesian approach if one acknowledges the existence of infinite dimensional unknown parameter in the model, e.g. assuming $p(b) : 1, \ldots, B \mapsto (0,1)$ is known but not the conditional distribution of $[\theta \mid p]$. Sims

suggested to break up the range of the known $p(\cdot)$ into k segments, such that $\theta(b)$ and $p(b)$ are independent in each segment, and estimate the unknown $\theta(b)$ via a step function, with a constant for each segment. To complete Sim's Bayesian sieve, one needs to estimate the probability distribution on p_j's induced by the partition, which will converge to a Dirichlet distribution for a large sample, making analytical calculation for the posterior mean of ψ feasible.

3.3.4 Gaussian Likelihood (Harmeling and Touissant, 2007)

Harmeling and Touissant (2007) consider a slightly modified version of the Wasserman's model (3.5). Instead of a Bernoulli, they consider:

$$X_i \sim \mathcal{U}\{1,\ldots,B\},\ Y_i \mid \Theta_{X_i} \sim \mathcal{N}(\theta_{X_i},1),\ 1 \le i \le n,$$

and $\mathcal{N}(\mu,1)$ prior for each θ_b, $b=1,\ldots,B$, and a $\mathcal{N}_\mu(0,\sigma)$ hyper-prior on μ. The maximum likelihood estimate and the posterior mean Bayes' estimates for $B\to\infty$ are given by

$$\hat{\psi}_{\text{MLE}} = \left(\sum_{i=1}^n R_i\right)^{-1}\left(\sum_{i=1}^n R_iY_i\right), \quad \text{and}$$
$$\hat{\psi}_{\text{Bayes}} = \left(2/\sigma + \sum_{i=1}^n R_i\right)^{-1}\left(\sum_{i=1}^n R_iY_i\right),$$

respectively. Although these estimators are not directly comparable to (3.9), Harmeling and Touissant (2007)'s likelihood-based estimators also achieve a lower MSE compared to the HT estimator, as expected.

We also note that the idea of binning and smoothing p_j's is also connected to importance sampling ideas in subsection 2.2, in particular, the trapezoidal rule (also called weighted Monte Carlo) by Yakowitz et al. (1978), or its generalizations, called the Riemann summation approach Philippe (1997); Philippe and Robert (2001). These trapezoidal rules based on ordered samples reduces the variance drastically and achieves faster convergence and better stability. Along these lines, a later development, called the nested sampling approach by Skilling (2006) also relies on "dividing the unit prior mass into tiny elements, and sorting them by likelihood." For a comprehensive discussion of the different strategies for variance reduction for the evidence estimation problem, we refer the readers to (Polson and Scott, 2014; Datta and Polson, 2025).

3.4 Properties of Li's Bayesian Estimator

While a closed form analytical expression for the variance of the Li's estimator (3.9) is difficult, we can use the linearization strategy aka the delta theorem to derive an approximation for the mean and variance of the Bayes estimator (3.9). Doing so, we provide a proof and a closer look at an assertion in Ritov et al. (2014) that Li's estimator is consistent only if $\mathbb{E}(Y \mid R=1) = \mathbb{E}(Y)$, and illustrate this phenomenon via simulation studies.

Recall that for a pair of random variables X, Y with mean μ_X and μ_Y and variances $\mathbb{V}(X)$, $\mathbb{V}(Y)$ and covariance $\text{Cov}(X,Y)$, we have the following approximations:

$$\mathbb{E}(X/Y) \approx \mu_X/\mu_Y \tag{3.13}$$
$$\mathbb{V}(X/Y) \approx \left(\frac{\mu_X}{\mu_Y}\right)\left(\frac{\mathbb{V}(X)}{\mu_X^2} + \frac{\mathbb{V}(Y)}{\mu_Y^2} - 2\frac{\text{Cov}(X,Y)}{\mu_X\mu_Y}\right) \tag{3.14}$$

To calculate the approximate mean and variance for the Li's posterior mean estimator (3.9), we first calculate the mean and variances for the numerator and denominator separately as follows. Recall once again that the posterior mean estimator is given by $\hat{\psi}_{Li} = (\sum_i R_iY_i + 1)/(\sum_i R_i + 2)$ assuming $\alpha_F = 1$, i.e., a Beta$(1,1)$ or Uniform prior on ψ but the exact choice of shape parameter α_F does not influence the asymptotic nature of the variance. For notational convenience, we define the following quantities:

$$\bar{\theta} = \frac{1}{B}\sum_{b=1}^B \theta_b \doteq \psi, \quad \bar{p} = \frac{1}{B}\sum_{b=1}^B p_b, \quad \bar{\theta.p} = \frac{1}{B}\sum_{b=1}^B \theta_b p_b,$$
$$\sigma_{\theta,p} = \bar{\theta.p} - \bar{p}\bar{\theta} = \bar{\theta.p} - \bar{p}\psi.$$

First, the expectation for the numerator and denominator follows from applying iterated expectations as:

$$\begin{aligned}\mathbb{E}\left(\sum_i R_iY_i + 1\right) &= \mathbb{E}_X\left\{\sum_{i=1}^n \mathbb{E}(R_i \mid X_i)\mathbb{E}(Y_i \mid X_i)\right\} + 1 \\ &= \mathbb{E}_X\left\{\sum_{i=1}^n \theta_{X_i}p_{X_i}\right\} + 1 \\ &= n\bar{\theta.p} + 1 = n\psi\bar{p} + n\sigma_{\theta,p} + 1 \qquad (3.15)\end{aligned}$$
$$\mathbb{E}\left(\sum_i R_i + 2\right) = n\bar{p} + 2. \tag{3.16}$$

Hence, the expectation for the Bayes' estimator can be approximated by taking the ratio of right hand sides from (3.15) and (3.16) as:

$$\mathbb{E}(\hat{\psi}_{Li}) \approx \frac{n\psi\bar{p} + n\sigma_{\theta,p} + 1}{n\bar{p} + 2} \to \psi + \sigma_{\theta,p}/\bar{p},\ \text{as } n\to\infty,$$

showing the $\hat{\psi}_{Li}$ is asymptotically unbiased if $\sigma_{\theta,p} = 0$, i.e., if $\theta_X \perp p_X \mid X$, or equivalently if $\mathbb{E}(Y \mid R=1) = \mathbb{E}(Y)$, as claimed by (Ritov et al., 2014). Now, the variances for the numerator and denominator can be calculated using the conditional variance identity. The expressions are given below.

$$\begin{aligned}\mathbb{V}\left(\sum_i R_i + 2\right) &= \sum_i \mathbb{V}(R_i) \\ &= \sum_i \mathbb{E}[\mathbb{V}(R_i \mid X_i))] + \mathbb{V}[\mathbb{E}(R_i \mid X_i)]\end{aligned}$$

$$= \sum_i [\mathbb{E}(p_{x_i}(1-p_{X_i})) + \mathbb{V}(p_{X_i})] = n(\bar{p} - \bar{p}^2). \tag{3.17}$$

Similarly,

$$\mathbb{V}\left(\sum_i R_i Y_i + 1\right) = \sum_i \mathbb{V}(R_i Y_i) = n(\bar{\theta.p} - \bar{\theta.p^2}),$$
$$\text{where } \bar{\theta.p} = \frac{1}{B}\sum_{b=1}^{B} \theta_b p_b. \tag{3.18}$$

Finally, the covariance term is:

$$\begin{aligned}\text{Cov}\left(\sum_i R_i Y_i + 1,\ \sum_i R_i + 2\right) &= \text{Cov}\left(\sum_i R_i Y_i,\ \sum_i R_i\right) \\ &= \sum_i \text{Cov}(R_i Y_i,\ R_i) \\ &= n\bar{\theta.p}\ (1-\bar{p}). \end{aligned} \tag{3.19}$$

Putting everything together from (3.15), (3.16), (3.18), (3.17) and (3.19), we get the following formula for approximate variance for the Bayes estimator as:

$$\mathbb{V}(\hat{\psi}_{Li}) \approx \left(\frac{n\bar{\theta.p}+1}{n\bar{p}+2}\right)^2 \times \left[\frac{n(\bar{\theta.p} - \bar{\theta.p^2})}{(n\bar{\theta.p}+1)^2} + \frac{n(\bar{p}-\bar{p}^2)}{(n\bar{p}+2)^2} - \frac{2n\bar{\theta.p}(1-\bar{p})}{(n\bar{p}+2)(n\bar{\theta.p}+1)}\right]. \tag{3.20}$$

A couple of immediate implications of the formula in (3.20) are as follows.

(i) Li's Bayesian estimator is consistent if θ_{X_i} and p_{X_i} are uncorrelated, as $\mathbb{E}(\hat{\psi}_{Li}) \to \psi$ and $\mathbb{V}(\hat{\psi}_{Li}) = O(1/n) \to 0$ as the sample size $n \to \infty$.

(ii) Unlike the variance of the HT estimator, given in (3.10), the variance of the Li's Bayes estimator does not have the individual p_b terms in the denominator and will not inflate for very small p_b values. In other words, the restriction (3.2) is not needed for the Bayesian solution.

4. SIMULATION EXAMPLES

4.1 Comparing IPW Estimators for Wasserman's Example

4.1.1 Missing Completely at Random

We compare the estimation performance for four different candidate estimators for the Robins-Ritov-Wasserman's problem. The candidates are: the original Horvitz–Thompson (2.2), Hájek (2.3), the Li's estimator, *i.e.,* the Bayes posterior mean under a Beta hyperprior (3.9) and Ghosh (2015)'s binning and smoothing idea applied to the Hájek estimator. We choose the bounds for p_i's $\delta = 0.01$, and parameter space dimension $B = 1000$, *i.e.,* the p_j's in our experiment are B equidistant grid-points in $[\delta, 1-\delta]$. We vary the support of generative distribution for $\theta \in \mathcal{U}[a,b]$ to four different values, viz. $[a,b] \in \{[0.1,0.9],[0.1,0.4],[0.35,0.65],[0.6,0.9]\}$. Finally, we take sample size $n = 100$, as was originally intended in the RRW example to make it analogous to a high-dimensional problem with $B \gg n$.

Table 1 and Fig. 2 shows the mean squared errors calculated over 100 replicates, and shows the following: (1) the Li's estimator leads to the lowest mean squared error over replicates, (2) the Horvitz–Thompson estimator performs the worst across all situations considered and finally, (3) the Hájek estimator and the Binning-Smoothing estimator (Ghosh, 2015) achieves very similar performance and generally occupies a middle ground between the Bayes' and the Horvitz–Thompson in terms of achieved mean-squared error.

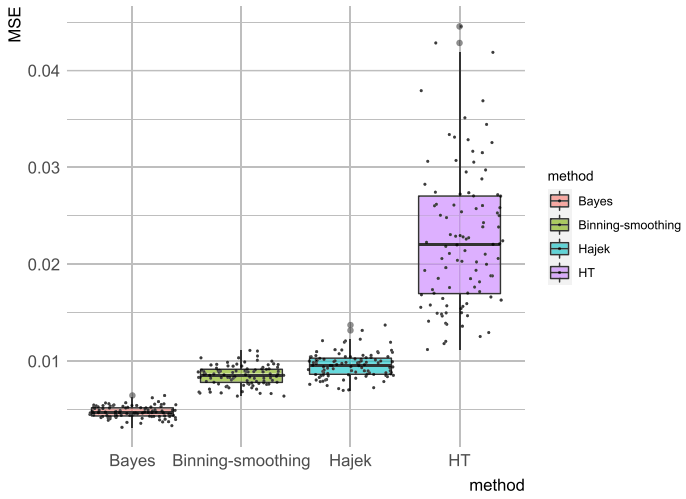
Table 1. Mean squared error comparison between the four candiate estimators: Horvitz–Thompson (2.2), Hájek (2.3), the Li's Bayesian estimator (3.9) and Ghosh (2015)'s estimator for different values of $[a,b]$, where $\theta \in [a,b]$. The numbers on the table are reported after multiplying the MSE by 10^2 for better comparison and the column winner are denoted by boldfaced entries.

	[0.6, 0.9]		[0.1, 0.9]	
method	mean	sd	mean	sd
Bayes' (Li's)	**0.37198**	0.05093	**0.47541**	0.05980
Binning-smoothing	0.63991	0.08583	0.85225	0.10204
Hájek	0.71787	0.11643	0.95824	0.12818
HT	3.09007	0.83348	2.27093	0.72187
	[0.1, 0.4]		[0.35, 0.65]	
method	mean	sd	mean	sd
Bayes' (Li's)	**0.35743**	0.04758	**0.47329**	0.06272
Binning-smoothing	0.64379	0.08114	0.86459	0.10523
Hájek	0.73526	0.13494	0.96710	0.13523
HT	1.17543	0.51219	2.22213	0.69317

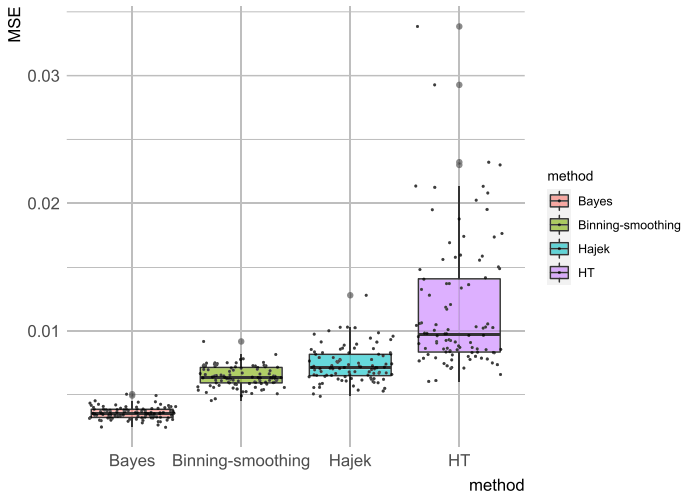
For a single replication with 1,000 evaluations, Fig. 3 shows the bias and variance for the four candidate estimators. Fig. 3 demonstrates the nature of variance reduction by Bayes' estimator and the Binning-smoothing idea in estimating $\psi = \mathbb{E}(\boldsymbol{\theta})$ for different ranges of θ, without any significant increase in bias.

4.1.2 Missing at Random

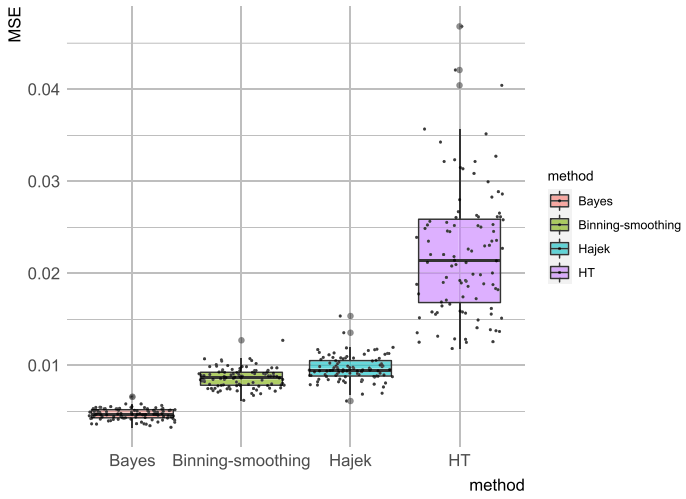
As discussed earlier, Li's estimator can be biased when there is a non-zero correlation between θ_X and p_X, *e.g.,* when the missingness might arise due to confounding. We show an example here to illustrate the effect of this situation, and compare the four candidates. We take $\delta = 0.1$, *i.e.,* p_b's to be equidistant points in $[0.1, 0.9]$, and instead of θ_b's to be uniformly distributed in $[a,b]$, we take:



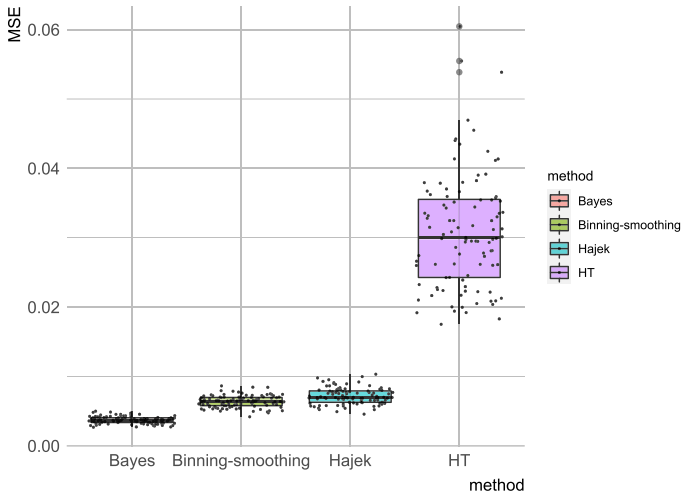
(a) $\theta_b \sim \mathcal{U}(0.1, 0.9)$



(b) $\theta_b \sim \mathcal{U}(0.1, 0.4)$



(c) $\theta_b \sim \mathcal{U}(0.35, 0.65)$



(d) $\theta_b \sim \mathcal{U}(0.6, 0.9)$

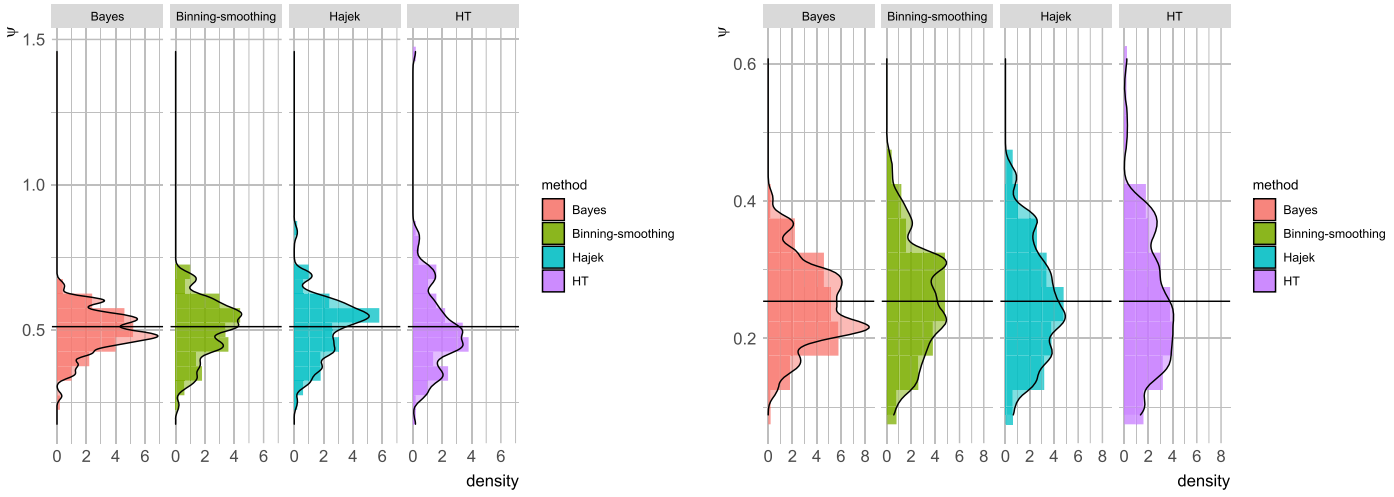
Figure 2: Mean square error comparison for Horvitz-Thompson, Hájek and the Li's estimator (3.9) for different generative distributions of $\theta_1, \ldots, \theta_B$, and fixed values of n, B and δ. In each case, Bayes' estimator has the lowest MSE, followed by Hájek, and HT has the largest MSE.

$$\theta_b = \gamma \times p_b + (1 - \gamma) \times U_b,$$
$$\text{where } U_b \sim \mathcal{U}(a, b),\ \gamma = 0.5, b = 1, \ldots, B.$$

The remaining set-up is same as before. Clearly, in this case the true value of $\psi = \mathbb{E}(\theta) = 0.5$. Figure 4a shows that the Li's estimator has an upward bias, as expected, because of the positive correlation between θ and p, but the Hájek and Binned-Smoothed estimators are unaffected. On the other hand, the Binned-Smoothed is significantly better than all the other candidates (HT, Hájek and Li's) in terms of MSE (Fig. 4b). This shows that the Li's posterior mean estimator can be biased when there is confounding and one needs to be careful when using it. The binned-smoothed method, on the other hand, performs well irrespective of the correlation between θ_X and p_X, *i.e.,* for both MAR and MCAR-type missingness.

5. DISCUSSION

Our main goal in this paper was to shed light on various aspects of the family of inverse probability weight estimators, discuss its weakness and strengths and highlight



(a) $\theta_b \sim \mathcal{U}(0.1, 0.9)$

(b) $\theta_b \sim \mathcal{U}(0.1, 0.4)$

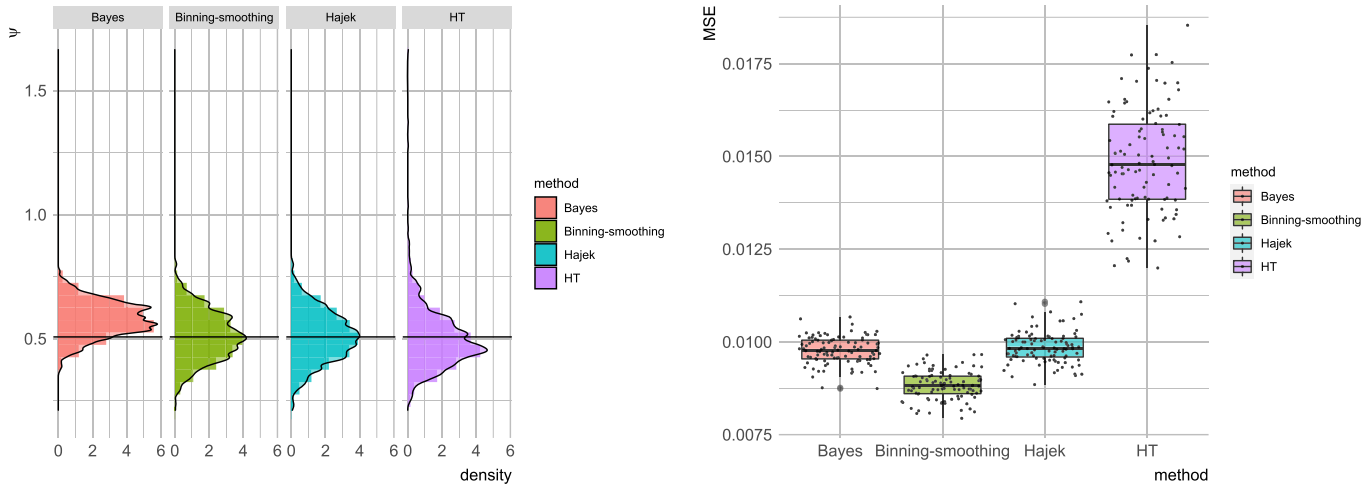
(c) $\theta_b \sim \mathcal{U}(0.35, 0.65)$

(d) $\theta_b \sim \mathcal{U}(0.6, 0.9)$

Figure 3: Histogram and Kernel Density plots showing the distribution of Horvitz–Thompson, Hájek and the Bayes estimators for different generative distributions of $\theta_1, \ldots, \theta_B$, and fixed values of n, B and δ.

the connections between survey sampling and Monte Carlo integration via these popular tool pervasive in Statistical literature. For survey sampling, we consider some of the 'weak paradoxes' (Ghosh, 2015) that arise from using inverse probability estimators using the well-known examples from Wasserman (2004) and Basu (1988), and review the merits and demerits of popular IPW estimators. In particular, we show that the hierarchical Bayes' estimator due to (Li, 2010) leads to robustness against pathological situations but admits bias in presence of confounding. We provide sufficient conditions for consistency of the Li's posterior mean estimate in Wasserman's example and show that, under certain conditions, both the Li's estimator and the binning-smoothing idea (Ghosh, 2015) achieves lower variation in mean squared errors. We then highlight the analogous tools in Monte Carlo integration, revisiting a few earlier works (e.g., Hesterberg, 1988), where the Horvitz–Thompson estimator is likened to the naïve importance sampling, and much like survey sampling, self-normalized weights or using control variates lead to better estimators. We conclude with a brief discussion of applications in the context of causal inference, a key use of IPW estimators, and a few possible directions for future work.

A possible future direction, borrowing from (Kong et al.,



(a) $\theta_b = \gamma \times p_b + (1-\gamma) \times \mathcal{U}(0,1)$ (b) $\theta_b = \gamma \times p_b + (1-\gamma) \times \mathcal{U}(0,1)$

Figure 4: Histogram, Kernel density plots and MSE comparison for Horvitz–Thompson, Hájek and the Bayes estimators for the missing-at-random scenario.

2003) is to use the general semiparametric models and the associated estimators or computational algorithm for $k > 1$ in the context of survey sampling that would allow one to combine data from one or more surveys that use different but known inclusion probabilities. We conjecture that this might lead to improved estimators for average treatment effect estimation, which we have briefly described in section 2.3.

In the context of causal inference, a second possible direction for future research is comparing a suitable modification of the Bayes estimator in (3.9) for the ATE problem with that of the adaptive estimators developed by Khan and Ugander (2023) and both empirically and theoretically to investigate consistency properties. It will be also worthwhile to consider the semiparametric model in (Kong et al., 2003) for ATE and compare the results of simultaneous estimation using MLE with these candidates.

Finally, the problems presented in (Wasserman, 2004) and discussed in (Sims, 2007, 2010) are inherently high-dimensional in nature, where sparsity is pervasive. There is a large and growing literature on promoting sparsity in causal inference, *e.g.,* in presence of a large number of confounders or baseline variables while estimating the effect of an exposure on an outcome. We refer the readers to Shortreed and Ertefaie (2017); Wang et al. (2012, 2015); Kim et al. (2022) for recent advances in this area. In our simple focal example, the high-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_B)$ or $\mathbf{p} = (p_1, \ldots, p_B)$ could exhibit sparsity with a small ℓ_0 norm or other notions of sparsity. To handle sparsity in higher dimensions while maintaining tail-robustness and accuracy, the state-of-the-art Bayesian solution would be augmenting (3.5) with global-local shrinkage priors (Polson and Scott, 2010, 2012; Bhadra et al., 2019), *i.e.,*

$$\begin{aligned}
\theta_b \mid \alpha_T, \psi &\overset{\text{ind}}{\sim} f_{\text{HS}}(\theta; \alpha_T \psi,\ \alpha_T(1-\psi),\ \tau) \\
f_{\text{HS}}(\theta; a, b, \tau) &= \theta^{a-1}(1-\theta)^{b-1}\{1-(1-\theta)\tau^2\}^{-(a+b)}, \\
(\psi, \alpha_T \mid \alpha_F) &\sim \text{Beta}(\alpha_F, \alpha_F) \times \pi(\alpha_T), \\
\theta, \psi &\in [0,1].
\end{aligned}$$

Here, as before, ψ is the mean of θ, α_T and α_F are the and the shape parameters for $\boldsymbol{\theta}$ and ψ respectively and τ is a global shrinkage parameter adjusting to sparsity. Such Bayesian regularization priors have been used successfully for treatment effect estimation with more control variates than observations (Hahn et al., 2018). Designing built-in sparsity priors that can also incorporate selection mechanism for confounder selection is an interesting problem that we plan to address in a future endeavor.

ACKNOWLEDGEMENTS

The first author was inspired to study this problem after attending a talk by Professor Christopher Sims, who passed away on March 14, 2026. We dedicate this paper to his memory.

FUNDING

Dr. Datta acknowledges support from the National Science Foundation (NSF DMS-2015460 and NSF CAREER 2443282).

Accepted 3 February 2026

REFERENCES

Basu, D. *Statistical information and likelihood: a collection of critical essays by Dr. D. Basu.* (J. K. Ghosh, ed.) *Lecture Notes in Statistics* **45**. Springer Science & Business Media (1988). MR0953081

Bhadra, A., Datta, J., Polson, N. G. and Willard, B. Lasso meets horseshoe. *Statistical Science* **34**(3) 405–427 (2019). https://doi.org/10.1214/19-STS700. MR4017521

Chauvet, G. A note on the consistency of the narain-horvitz-thompson estimator (2014). *arXiv preprint arXiv:1412.2887*.

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J., Sedransk, J. and Thompson, M. Approaches to improving survey-weighted estimates. *Statistical Science* **32**(2) 227–248 (2017). https://doi.org/10.1214/17-STS609. MR3648957

Chopin, N. and Robert, C. P. Properties of nested sampling. *Biometrika* **97**(3) 741–755 (2010). https://doi.org/10.1093/biomet/asq021. MR2672495

Cunningham, S. Causal inference. In *Causal Inference*. Yale University Press (2021).

Datta, J. and Polson, N. G. Quantile importance sampling (2025). *arXiv preprint arXiv:2305.03158*. https://doi.org/10.1214/25-bjps638. MR5007225

Delevoye, A. and Säve, F. Consistency of the horvitz–thompson estimator under general sampling and experimental designs. *Journal of Statistical Planning and Inference* **207**. 190–197 (2020). https://doi.org/10.1016/j.jspi.2019.12.002. MR4066130

Diaconis, P. Bayesian numerical analysis. *Statistical decision theory and related topics IV* **1**. 163–175 (1988). MR0927099

Efron, B. and Morris, C. Stein's paradox in statistics. *Scientific American* **236**(5) 119–127 (1977).

Firth, D. On improved estimation for importance sampling. *Brazilian Journal of Probability and Statistics* **25**(3) 437–443 (2011). https://doi.org/10.1214/11-BJPS155. MR2832895

Ghosh, J. K. Weak paradoxes and paradigms. In *Statistical Paradigms: Recent Advances and Reconciliations* 3–12. World Scientific (2015). https://doi.org/10.1142/9789814343961_0001. MR3308074

Hahn, P. R., Carvalho, C. M., Puelz, D. and He, J. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis* **13**(1) 163–182 (2018). https://doi.org/10.1214/16-BA1044. MR3737947

Harmeling, S. and Touissant, M. Bayesian estimators for robins-ritov's problem. Technical report, (2007).

Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. https://doi.org/10.1214/16-STS608. MR3648956

Hesterberg, T. Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**(2) 185–194 (1995).

Hesterberg, T. C. *Advances in importance sampling*. PhD thesis, Stanford University (1988). MR2637036

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* **47**(260) 663–685 (1952). MR0053460

Hájek, J. Comment on "an essay on the logical foundations of survey sampling" by basu. (V. P. Godambe and D. A. Sprott, eds.) In *Foundations of Statistical Inference* 236–242. Holt, Rinehart and Winston (1971). MR0423625

Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics* **86**(1) 4–29 (2004).

Isaki, C. T. and Fuller, W. A. Survey design under the regression superpopulation model. *Journal of the American Statistical Association* **77**(377) 89–96 (1982). MR0648029

James, W. and Stein, C. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* 361–379. University of California Press (1961). MR0133191

Khan, S. and Ugander, J. Adaptive normalization for ipw estimation. *Journal of Causal Inference* **11**(1) 20220019 (2023). https://doi.org/10.1515/jci-2022-0019. MR4545024

Kim, C., Tec, M. and Zigler, C. M. Bayesian nonparametric adjustment of confounding (2022). *arXiv preprint arXiv:2203.11798*. https://doi.org/10.1111/biom.13833. MR4680719

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. A theory of statistical models for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(3) 585–604 (2003). https://doi.org/10.1111/1467-9868.00404. MR1998624

Li, L. Are bayesian inferences weak for wasserman's example? *Communications in Statistics—Simulation and Computation®* **39**(3) 655–667 (2010). https://doi.org/10.1080/03610910903576540. MR2784548

Linero, A. R. In nonparametric and high-dimensional models, bayesian ignorability is an informative prior. *Journal of the American Statistical Association* **119**(548) 2785–2798 (2024). https://doi.org/10.1080/01621459.2023.2278202. MR4833915

Little, R. J. Weighting and prediction in sample surveys. *Calcutta Statistical Association Bulletin* **60**(3–4) 147–167 (2008). https://doi.org/10.1177/0008068320080301. MR2553424

Madrid-Padilla, O.-H., Polson, N. G. and Scott, J. A deconvolution path for mixtures. *Electronic Journal of Statistics* **12**(1) 1717–1751 (2018). https://doi.org/10.1214/18-ejs1430. MR3806437

Narain, R. On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **3**(2) 169–175 (1951). MR0045354

Neyman, J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society Series A: Statistics in Society* **97**(4) 558–606 (1934).

Owen, A. B. Comment: Unreasonable effectiveness of monte carlo. *Statistical Science* **34**(1) 29–33 (2019). https://doi.org/10.1214/18-STS676. MR3938960

Philippe, A. Processing simulation output by riemann sums. *Journal of Statistical Computation and Simulation* **59**(4) 295–314 (1997). https://doi.org/10.1080/00949659708811863. MR1623186

Philippe, A. and Robert, C. P. Riemann sums for mcmc estimation and convergence monitoring. *Statistics and Computing* **11**(2) 103–115 (2001). https://doi.org/10.1023/A:1008926514119. MR1837130

Polson, N. G. and Scott, J. G. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics* **9**. 501–538 (2010). https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. MR3204017

Polson, N. G. and Scott, J. G. Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(2) 287–311 (2012). https://doi.org/10.1111/j.1467-9868.2011.01015.x. MR2899864

Polson, N. G. and Scott, J. G. (2014). Vertical-likelihood monte carlo. *arXiv preprint arXiv:1409.3601*.

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (2006). MR2433201

Ramakrishnan, M. An alternative proof of the admissibility of the horvitz-thompson estimator. *The Annals of Statistics* **1**(3) 577–579 (1973). MR0408061

Rao, J. N., Chaudhuri, A., Eltinge, J., Fay, R. E., Ghosh, J., Ghosh, M., Lahiri, P. and Pfeffermann, D. Some current trends in sample survey theory and methods (with discussion). *Sankhyā: The Indian Journal of Statistics, Series B* 1–57 (1999). MR1720726

Ritov, Y., Bickel, P. J., Gamst, A. C. and Kleijn, B. J. K. The bayesian analysis of complex, high-dimensional models: Can it be coda? *Statistical Science* **29**(4) 619–639 (2014). https://doi.org/10.1214/14-STS483. MR3300362

Robins, J. M. and Ritov, Y. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine* **16**(3) 285–319 (1997).

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66**(5) 688–701 (1974).

Särndal, C.-E., Swensson, B. and Wretman, J. *Model assisted survey sampling*. Springer Science & Business Media (2003). https://doi.org/10.1007/978-1-4612-4378-6. MR1140409

Shortreed, S. M. and Ertefaie, A. Outcome-adaptive lasso: variable

selection for causal inference. *Biometrics* **73**(4) 1111–1122 (2017). https://doi.org/10.1111/biom.12679. MR3744525

Sims, C. *Understanding non-bayesians*. Unpublished chapter, Department of Economics, Princeton University (2010).

Sims, C. A. *Thinking about instrumental variables*. manuscript, Department of Economics, Princeton University (2007) manuscript.

Sims, C. A. On an example of larry wasserman, round 2 (2012).

Skilling, J. Nested sampling for general bayesian computation. *Bayesian analysis* **1**(4) 833–859 (2006). https://doi.org/10.1214/06-BA127. MR2282208

Stigler, S. M. The 1988 neyman memorial lecture: a galtonian perspective on shrinkage estimators. *Statistical Science* 147–155 (1990). MR1054859

Trotter, H. F. and Tukey, H. Conditional monte carlo for normal samples. In *Proc. Symp. on Monte Carlo Methods* 64–79. John Wiley and Sons (1956). MR0079825

Wang, C., Dominici, F., Parmigiani, G. and Zigler, C. M. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* **71**(3) 654–665 (2015). https://doi.org/10.1111/biom.12315. MR3402601

Wang, C., Parmigiani, G. and Dominici, F. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68**(3) 661–671 (2012). https://doi.org/10.1111/j.1541-0420.2011.01731.x. MR3055168

Wasserman, L. Bayesian inference. In *All of Statistics* 175–192. Springer (2004). https://doi.org/10.1007/978-0-387-21736-9. MR2055670

Yakowitz, S., Krimmel, J. and Szidarovszky, F. Weighted monte carlo integration. *SIAM Journal on Numerical Analysis* **15**(6) 1289–1300 (1978). https://doi.org/10.1137/0715088. MR0512700

Jyotishka Datta. Jyotishka Datta is Associate Professor, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. E-mail address: jyotishka@vt.edu

Nicholas Polson. Nicholas G. Polson is Professor, Booth School of Business, University of Chicago, USA. E-mail address: ngp@chicagobooth.edu