Self-Normalized IPW

David S. Rosenberg

NYU: CDS

February 3, 2021

Contents

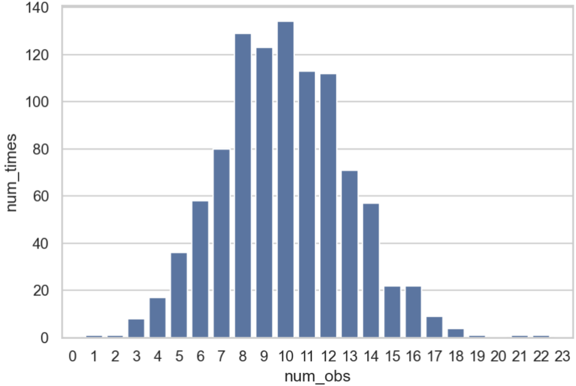
Self-normalized IPW for MCAR

Recap and what's next?

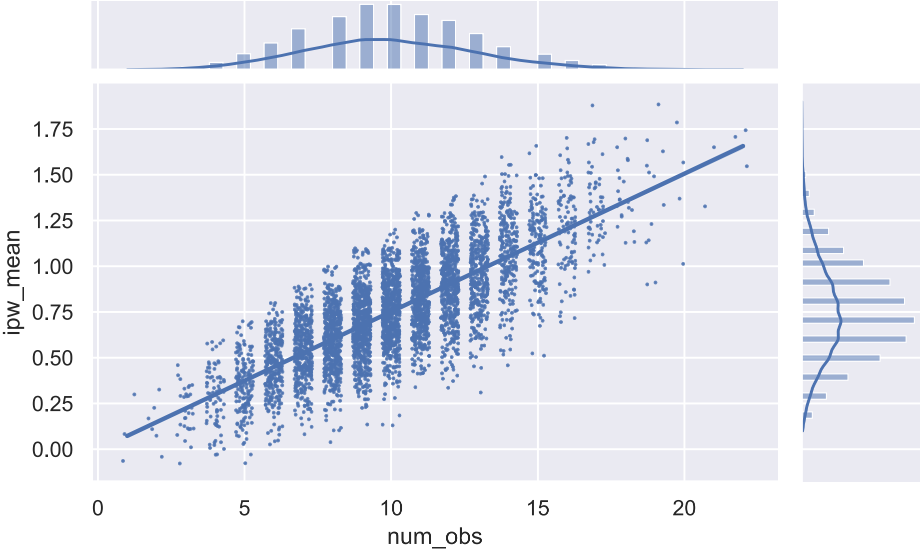
- The complete-case mean $\hat{\mu}_{\text{cc}}$ can be very biased for MAR setting
- The IPW mean $\hat{\mu}_{\text{ipw}}$ is unbiased and consistent for the MAR setting.
- In MCAR setting, found that $\hat{\mu}_{\text{cc}}$ performed much better than $\hat{\mu}_{\text{ipw}}$.
- We'll now do a deeper dive into $\hat{\mu}_{\text{ipw}}$ in the MCAR setting.
- Goal: try to find a tweaked version of $\hat{\mu}_{\text{ipw}}$ that performs better in MCAR setting.
- (Then hope that tweaked version also performs better in MAR setting.)

Probability review: how many responses will we get?

- $R, R_1, \ldots, R_n \in \{0,1\}$ are i.i.d. with $\mathbb{P}(R=1) = 0.1$.
- Number of observations: $N = \sum_{i=1}^{n} R_i$.
- Expected number of responses is $\mathbb{E}N = 0.1n$ and $N \sim \text{Binom}(n, p = 0.1)$.
- Histogram of N from 1000 simulations of our setup with $n = 100$:

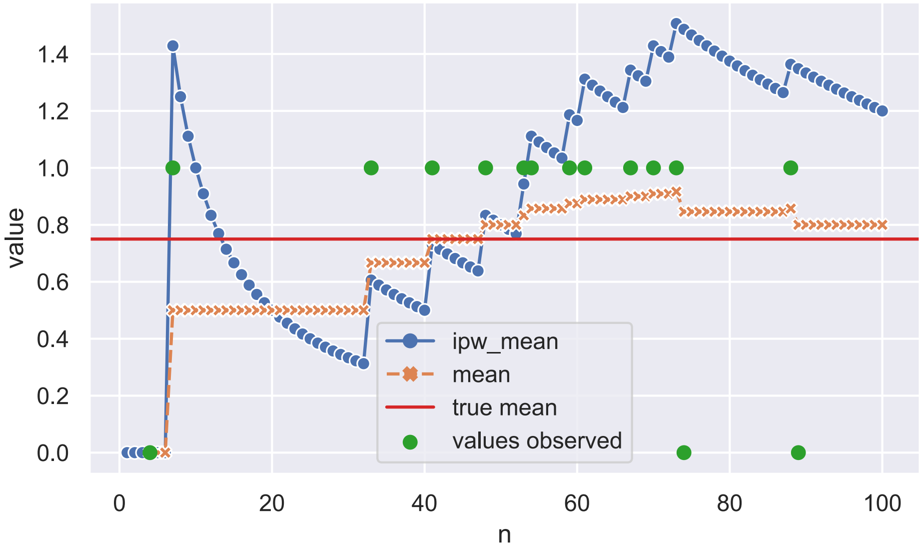


IPW mean vs number of observations



- Before we get into the math, let's first visualize whether there's a relationship between the number of observations we get $N = \sum_{i=1}^{n} R_i$ and the IPW mean estimate $\hat{\mu}_{\mathrm{ipw}}$.
- Visually, the linear relationship is quite striking.
- Note that we've added "jitter" to the x-value on the plot to see more clearly what's going on. The actual x value is the number of observations, so of course it's an integer.

IPW mean for "too many" observations



- In this example, by random chance we got 15 observations, a large deviation from the expected number of 10.
- The IPW estimate is significantly higher than the true mean.
- The complete case mean estimator is only slightly higher than the true mean.
- The conclusion is that almost all of the error for IPW is due to having "too many" observations, rather than an unlucky sample of observations of Y (since the complete case estimator is doing quite reasonably).

IPW: What if we get many more/fewer observations than expected?

- For our scenario, the IPW estimator is

$$\hat{\mu}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{0.1} = \sum_{i:R_i=1}\frac{Y_i}{0.1n}$$

- Now $\mathbb{E}\left[Y_i/0.1n\right] = \frac{\mu}{.1n}$. So

$$\mathbb{E}\left[\hat{\mu}_{\text{ipw}} \mid \sum_i R_i = N\right] = \frac{\mu N}{0.1n}$$

- Our estimate has a strong dependence on N.
- All that variance in N is increasing the variance of our estimator.

With a bit more formality:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\mu}_{\mathrm{ipw}} \mid \sum R_i = N\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi} \mid \sum R_i = N\right] \\
&= \frac{1}{n\pi}\sum_{i=1}^{n}\mathbb{E}\left[R_i Y_i \mid \sum R_i = N\right] \\
&= \frac{1}{n\pi}\sum_{i=1}^{n}\mathbb{E}\left[Y_i \mid \sum R_i = N\right]\mathbb{E}\left[R_i \mid \sum R_i = N\right] \text{ MCAR} \\
&= \frac{1}{n\pi}\sum_{i=1}^{n}\mathbb{E}\left[Y_i\right]\frac{N}{n} \text{ MCAR} \\
&= \frac{\mu N}{\pi},
\end{aligned}
$$

where $\pi = \mathbb{P}(R_i = 1) = 0.1$.

IPW: Can we improve this estimator?

- Since

$$\mathbb{E}\left[\hat{\mu}_{\text{ipw}} \mid \sum_i R_i = N\right] = \frac{\mu N}{0.1n},$$

- $\hat{\mu}_{\text{ipw}}$ is off in expectation by a factor of $N/.1n$.
- What if we "fix" $\hat{\mu}_{\text{ipw}}$ by dividing by that factor?

$$\begin{aligned}
\frac{0.1n}{N}\hat{\mu}_{\text{ipw}} &= \frac{0.1n}{N}\sum_{i=1}^{n}\frac{R_i Y_i}{0.1n} \\
&= \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}
\end{aligned}$$

- And that's exactly $\hat{\mu}_{\text{cc}}$, the average of the observed Y's, that we started with!
- Have we gone in a useless circle?
- Not at all! Let's try to apply this "correction" to the more general MAR case...

Self-normalized IPW for MAR

Recap and what's next?

- The complete-case mean $\hat{\mu}_{\text{cc}}$ can be very biased for MAR setting
- The IPW mean $\hat{\mu}_{\text{ipw}}$ is unbiased and consistent for the MAR setting,
  - but seems to have high variance and is
  - too large (positively biased) when we have "too many" observations and
  - too small (negatively biased) when we have "too few" observations.
- For MCAR setting, we did a deep dive and proposed an adjustment to $\hat{\mu}_{\text{ipw}}$,
  - and it ended up being equivalent to $\hat{\mu}_{\text{cc}}$.
- Can we make an analogous modification to $\hat{\mu}_{\text{ipw}}$ that works for the MAR setting?

Weights

- We can write
$$\hat{\mu}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i)} = \frac{1}{n}\sum_{i=1}^{n} W_i R_i Y_i,$$
where
$$W_i := \frac{1}{\pi(X_i)} = \frac{1}{p(R_i = 1 \mid X_i)}.$$
- We'll refer to W_i as the **weight** for observation Y_i.
- It's like each observed response Y_i (with $R_i = 1$) represents W_i responses in the full data.
- Upweighting by W_i makes up for the zeros when $R_i = 0$.

IPW in MAR: very large or small number of observations?

- If each observed response Y_i represents W_i responses in the full data,
  - then our observed data represents $\sum_{i=1}^n W_i R_i$ people.
- The IPW estimate normalizes by n: $\hat{\mu}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^n W_i R_i Y_i$.
- We'll show in homework that

$$\mathbb{E}\left[\sum_{i=1}^n W_i R_i\right] = n.$$

- So not totally unreasonable to divide by n.
- But what if $\sum_{i=1}^n W_i R_i$ is much smaller or larger than n?
- Then it seems like we're normalizing by the wrong thing...

The self-normalized IPW estimator

- If we normalize by $\sum_{i=1}^{n} W_i R_i$ instead of n, we get

Definition (Self-normalized IPW mean)

For a dataset $(W_1, R_1, Y_1), \ldots, (W_n, R_n, Y_n)$ as described above,

$$\hat{\mu}_{\text{sn\_ipw}} = \frac{\sum_{i=1}^{n} W_i R_i Y_i}{\sum_{i=1}^{n} W_i R_i}$$

Notes

- When we take $W_i \equiv k$, for any constant $k \neq 0$, we get back $\hat{\mu}_{\text{cc}}$.
- In the MCAR case with $\pi(x) \equiv p$, $\hat{\mu}_{\text{sn\_ipw}} = \hat{\mu}_{\text{cc}}$ and seems preferable to $\hat{\mu}_{\text{ipw}}$.
- Let's investigate how $\hat{\mu}_{\text{sn\_ipw}}$ compares to $\hat{\mu}_{\text{ipw}}$ a more complicated MAR scenario.

IPW vs. self-normalized IPW on MAR: Simulation

MAR: "SeaVan1" distribution

- X is drawn uniformly from $\{0,1,2\}$.
- $Y \mid X = x \sim \mathcal{N}(x,1)$
- $R \mid X = x \sim \text{Bern}(\text{expit}(4-4x))$, where $\text{expit}(x) = 1/(1+e^{-x})$:

x	$\pi(x) = \mathbb{P}(R=1 \mid X=x)$
0	.982
1	.500
2	.018

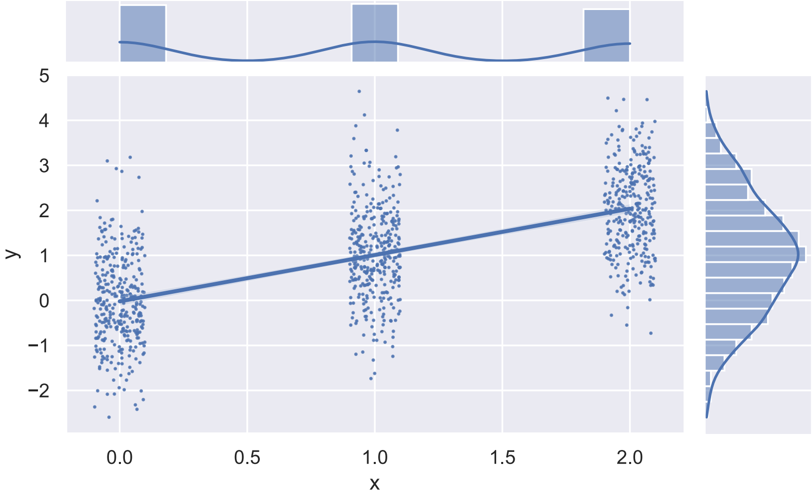
- $(X,R,Y),(X_1,R_1,Y_1),\ldots,(X_n,R_n,Y_n)$ are i.i.d. with distribution described above.
- We'll refer to this distribution as "**SeaVan1**", based on the names of the authors who created it[1]

[1] Based on Example 1 in [SV18]

- This distribution corresponds to a massive response bias: an individual with $X = 0$ is 55 times more likely to respond than an individual with $X = 2$.

MAR: "SeaVan1" distribution illustrated

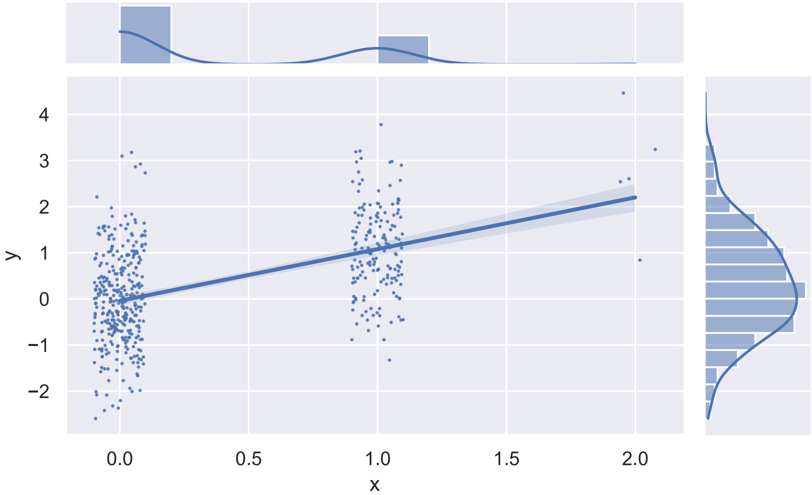
A sample of size 1000 from the full data distribution is shown below:



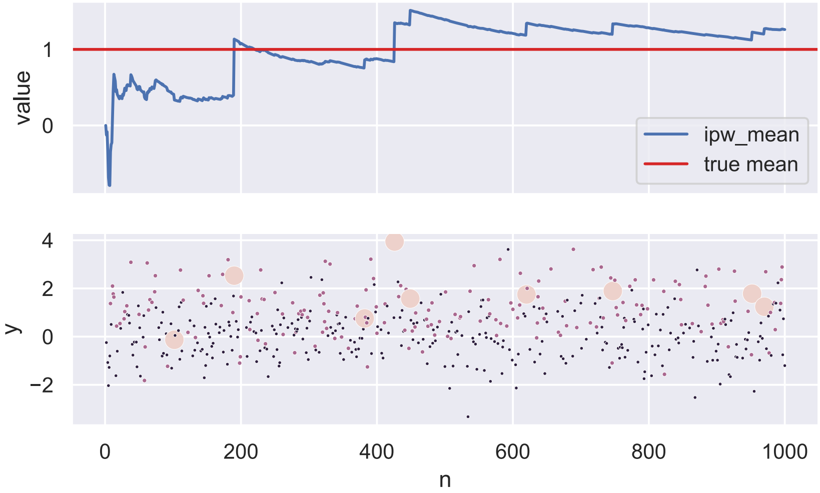
We've added jitter to the x values so that it's easier to see the distribution.

MAR: "SeaVan1" distribution illustrated

(X_i, Y_i) for which $R_i = 1$, i.e. the complete cases.

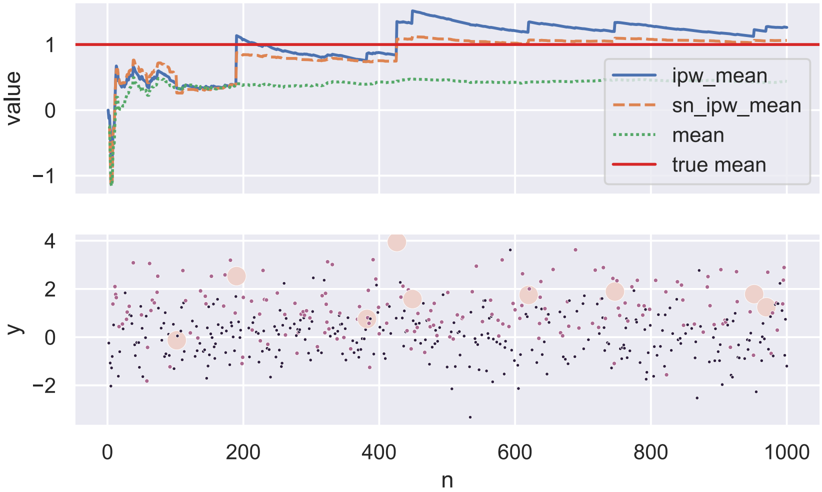


IPW estimator on SeaVan1



- On the bottom graph, we have a scatter plot of the observed Y values as they showed up in a sequence of 1000 draws from the data generating distribution.
- The color corresponds to the X value, which also determines the weight of that observation. The size of the plot point is correlated with the weight. The biggest circles, corresponding to $X = 2$, are the most rare, and thus have the most weight and are drawn the largest.
- Note that a lot of large jumps in the IPW estimator occur when we observe one of the rare Y's that correspond to $X = 2$, as these points have a lot of weight.

Self-normalized IPW estimator on SeaVan1



- Here we've add the self-normalized IPW estimate, denoted sn_ipw_mean, as well as the mean of the observed values
- We can see that complete-case mean (the green line) stabilizes rather quickly to the wrong value. (Exercise: what value does it converge to?)
- We see that the self-normalized estimator doesn't have such a pronounced jump when each of the rare points is observed.
- We also don't see as pronounced a decay between these observations.
- Visually, it seems like the self-normalized estimator has lower variance, but we'll investigate this more thoroughly with simulations in later slides.

IPW vs self-normalized IPW: 5000x

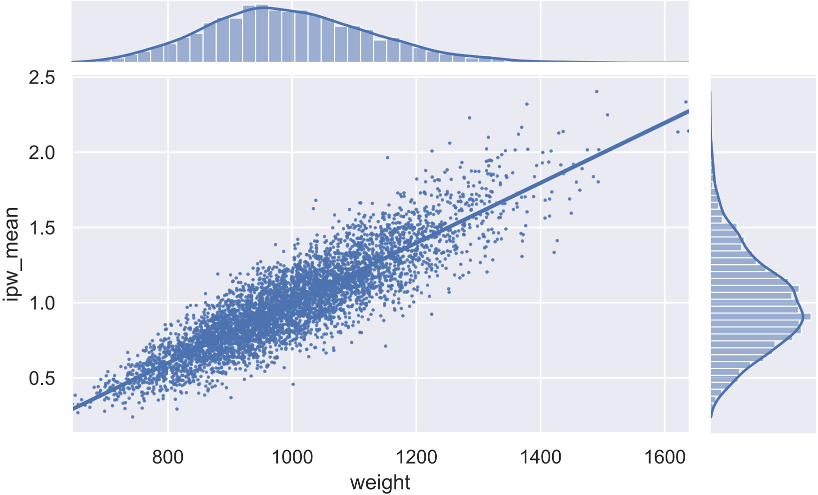
- We repeat the experiment above 5000 times (1000 samples each) and get the following.
- Recall that the true mean is $\mu = 1.0$.

estimator	mean	SD	SE	bias	RMSE
mean ($\hat{\mu}_{cc}$)	0.357244	0.050305	0.000711	-0.643534	0.645497
ipw_mean ($\hat{\mu}_{ipw}$)	0.995142	0.308634	0.004365	-0.005635	0.308686
sn_ipw_mean ($\hat{\mu}_{sn\_ipw}$)	0.978119	0.197319	0.002791	-0.022659	0.198615

- As expected, the mean [of the observed Y's] has a large bias and that drives almost all the RMSE (root mean squared error). The SD is relatively small.
- From theory, we know the IPW mean is unbiased, and indeed our bias estimate is just a bit more than 1 SE from 0, which concurs with expectations. Almost all of the RMSE comes from the SD.
- The self-normalized IPW estimator does have a bias, and indeed our bias estimate is many SEs from 0 – so clearly there's a significant bias.
- However, the SD of the self-normalized IPW estimate is almost 10x the bias, and so the RMSE is dominated by the variance of the estimator, rather than the bias.
- Overall, it seems that, at least for this data generating distribution, the self-normalized estimator IPW makes a much better tradeoff between bias and variance than the ipw_mean and mean estimators.

IPW vs total weight of observations

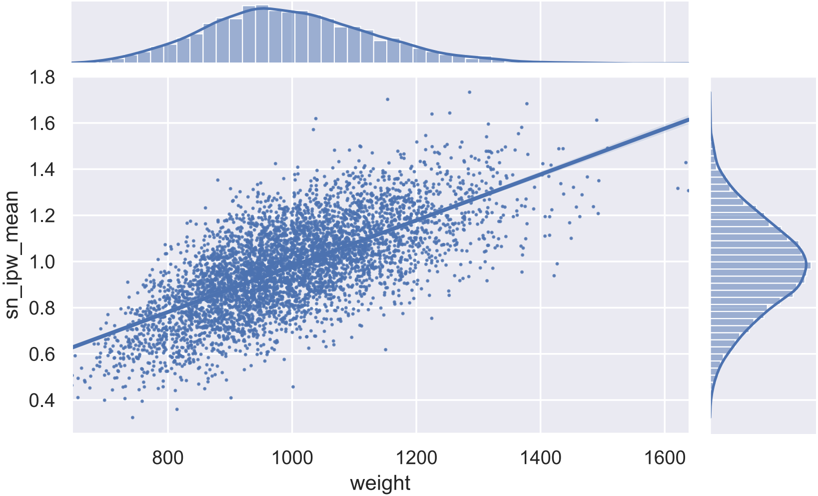
- The points below have correlation 0.885.



- Above we suggested that normalizing by $\sum_{i=1}^{n} W_i R_i$ instead of n might give better estimates when $\sum_{i=1}^{n} W_i R_i$ is quite different from n, its expectation.
- Here we plot the IPW estimate vs the total weight across 5000 trials (sample size still 1000).
- Here we see a very strong correlation between the total weight of the observed instances and the ipw_mean estimate.
- In words, when we have small weight, we typically underestimate the mean, and when we have large weight, we typically overestimate the mean.

SN-IPW vs total weight of observations

- 5000 trials; sample size 1000. Correlation is 0.690.



References

References I

[SV18] Shaun R. Seaman and Stijn Vansteelandt, *Introduction to double robust methods for incomplete data*, Statistical Science **33** (2018), no. 2, 184–197.