Propensity score stratification: New insights to an old problem.

Roland A. Matsouaka

Department of Biostatistics and Bioinformatics
&
Program for Comparative Effectiveness Methodology, Duke Clinical Research Institute

Duke University, Durham, North Carolina.

Presented at: International Conference on Health Policy Statistics (ICHPS)

San Diego, CA

Tuesday January 7, 2020 at 11:00 AM

Outline

1. PS stratification: an overview
2. It's all about the weights
3. What are we overlooking?
4. What weights to use?
5. Illustrative examples

Estimating the average treatment effect (ATE)

Consider a trt $Z = \{0, 1\}$, a covariate-vector $\boldsymbol{X}$, and an outcome Y.

Aim of most studies: estimate the effect of Z on Y.

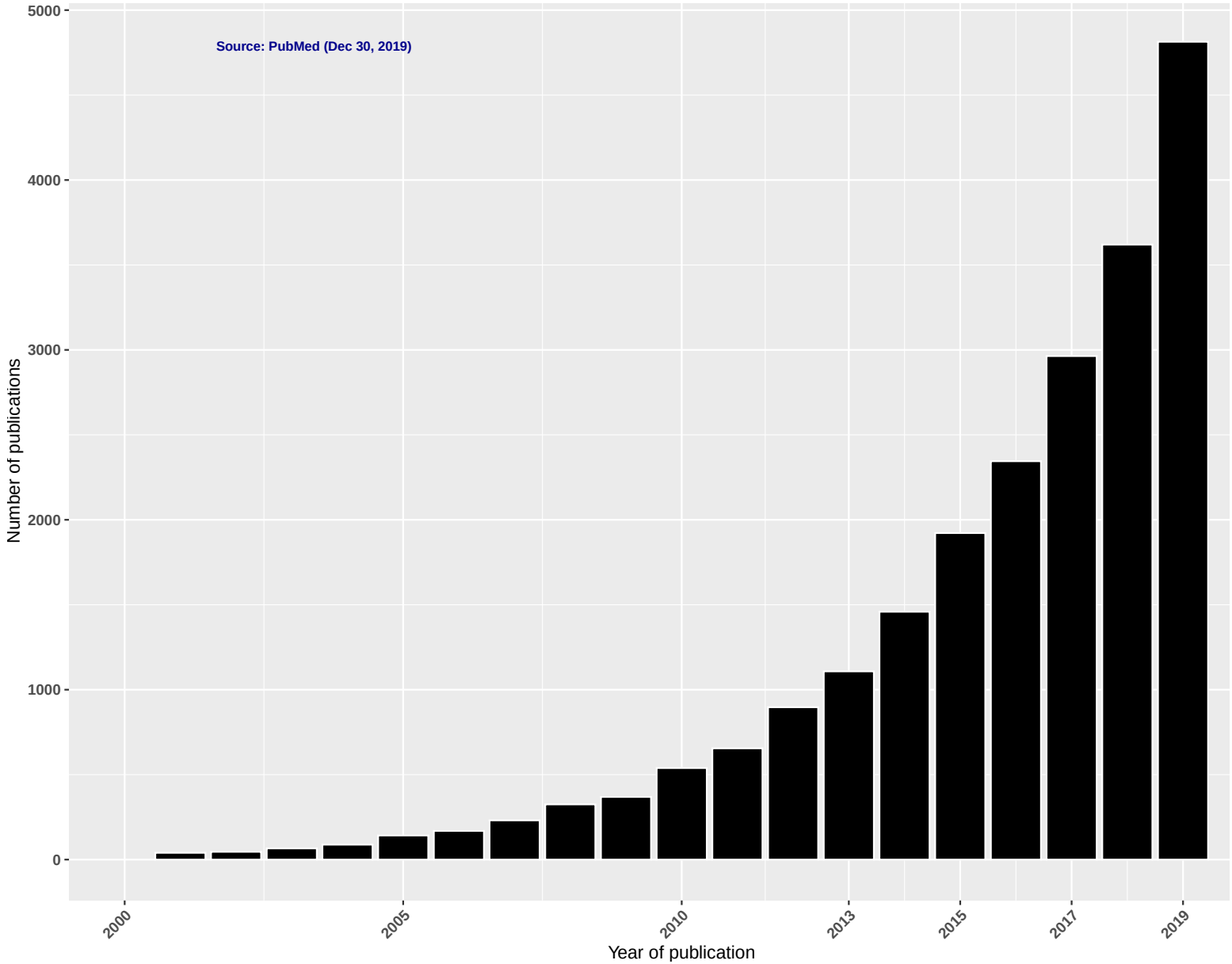
- Rubin-Neyman's potential outcome: each individual has $(Y(0), Y(1))$
- We observe $Y = ZY(1) + (1 - Z)Y(0)$
- Objective: estimate $\mu = E[Y(1) - Y(0)]$.

Confounding in non-randomized studies

Aim : estimate the effect of Z on Y, i.e., $\mu = E\left[Y(1) - Y(0)\right]$

- RCT ensures covariate balance; but may still control for $\boldsymbol{X}$
- For non-RCT: we need to adjust for confounding
- PS methods are increasingly used to evaluate such trt effects.

Propensity score analysis (PSA): State of affairs



Propensity score analysis (PSA)

$$\text{Propensity score} = e(\boldsymbol{X}) = P(Z = 1|\boldsymbol{X}):$$

reflects the propensity to receive $Z = \{0, 1\}$, based on observed covariates

- $(Y(1), Y(0)) \perp\!\!\!\perp Z|e(\boldsymbol{X})$ whenever $(Y(1), Y(0)) \perp\!\!\!\perp Z|\boldsymbol{X}$
- It is a balancing score: i.e. $E[E(Y|Z = z, e(\boldsymbol{X}))] = E[Y(z)]$.

Propensity score analysis (PSA)

All **PS Methods** take advantage of the balancing score property.

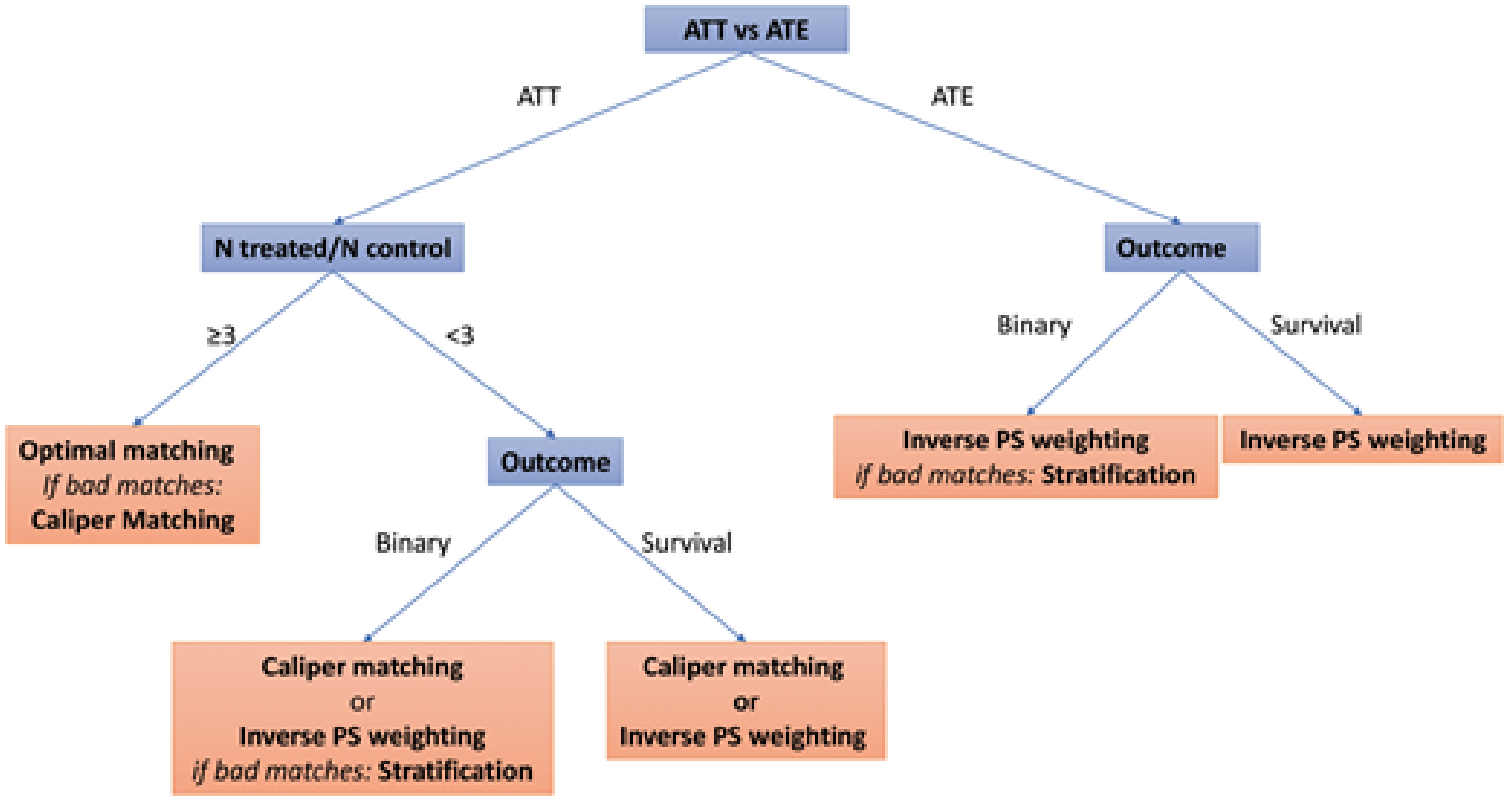
PSA is conducted in two steps

- Step I: estimate PS's (logistic reg., GAM, GBM, BART, etc.)
- Step II: estimate trt effects of interest using a chosen PS method.

PS Methods: PS regression, matching, weighting, stratification.
(we can also combine with regression $\Leftrightarrow$ double robustness)

Algorithm to select the most appropriate PS method*

PS methods: PS regression, matching, weighting, stratification.



* Statistical primer: propensity score matching and its alternatives [Benedetto et al., 2018]

PS stratification: Modus operandi

PS stratification idea: Leverage PS balancing score property

- Partition the sample into PS strata S_k, $k = 1, \ldots, K$
- Calculate $\widehat{\mu}_k = \sum_{i=1}^{N} I(e_{ik} \in S_k) \left[\frac{Z_i Y_i}{N_{1k}} - \frac{(1 - Z_i) Y_i}{N_{0k}} \right]$
- Estimate μ as a weighted average $\widehat{\mu} = \sum_{k=1}^{K} w_k \widehat{\mu}_k$

where N_{zk} = number of patients in trt $Z = z$, for $z = 0, 1$.

PS stratification: Modus operandi

True weights are known; need to be specified using the data

Commonly-used weights

- Sample-fraction weights (SFW): $\widehat{w}_k^{(sf)} = \dfrac{N_k}{N}$,

with $N_k = N_{0k} + N_{1k}$ = number of patients in stratum S_k

PS stratification: Granularities

- **Usual assumptions are made**: SUTVA, SITA, Positivity, Balance[1]
- PS estimation often ignored in inference; although:
  1. Number and choice of strata boundaries influenced by PS model
  2. Estimator depends on the PS estimation
- Rationale for weights choice?

.

[1] SUTVA: Stable unit trt value assumption; SITA: Strongly ignorable trt assignment [Rosenbaum and Rubin, 1983]

PS stratification: Balance and weights

Justification for the SF weights $\widehat{w}_k^{(sf)} = \dfrac{N_k}{N}$

The choice for the weights w_k is made assuming that

> *"... there is* **little variation within a stratum** *or block, and one can analyze the data as if the propensity score is constant, and thus as if the data within a block were generated by a completely* **randomized experiment**." [Imbens and Wooldridge, 2009]

PS stratification: Balance and weights

An almost **block randomization** is ideal, but untenable. [Morgan and Winship, 2014].

In reality, a more coarse stratification is used to avoid sparse strata.

Moreover, it's been suggested the use

- outcome regression models to reduced residual bias
- alternative weights, including inverse-variance weights (IVW)

PS weights: either . . . or

sample-fraction weights (SFW):

$$\widehat{w}_k^{(sf)} = \frac{N_k}{N}$$

or

inverse-variance weights (IVW):

$$\widehat{w}_k^{(iv)} = \left(\sum_{k=1}^{K} 1/\widehat{\sigma}_k^2 \right)^{-1} \left(1/\widehat{\sigma}_k^2\right)$$

PS weights: one vs. the other?

Rudolph et al.[1] compared

$$\widehat{w}_k^{(sf)} = \frac{N_k}{N} \qquad \text{vs.} \qquad \widehat{w}_k^{(iv)} = \left(\sum_{k=1}^{K} 1/\widehat{\sigma}_k^2 \right)^{-1} \left(1/\widehat{\sigma}_k^2\right)$$

and showed that,

- under assumptions of positivity and **constant trt effect**,
  - both methods perform well;
  - IVW performs slightly better.
- However, under **trt heterogeneity**, SFW outperforms IVW

[1] Optimally combining propensity score subclasses [Rudolph et al., 2016]

The inverse-variance weights

Why the inverse-variance weights?

Optimal: Minimize variance, AMSE; maximize power, signal-to-noise ratio.

Rationale for IVW, under constant treatment effect:

- convey the info underlying trt effect in each stratum;
- strata with smaller variance must weigh more (precision)
- IVW better borrow strength across strata to estimate trt effect

Inverse-variance weights: special cases (Part I)

Consider $w_k^{(iv)} = \left(\sum_{k=1}^{K} 1/\sigma_k^2\right)^{-1} (1/\sigma_k^2)$ with $\sigma_k^2 = \dfrac{N_{0k}\sigma_{1k}^2 + N_{1k}\sigma_{0k}^2}{N_{0k}N_{1k}}$

- If $N_{1k}\sigma_{0k}^2 + N_{0k}\sigma_{1k}^2 = aN_k$, ($a \in \mathbb{R}^+$), we have $\sigma_k^2 = \dfrac{aN_k}{N_{0k}N_{1k}}$ and

$$w_k^{(iv)} = \left(\sum_{k=1}^{K} \frac{N_{0k}N_{1k}}{N_k}\right)^{-1} \frac{N_{0k}N_{1k}}{N_k}$$

i.e., $w_k^{(iv)}$ = the Mantel-Haenszel weights (MHW)

Inverse-variance weights: special cases (Part II)

Let $p_k = \dfrac{N_{1k}}{N_k}$ and consider $\widehat{w}_k^{(mh)} = \left(\sum_{k=1}^{K} \dfrac{N_{0k} N_{1k}}{N_k} \right)^{-1} \dfrac{N_{0k} N_{1k}}{N_k}$

- We can write $\widehat{w}_k^{(mh)} = \left(\sum_{k=1}^{K} N_k p_{1k}(1 - p_{1k}) \right)^{-1} N_k p_{1k}(1 - p_{1k})$
- $\widehat{w}_k^{(mh)}$ is equal to $\widehat{w}_k^{(mh)} = \dfrac{N_k}{N}$, if $p_{1k}(1 - p_{1k}) = b$, $b \in \mathbb{R}^+$

i.e., Mantel-Haenszel weights simplify to the sample-fraction weights

SFW and MHW are special cases for IVW

1. **MHW**: whenever $N_{1k}\sigma^2_{0k} + N_{0k}\sigma^2_{1k} = aN_k$
2. **SFW**: if $N_{1k}\sigma^2_{0k} + N_{0k}\sigma^2_{1k} = aN_k$ and $p_{1k}(1-p_{1k}) = b$

Questions

- Are the **SFW assumptions** plausible?
- Why were **Rudolf et al.'s results** conflicting? (constant vs. heterogeneous trt)
- Why IVW not adopted throughout, like in Meta-analysis methods?

PS inverse-variance weights: the issues

Big picture

Homogeneity, independence, consistency and unbiasedness

Issues may occur when there is

- (strong) heterogeneity of trt across strata
- small strata or sparse strata
- correlation since $E\left(\sum_{k=1}^{K} \widehat{w}_k \widehat{\mu}_k\right) = \sum_{k=1}^{K} [E(\widehat{w}_k)E(\widehat{\mu}_k) + Cov(\widehat{w}_k, \widehat{\mu}_k)]$
- $\widehat{w}_k$ is a consistent, but not an unbiased estimator of w_k

PS inverse-variance weights: the truth is . . .

1. $\widehat{\mu}_k \perp\!\!\!\perp \widehat{\sigma}_k^2$ if and only if $\widehat{\mu}_k \sim N(\mu_k, \sigma_k^2)$.
2. In general, $E\left(\frac{1}{\widehat{\sigma}_k^2}\right) \geq \frac{1}{\sigma_k^2}$ (by Jensen's inequality)
3. with $\widehat{w}_k$, $Var(ATE)$ is understimated, even if $\widehat{\mu}_k \sim N(\mu_k, \sigma_k^2)$.
4. If $\widehat{\mu}_k$ is not normal, we don't always know what we're getting

($\widehat{\mu}_k$ and $\widehat{\sigma}_k^2$ are not independent; the weights and the variance Var(ATE) are underestimated)

☞ What should we do?

Calibrate the weights ...

Calibrate the weights and re-evaluate the variance Var(ATE)

When $\widehat{\mu}_k \sim N(\mu_k, \sigma_k^2)$, we have $E\left[\frac{c_k}{\widehat{\sigma}_k^2}\right] = \frac{1}{\sigma_k^2}$ where $c_k = \frac{N_k - 3}{(N_k - 1)}$

- Hence, we calibrate the weights, ATE estimate and variance* as:

1. $\widehat{\mu}_{*ate} = \sum_{k=1}^{K} \widehat{w}_{*k}\widehat{\mu}_k \quad \text{with} \quad \widehat{w}_{*k} = \left[\sum_{k=1}^{K} \frac{c_k}{\widehat{\sigma}_k^2}\right]^{-1} \frac{c_k}{\widehat{\sigma}_k^2}$

2. $Var(\widehat{\mu}_{*ate}) = \left[\sum_{k=1}^{K} \frac{c_k}{\widehat{\sigma}_k^2}\right]^{-1} \left[1 + 4\sum_{k=1}^{K} \frac{w_{*k}(1 - w_{*k})}{N_k - 1}\right]$

*Variance of a weighted mean [Meier, 1953]

. . . and call a wild bootstrap to the rescue

When $\widehat{\mu}_k$ is not normal or we just want to generalize,

use a wild bootstrap[1] to estimate the weights

☞ Obtain B bootstrap replicates by perturbing the original sample

(a.k.a perturbation-resampling method)

Wild bootstrap algorithm

☞ Estimate $\widehat{\mu}^*_{bk}$ and $\widehat{\sigma}^{*2}_{bk}$ via perturbation of the original sample

- For each $b = 1, \ldots, B$, generate $v_i \sim exp(1)$
  1. perturb Y in the original sample
  2. estimate the PS using a v-weighted logistic model
  3. split the bootstrap sample into strata
  4. calculate v-weighted $\widehat{\mu}^*_{bk}$, $\widehat{\sigma}^*_{bk}$, and $w^*_{bk} \propto \frac{c_k}{\widehat{\sigma}^{*2}_{bk}}$
- use as weights w^*_k the mean of w^*_{bk}, $b = 1, \ldots, B$

[1] A simple resampling method by perturbing the minimand [Jin et al., 2001]

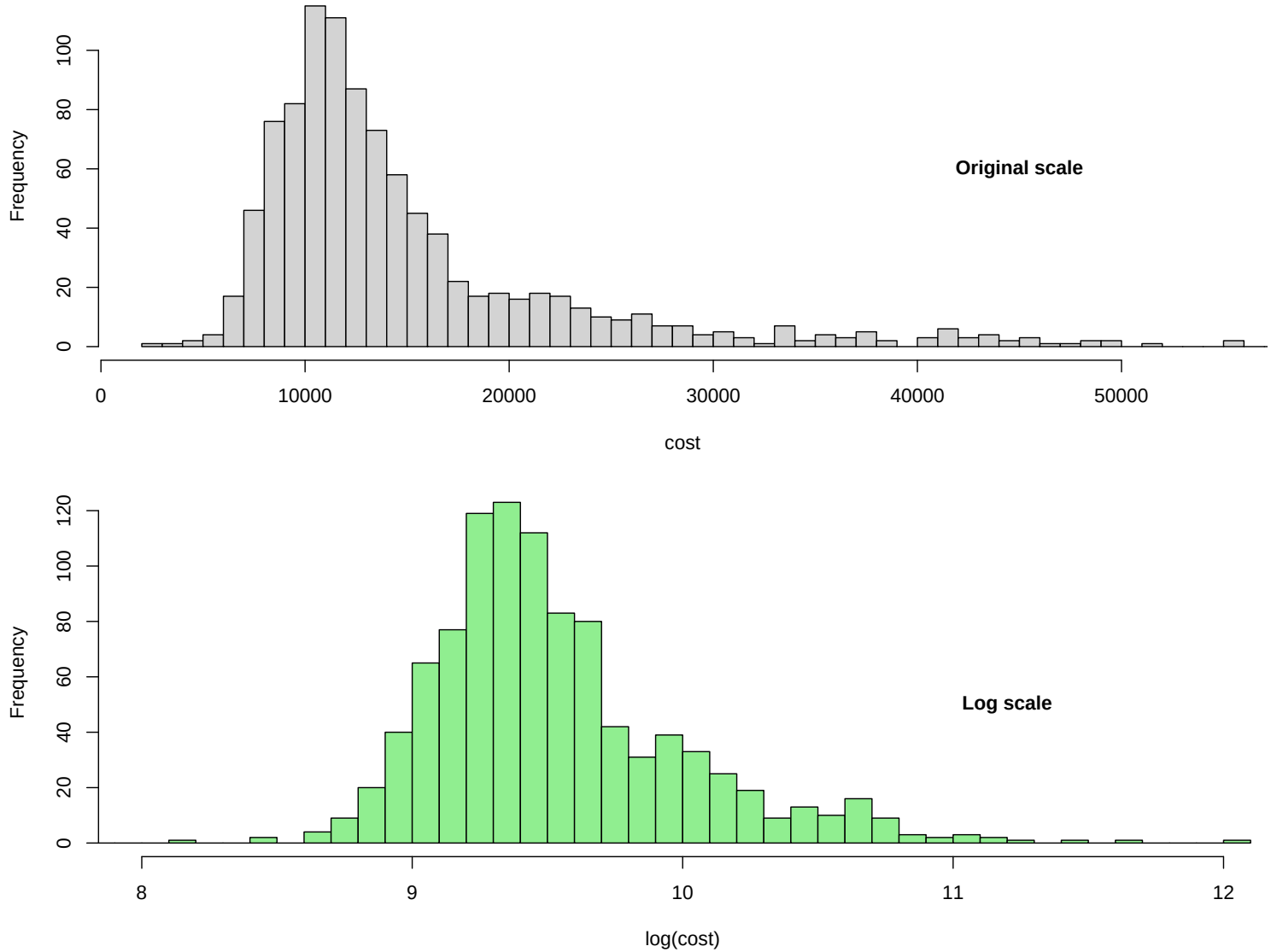
Example 1: The Lindner data set (1997)

Dataset from Lindner Center, Christ Hospital, Cincinnati, OH[1]

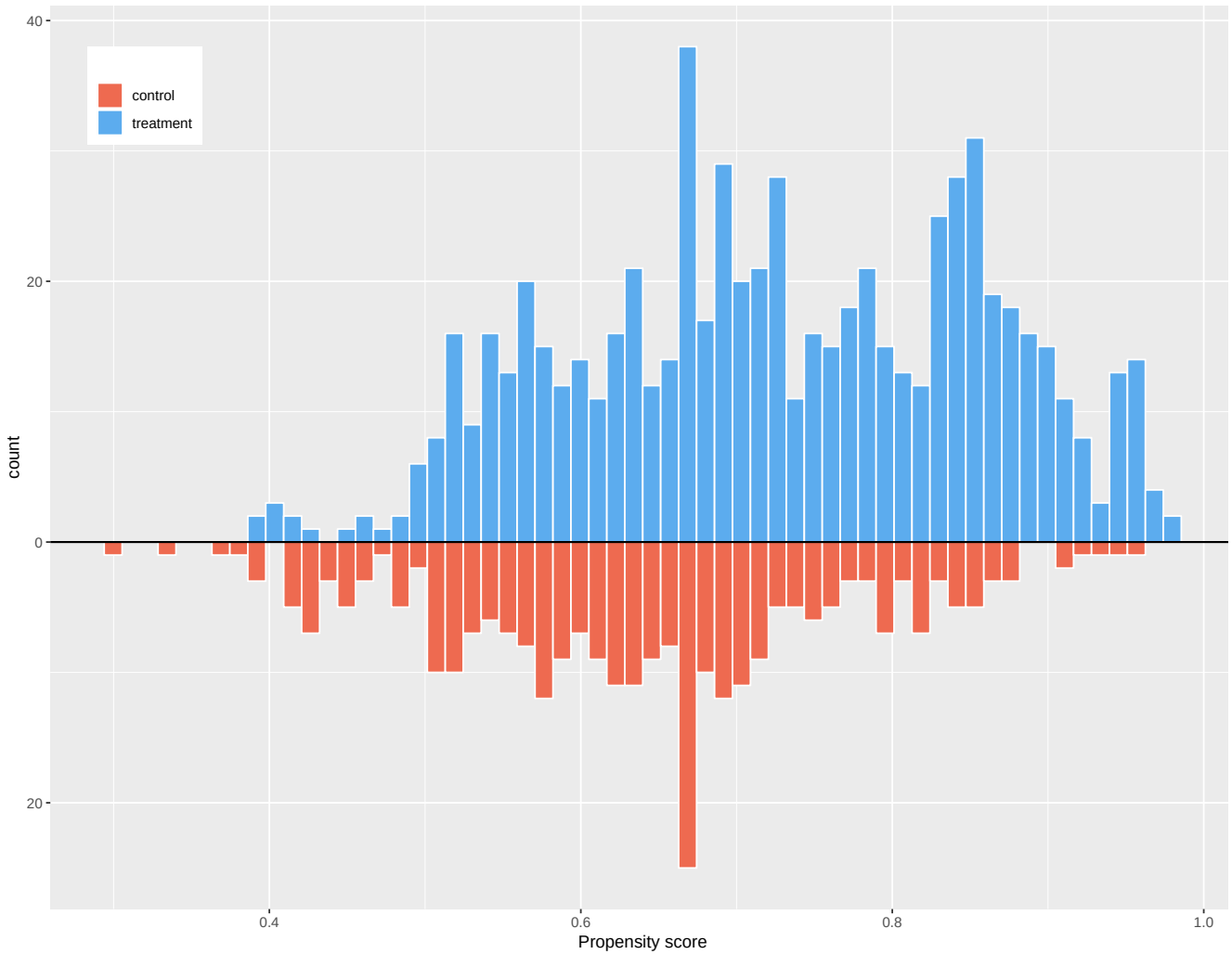
- 996 patients who received Percutaneous Coronary Intervention (PCI)
- Outcomes: lifepres (dead or alive) and cardbill (6-month cost in $)
- Trt: PCI vs. PCI+abciximab (298 patients in PCI group)
- 26 patients died (15 in the PCI group)
- 7 covariates including gender, height, stent, diabetic, acute MI.

[1] Come with R packages such as **USPS, PSAgraphics, twang**

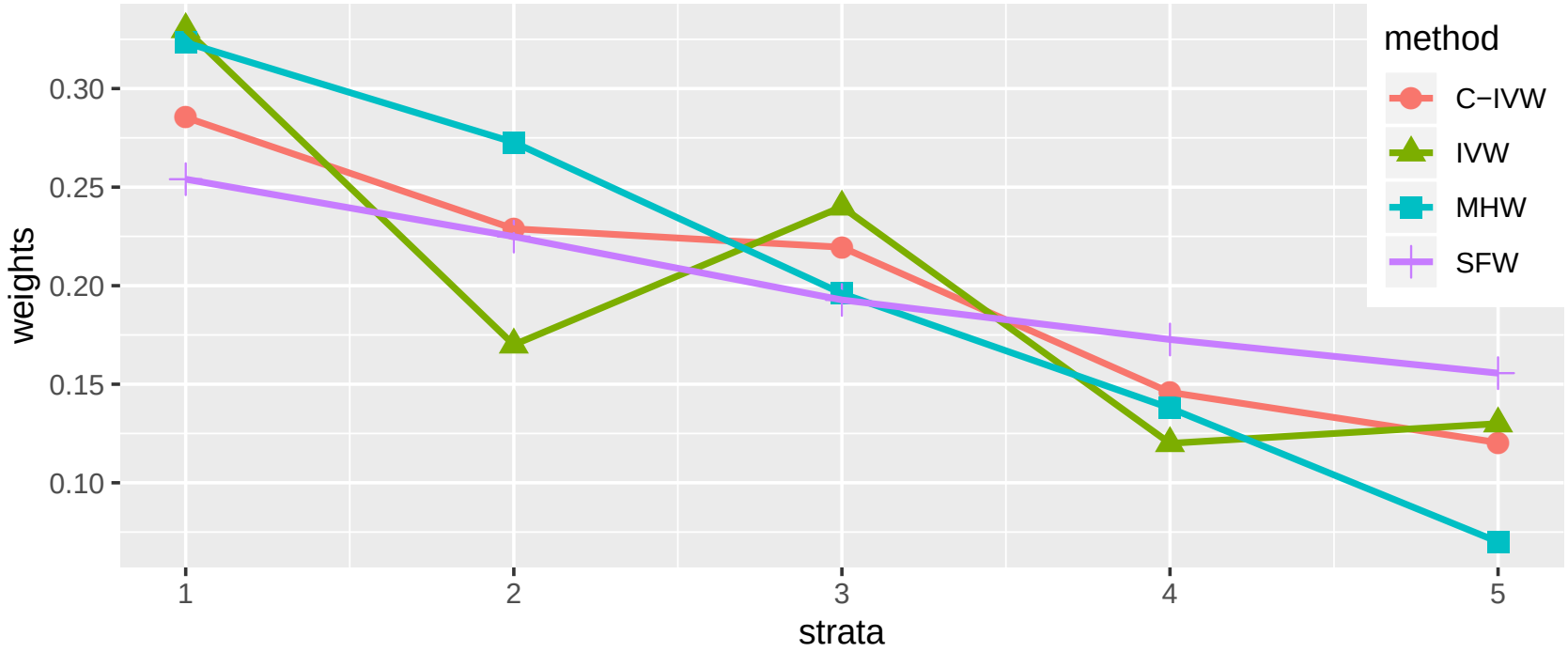
The Lindner data set: Outcome distribution (Cardbill)



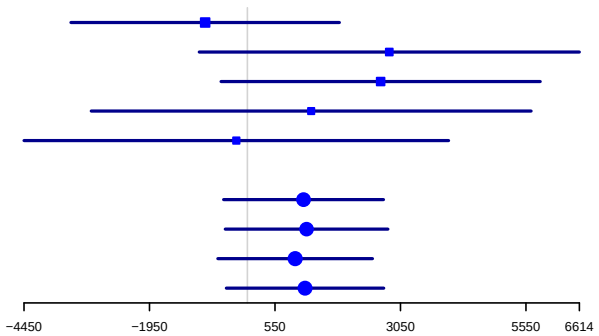
The Lindner data set: Propensity scores



The Lindner data set: cardbill with 5 strata

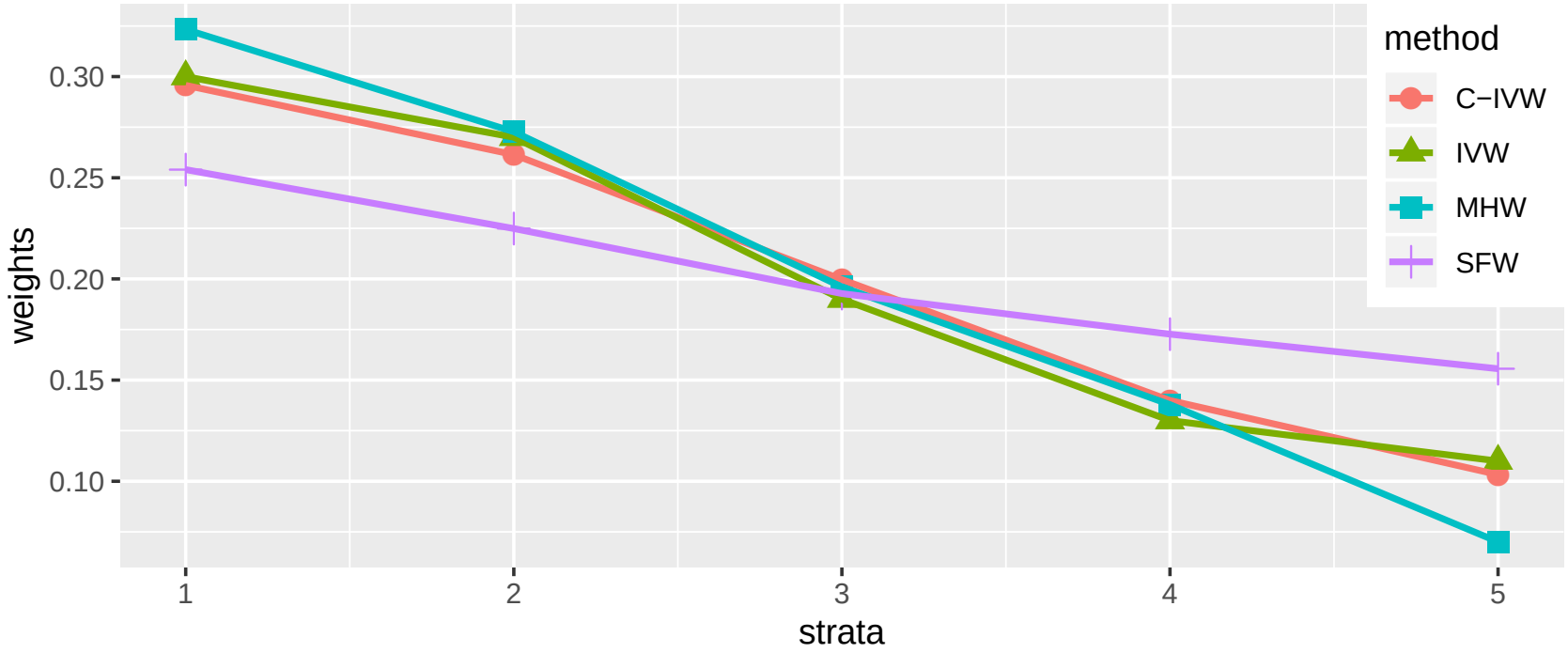


Stratum	ATE
1	−841.3
2	2827.12
3	2654.6
4	1270.63
5	−220.6
Summary	
SFW	1118.94
MHW	1179.17
IVW	951.73
C−IVW	1148.46

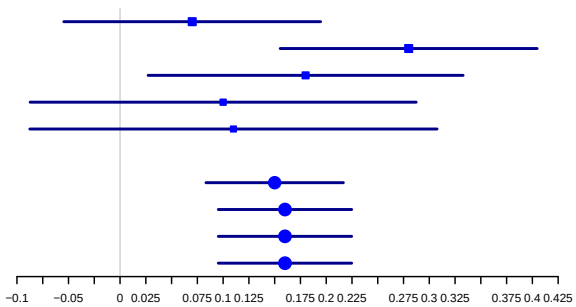


Method	ATE	Std. Error	p-value
SFW	1,118.94	812.45	0.17
MHW	1,179.17	826.54	0.15
IVW	951.73	785.74	0.23
C-IVW	1,148.46	799.92	0.15

The Lindner data set: cardbill with 5 strata (log scale)

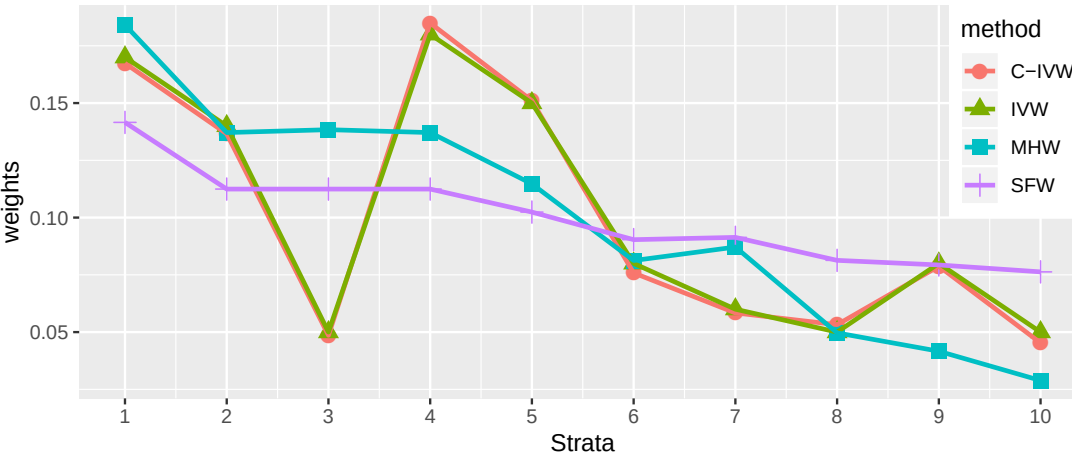


Stratum	ATE
1	0.07
2	0.28
3	0.18
4	0.1
5	0.11
Summary	
SFW	0.15
MHW	0.16
IVW	0.16
C−IVW	0.16



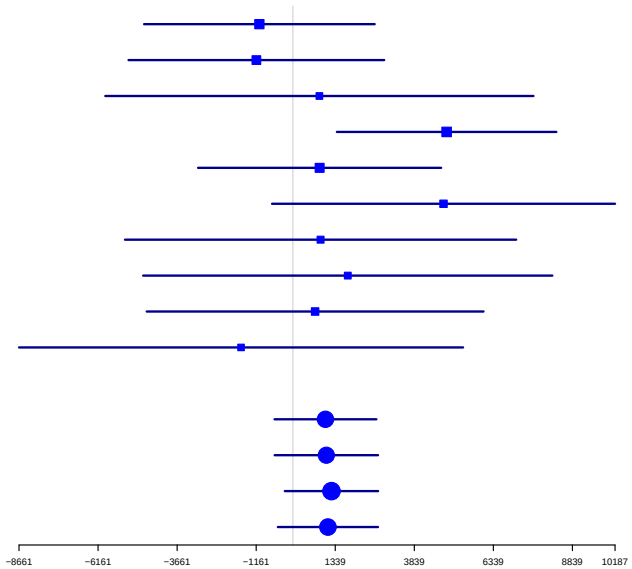
Method	ATE	Std. Error	p-value
SFW	0.15	0.034	9.2×10^{-6}
MHW	0.16	0.033	2.4×10^{-6}
IVW	0.16	0.033	1.8×10^{-6}
C-IVW	0.16	0.033	2.3×10^{-6}

The Lindner data set: cardbill with 10 strata

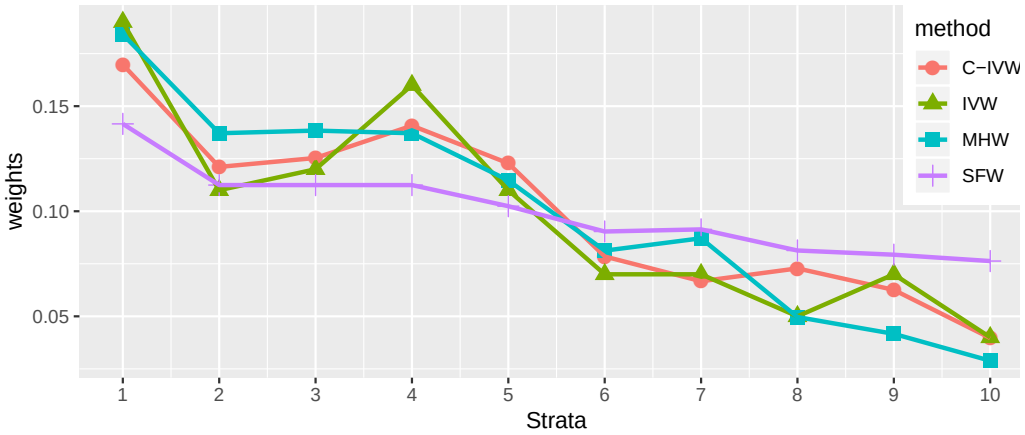


Method	ATE	Std. Error	p-value
SFW	1,029.34	822.10	0.21
MHW	1,057.04	834.22	0.20
IVW	1,219.51	752.85	0.10
C-IVW	1,107.30	807.66	0.17

Stratum	ATE
1	−1059.8
2	−1155.23
3	836.85
4	4860.85
5	844.85
6	4762.87
7	875.43
8	1734.73
9	702.49
10	−1639.55
Summary	
SFW	1029.34
MHW	1057.04
IVW	1219.51
C−IVW	1107.31

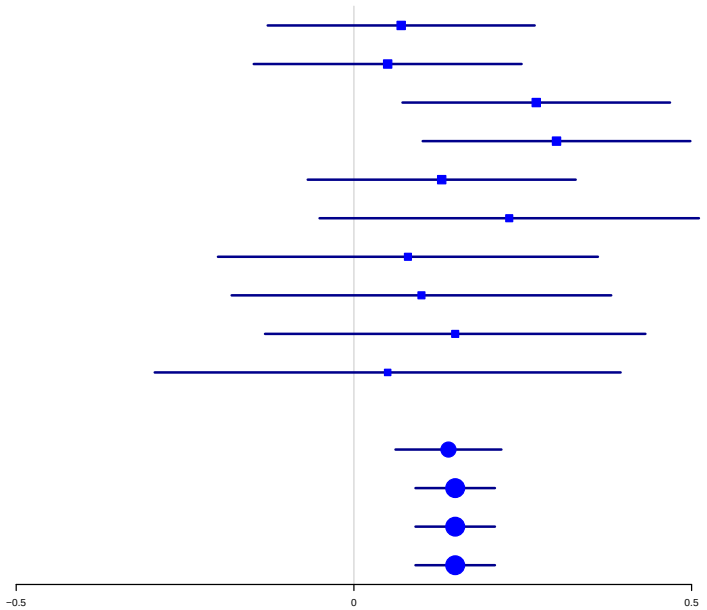


The Lindner data set: cardbill with 10 strata (log scale)



Method	ATE	Std. Error	p-value
SFW	0.14	0.04	3.2×10^{-5}
MHW	0.15	0.03	8.1×10^{-6}
IVW	0.15	0.03	4.1×10^{-6}
C-IVW	0.15	0.03	6.4×10^{-6}

Stratum	ATE
1	0.07
2	0.05
3	0.27
4	0.3
5	0.13
6	0.23
7	0.08
8	0.1
9	0.15
10	0.05
Summary	
SFW	0.14
MHW	0.15
IVW	0.15
C−IVW	0.15

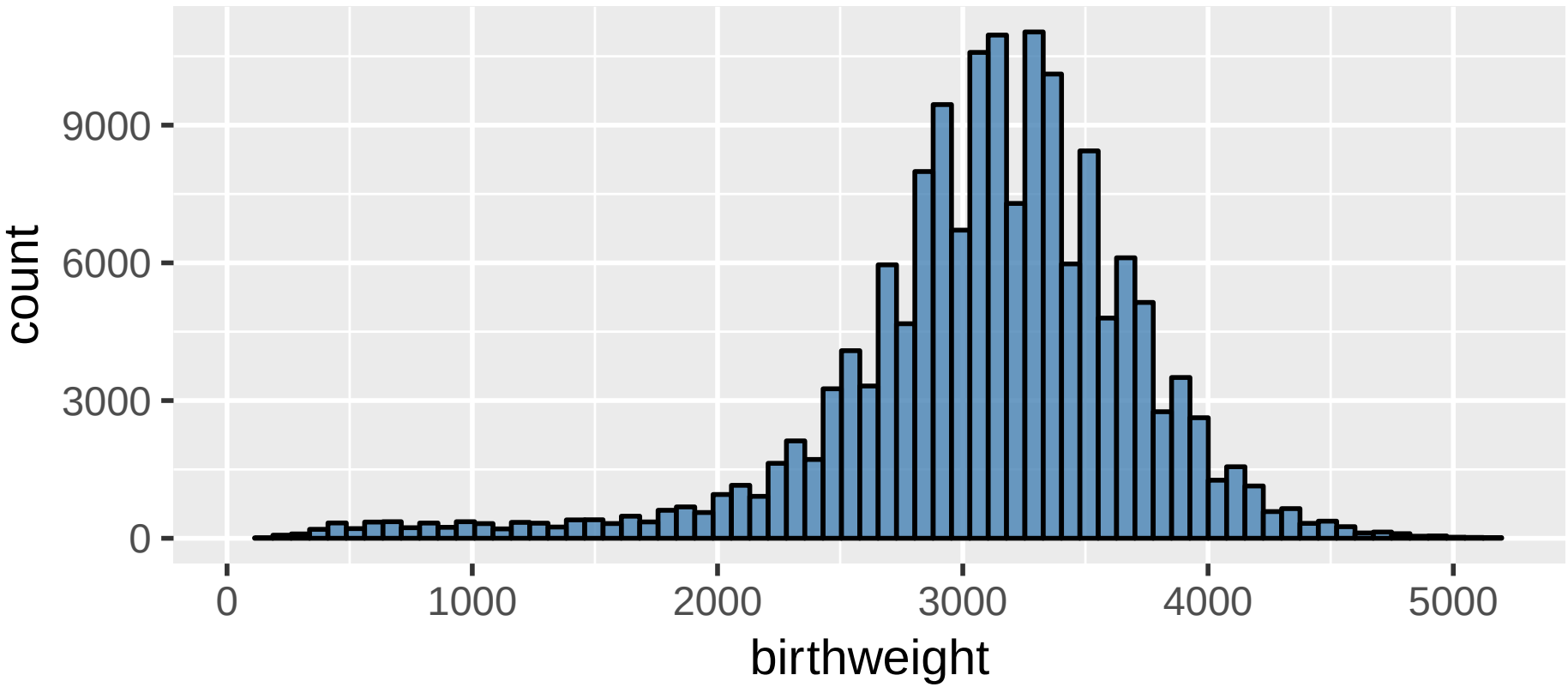


Example 2: North Carolina birth weights (1988–2002)

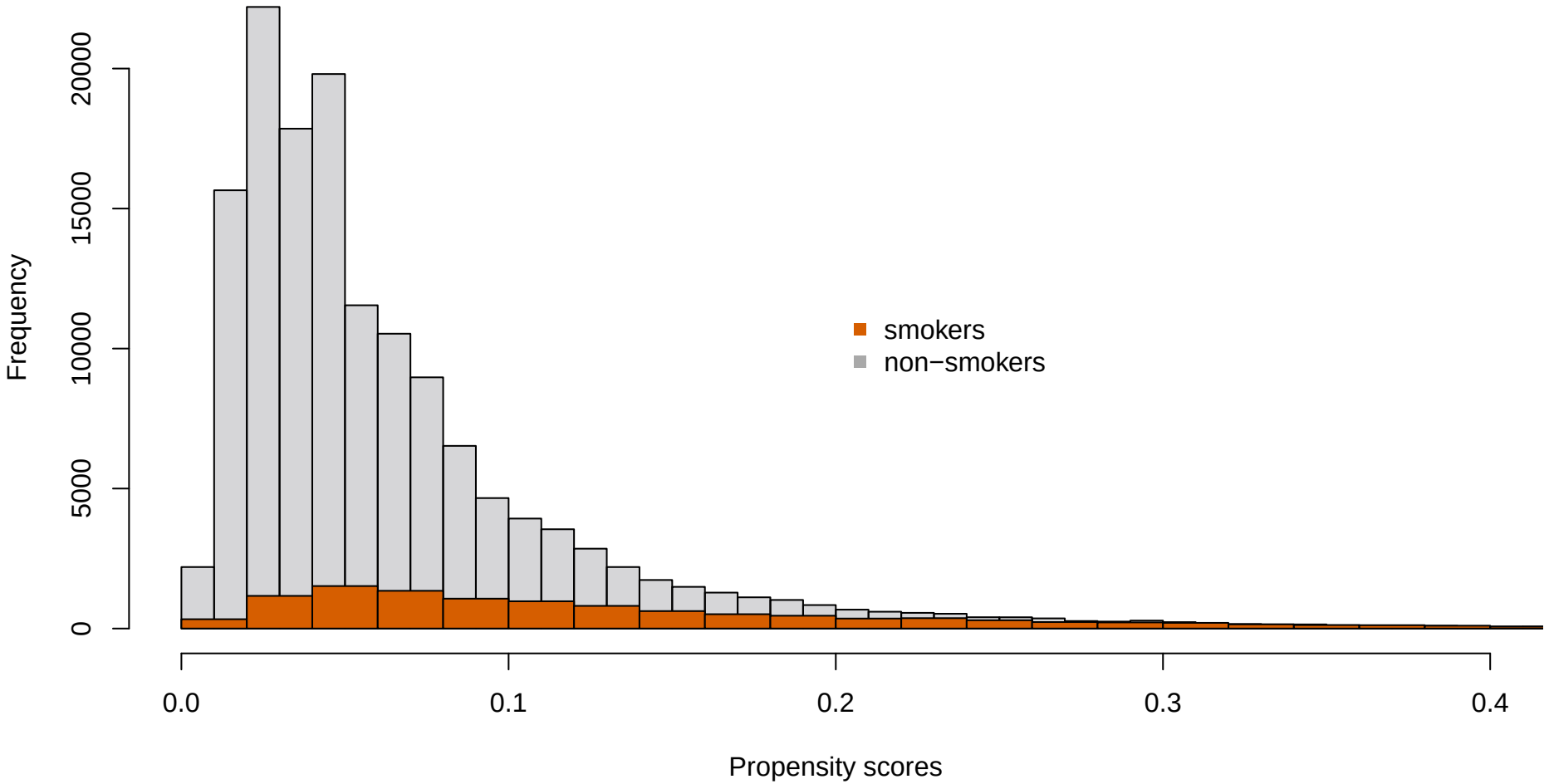
Data from Odum Institute, UNC, Chapel Hill

- 157,988 first-time black mothers
- Outcome: infants birth weights (in grams)
- Trt: smoking vs. non-smoking during pregnancy
- 1150 mothers ($\sim 7.3\%$) were smokers
- ~ 30 covariates available

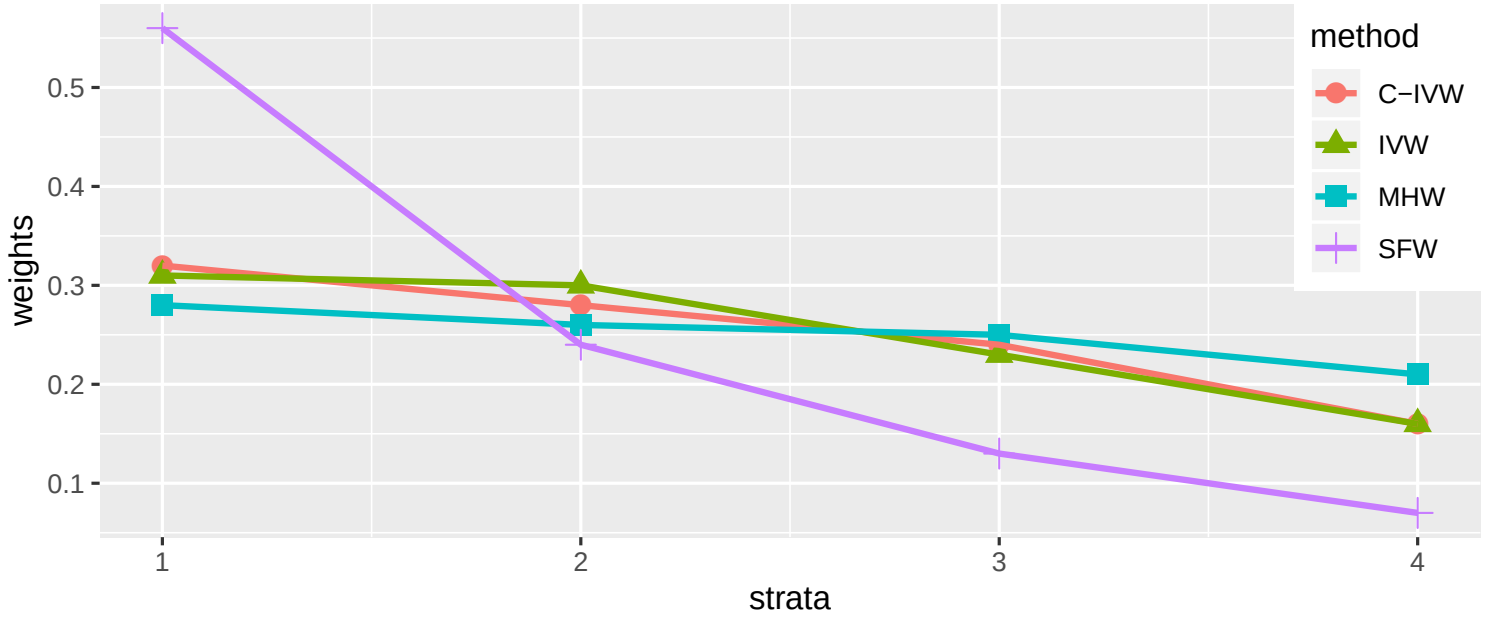
NC Birth weights: Outcome distribution

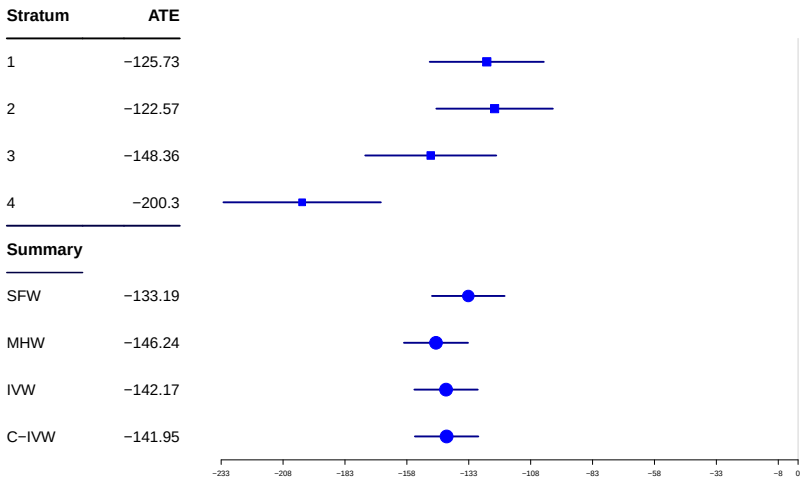


NC Birth weights: Propensity scores



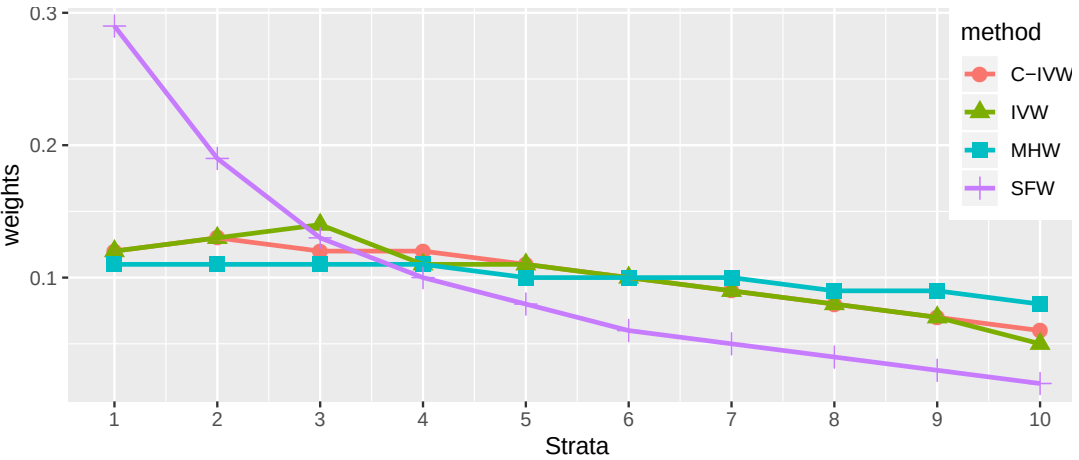
NC birth weights: 4 strata



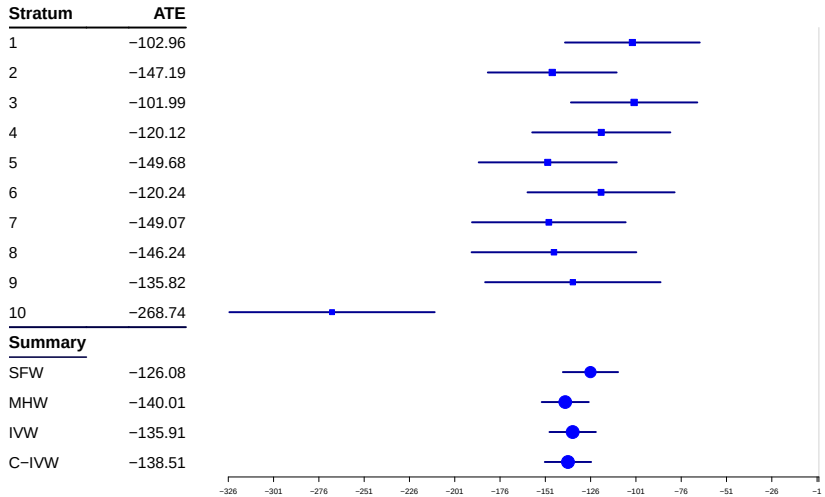


Method	ATE	Std. Error	p-value
SFW	-133.19	7.48	5.4×10^{71}
MHW	-146.24	6.60	8.3×10^{-109}
IVW	-142.17	6.53	3.5×10^{-105}
C-IVW	-141.95	6.53	9.7×10^{-107}

NC birth weights: 10 strata



Method	ATE	Std. Error	p-value
SFW	-126.08	7.77	3.3×10^{-59}
MHW	-140.01	6.62	2.9×10^{-99}
IVW	-135.91	6.53	4.03×10^{-96}
C-IVW	-138.51	6.52	4.2×10^{-100}



Summary

In **propensity score stratification,** the choice of weights is **crucial**

1. sample-fraction weights rely on stringent assumptions
2. inverse-variance weights are optimal; however
   - their implementation can go wrong (small strata, correlation mean-variance)
   - traditional bootstrap won't help
3. use the wild boostrap based on perturbation-resampling method
   (calibrate the weights and re-adjust Var(ATE))

Thank You

Roland A. Matsouaka

roland.matsouaka@duke.edu @matsouaka

References


Benedetto, U., Head, S. J., Angelini, G. D., and Blackstone, E. H. (2018). Statistical primer: propensity score matching and its alternatives. *European Journal of Cardio-Thoracic Surgery*, 53(6):1112–1117.


Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.


Jin, Z., Ying, Z., and Wei, L.-J. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, pages 381–390.


Meier, P. (1953). Variance of a weighted mean. *Biometrics*, 9(1):59–73.


Morgan, S. L. and Winship, C. (2014). *Counterfactuals and causal inference*. Cambridge University Press.


Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.


Rudolph, K. E., Colson, K., Stuart, E. A., and Ahern, J. (2016). Optimally combining propensity score subclasses. *Statistics in Medicine*, 35(27):4937–4947.