Supplementary Material for *Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments*

Miratrix, Sekhon, Yu

October 25, 2012

This supplementary material has an expanded version of the variance formula proof plus some other lemmas and smaller proofs not included in the main paper. Following these proofs are some simple toy examples that illustrate how post-stratification compares to blocking and simple-difference estimators in various circumstances.

Conditioning on $\mathcal{D}$ Maintains Assignment Symmetry

Assume the original randomization is Assignment Symmetric. The event $\mathcal{D}$ of $\hat{\tau}_{ps}$ being defined is a function of W, the vector of number of treated units in the strata:

$$\mathbf{1}_{\mathcal{D}} = f(W) \equiv \prod_{k=1}^{K} \mathbf{1}_{\{W_k>0\}} \mathbf{1}_{\{W_k<n_k\}}$$

Treatment assignment pattern T_k is independent of pattern T_i given W, so since $\mathcal{D}$ is a function of W, T_k is independent of T_i given $W, \mathcal{D}$: conditioning on $\mathcal{D}$ maintains independence of treatment assignment patterns.

Now let Ω_w be the space of possible values of W and consider two assignment patterns s and t in stratum k. We have

$$\mathbf{P}\{T_k = s|W = w\} = \mathbf{P}\{T_k = s|W_k = w_k\} = \mathbf{P}\{T_k = t|W_k = w_k\} = \mathbf{P}\{T_k = t|W = w\}$$

due to the unconditioned Assignment Symmetry. Then

$$\begin{aligned} \mathbf{P}\{T_k = s|W_k = \ell, \mathcal{D}\} &= \frac{1}{Z} \sum_{w \in \Omega_W} \mathbf{P}\{T_k = s|W = w\} \, \mathbf{1}_{\{w_k=\ell\}} \mathbf{1}_{\{f(w)=1\}} \mathbf{P}\{W = w\} \\ &= \mathbf{P}\{T_k = t|W_k = \ell, \mathcal{D}\} \end{aligned}$$

with $Z = \sum \mathbf{1}_{\{w_k=\ell\}} \mathbf{1}_{\{f(w)=1\}} \mathbf{P}\{W = w\}$. Therefore. conditioning on $\mathcal{D}$ maintains equiprobable treatment assignment patterns.

Full Derivation for Theorem 2.1

This section fills out details missing in the proof of Theorem 2.1 in Appendix A of the main paper. As discussed there, under Assignment Symmetry the chance of any given unit being treated is $W_k(1)/n_k$ so

$$\mathbb{E}\left[T_i|W_k(1)\right] = \frac{W_k(1)}{n_k}$$

for unit i in stratum k. Then

$$\mathbb{E}\left[\frac{T_i}{W_k(1)}\right] = \mathbb{E}\,\mathbb{E}\left[\frac{T_i}{W_k(1)}|W_k(1)\right] = \mathbb{E}\left[\frac{1}{n_k}\right] = \frac{1}{n_k}.$$

Rearrange $\beta_{1k} \equiv \mathbb{E}\left[W_k(0)/W_k(1)\right] = n_k\,\mathbb{E}\left[1/W_k(1)\right] - 1$ to get $\mathbb{E}\left[1/W_k(1)\right] = (\beta_{1k}+1)/n_k$ and

$$\mathbb{E}\left[\frac{T_i^2}{W_k^2(1)}\right] = \mathbb{E}\,\mathbb{E}\left[\frac{T_i}{W_k^2(1)}|W_k(1)\right] = \frac{1}{n_k}\,\mathbb{E}\left[\frac{1}{W_k(1)}\right] = \frac{\beta_{1k}+1}{n_k^2}.$$

These derivations are easier if we use $\alpha_{1k} \equiv \mathbb{E}\left[1/W_k(1)\right]$, but the β's are more interpretable and lead to nicer final formula. To continue, Assignment Symmetry gives

$$\begin{aligned}
\mathbb{E}\left[T_iT_j|W_k(1)=w\right] &= \mathbf{P}\{T_i=1 \wedge T_i=1|W_k(1)=w\} \\
&= \frac{\binom{n_k-2}{w-2}}{\binom{n_k}{w}} = \frac{(n_k-2)!}{(w-2)!(n_k-w)!}\cdot\frac{w!(n_k-w)!}{n_k!} \\
&= \frac{w(w-1)}{n_k(n_k-1)}
\end{aligned}$$

so

$$\mathbb{E}\left[\frac{T_iT_i}{W_k^2(1)}\right] = \mathbb{E}\left[\frac{W_k(1)(W_k(1)-1)}{W_k^2(1)}\right]\cdot\frac{1}{n_k(n_k-1)} = \frac{-\beta_{1k}+n-1}{n_k^2(n_k-1)}.$$

There are analogous formula for the control unit terms. Similarly,

$$\mathbb{E}\left[\frac{T_i(1-T_i)}{W_k(1)W_k(0)}\right] = \mathbb{E}\left[\frac{W_k(1)(n_k-W_k(1))}{W_k(1)W_k(0)}\right]\cdot\frac{1}{n_k(n_k-1)} = \frac{1}{n_k(n_k-1)}.$$

We use these relationships to compute means and variances for the strata-level estimators.

Unbiasedness. The strata-level estimators are unbiased:

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_k] &= \mathbb{E}\left[\sum_{i:b_i=k}\frac{T_i}{W_k(1)}y_i(1) - \sum_{i:b_i=k}\frac{1-T_i}{W_k(0)}y_i(0)\right] \\
&= \sum_{i:b_i=k}\mathbb{E}\left[\frac{T_i}{W_k(1)}\right]y_i(1) - \sum_{i:b_i=k}\mathbb{E}\left[\frac{1-T_i}{W_k(0)}\right]y_i(0) \\
&= \sum_{i:b_i=k}\frac{1}{n_k}y_i(1) - \sum_{i:b_i=k}\frac{1}{n_k}y_i(0) = \tau_k.
\end{aligned}$$

Variance. $\text{Var}[\hat{\tau}_k] = \mathbb{E}[\hat{\tau}_k^2] - \tau^2$. $\mathbb{E}[\hat{\tau}_k^2]$ breaks down into three big terms:

$$\begin{aligned}\mathbb{E}[\hat{\tau}_k^2] &= \underbrace{\mathbb{E}\left[\sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1)\right]^2}_{(a)} \\ &- \underbrace{2\,\mathbb{E}\left[\left(\sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1)\right)\left(\sum_{i:b_i=k} \frac{1-T_i}{W_k(0)} y_i(0)\right)\right]}_{(b)} + \underbrace{\mathbb{E}\left[\sum_{i:b_i=k} \frac{1-T_i}{W_k(0)} y_i(0)\right]^2}_{(c)}.\end{aligned}$$

Simplify the three parts of the above. For part (a):

$$\begin{aligned}(a) &= \mathbb{E}\left[\sum_{i:b_i=k} \frac{T_i^2}{n_k^2(1)} y_i^2(1) + \sum_{i\neq j} \frac{T_i T_i}{n_k^2(1)} y_i(1) y_i(1)\right] \\ &= \sum_{i:b_i=k} \mathbb{E}\left[\frac{T_i^2}{n_k^2(1)}\right] y_i^2(1) + \sum_{i\neq j} \mathbb{E}\left[\frac{T_i T_i}{n_k^2(1)}\right] y_i(1) y_i(1) \\ &= \frac{\beta_{1k}+1}{n_k^2} \sum_{i:b_i=k} y_i^2(1) + \frac{-\beta_{1k}+n_k-1}{n_k^2(n_k-1)} \sum_{i\neq j} y_i(1) y_i(1).\end{aligned}$$

Part (c) is similar. The cross-terms are:

$$\begin{aligned}(b) &= 2\,\mathbb{E}\left[\sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1) \frac{1-T_i}{W_k(0)} y_i(0)\right] + 2\,\mathbb{E}\left[\sum \frac{T_i}{W_k(1)} y_i(1) \frac{1-T_i}{W_k(0)} y_i(0)\right] \\ &= 0 + 2\sum_{i\neq j} \mathbb{E}\left[\frac{T_i}{W_k(1)} \frac{1-T_i}{W_k(0)}\right] y_i(1) y_i(0) \\ &= \frac{2}{n_k(n_k-1)} \sum_{i\neq j} y_i(1) y_i(0).\end{aligned}$$

The first term vanishes since $T_i(1-T_i) = 0$ always.

These are the three parts of the expectation of the square. We have related components in τ_k^2 when you expand the square:

$$\tau_k^2 = \underbrace{\left(\sum_{i:b_i=k} \frac{1}{n_k} y_i(1)\right)^2}_{(a')} - \underbrace{2\left(\sum_{i:b_i=k} \frac{1}{n_k} y_i(1)\right)\left(\sum_{i:b_i=k} \frac{1}{n_k} y_i(0)\right)}_{(b')} + \underbrace{\left(\sum_{i:b_i=k} \frac{1}{n_k} y_i(0)\right)^2}_{(c')}.$$

The variance is $\mathrm{Var}[\hat{\tau}_k] = (a)-(a')-(b)+(b')+(c)-(c')$, a sum of several ugly differences. Expanding (a') and plugging in gives the first difference:

$$\begin{aligned}
(a)-(a') &= \frac{\beta_{1k}+1}{n_k^2}\sum_{i:b_i=k} y_i^2(1) + \frac{-\beta_{1k}+n_k-1}{n_k^2(n_k-1)}\sum_{i\neq j} y_i(1)y_i(1) - \frac{1}{n_k^2}\sum_{i:b_i=k} y_i^2(1) - \frac{1}{n_k^2}\sum_{i\neq j} y_i(1)y_i(1) \\
&= \left(\frac{\beta_{1k}+1}{n_k} - \frac{1}{n_k^2}\right)\sum_{i:b_i=k} y_i^2(1) + \left(\frac{-\beta_{1k}+n_k-1}{n_k^2(n_k-1)} - \frac{1}{n_k^2}\right)\sum_{i\neq j} y_i(1)y_i(1) \\
&= \frac{\beta_{1k}}{n_k}\left[\frac{1}{n_k}\sum_{i:b_i=k} y_i^2(1) - \frac{1}{n_k(n_k-1)}\sum_{i\neq j} y_i(1)y_i(1)\right] \\
&= \frac{\beta_{1k}}{n_k}\sigma_k^2(1).
\end{aligned}$$

$(c)-(c')$ is similar. The cross terms are:

$$\begin{aligned}
(b)-(b') &= \frac{2}{n_k(n_k-1)}\sum_{i\neq j} y_i(1)y_i(0) - \frac{2}{n_k^2}\sum_{i:b_i=k} y_i(1)y_i(0) - \frac{2}{n_k^2}\sum_{i\neq j} y_i(1)y_i(0) \\
&= \left(\frac{2}{n_k(n_k-1)} - \frac{2}{n_k^2}\right)\sum_{i\neq j} y_i(1)y_i(0) - \frac{2}{n_k^2}\sum_{i:b_i=k} y_i(1)y_i(0) \\
&= -\frac{2}{n_k}\left[\frac{1}{n_k}\sum_{i:b_i=k} y_i(1)y_i(0) - \frac{1}{n_k(n_k-1)}\sum_{i\neq j} y_i(1)y_i(0)\right] \\
&= -\frac{2}{n_k}\gamma_k(0,1).
\end{aligned}$$

Sum the above to get the variance formula.

Proof of Lemma 10.2

We restate Lemma 10.2 for reference:

Lemma 1. *Let W be a Binomial (n,p) random variable or a hypergeometric (n,w,N) random variable, i.e., a sample of size n from coin flips with probability of heads p or an urn with $N = nc$ balls, $c > 1$, of which $w = ncp$ are white. Then for $Y = (n/W)\mathbf{1}_{\{W>0\}}$:*

$$-\frac{2(1-p)^n}{p} \leq \mathbb{E}[Y] - \frac{1}{p} \leq \frac{4}{p^2}\frac{1}{n} - \frac{1}{n+1}\frac{1}{p} + \max\left[\left(\frac{n}{2} - \frac{4}{p^2 n}\right)\exp\left(-\frac{p^2}{2}n\right), 0\right].$$

Proof: First we derive the lower bound on the expectations. For ease of notation, define $\mathbb{E}[X;\mathcal{A}]$ as the expectation of $X\mathbf{1}_{\{A\}}$. For both Bernoulli assignment or complete randomization,

$$np = \mathbb{E}[W] = \mathbb{E}[W; W>0] = \mathbb{E}[W|W>0]\,\mathbf{P}\{W>0\}.$$

Also, $\mathbf{P}\{W=0\} \leq (1-p)^n$. For a random variable $X > 0$, $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$. Therefore

$$\begin{aligned}
\mathbb{E}\left[\frac{n}{W}; W>0\right] &= \mathbb{E}\left[\frac{n}{W}|W>0\right]\mathbf{P}\{W>0\} \\
&\geq \frac{n}{\mathbb{E}[W; W>0]}\mathbf{P}\{W>0\}^2 \\
&= \frac{1}{p} + \frac{1}{p}\left(\mathbf{P}\{W>0\}^2 - 1\right) \\
&\geq \frac{1}{p} - \frac{2}{p}(1-p)^n
\end{aligned}$$

For the upper bound, expand $\mathbb{E}[n/W; W>0]$ into two terms and analyze each term. Namely, we will show that $\mathbb{E}[n/W; W>0] = I + D$ with

$$I \equiv \mathbb{E}\left[\frac{n}{W+1}; W>0\right] \leq \frac{1}{p} - \frac{1}{n+1}\frac{1}{p}$$

and

$$\begin{aligned}
D &\equiv \mathbb{E}\left[\frac{n}{W} - \frac{n}{W+1}; W>0\right] = \mathbb{E}\left[\frac{n}{W(W+1)}; W>0\right] \\
&\leq \min_{0<\alpha_n<p}\left\{\frac{1}{n}\left(\frac{1}{p-\alpha_n}\right)^2 + \max\left[\left(\frac{n}{2} - \frac{1}{n}\left(\frac{1}{p-\alpha_n}\right)^2\right)\exp(-2n\alpha_n^2), 0\right]\right\}
\end{aligned}$$

Instead of minimizing the bound across the possible values of α_n, we can simply fix $\alpha_n = p/2$ to obtain a looser, but more intelligible, bound:

$$D \leq \frac{4}{p^2}\frac{1}{n} + \max\left[\left(\frac{n}{2} - \frac{4}{np^2}\right)\exp(-\frac{p^2}{2}n), 0\right].$$

We show D first. Let α_n be in $(0, p)$. Then:

$$\begin{aligned}
D &= \mathbb{E}\left[\frac{n}{W(W+1)}; p - \frac{W}{n} < \alpha_n\right] + \mathbb{E}\left[\frac{n}{W(W+1)}\mathbf{1}_{\{W>0\}}; p - \frac{W}{n} \geq \alpha_n\right]. \\
&\leq \frac{1}{n}\,\mathbb{E}\left[\left(\frac{n}{W}\right)^2; p - \alpha_n < \frac{W}{n}\right] + \frac{n}{2}\mathbf{P}\left\{p - \frac{W}{n} \geq \alpha_n\right\}.
\end{aligned}$$

The $n/2$ is because $W > 0$ implies $W \geq 1$.

Hoeffding (1963) famously bounded the tail probabilities of sums of independent random variables, allowing us to control the probability of W/n being far from p. He also, in section 6 of the same work, generalized his bound to the hypergeometric. We use both of these results:

$$\mathbf{P}\left\{p - \frac{W}{n} \geq \alpha_n\right\} \leq \exp(-2n\alpha_n^2).$$

Because $0 < p - \alpha_n < W/n$ we have

$$\begin{aligned} D &\leq \frac{1}{n}\left(\frac{1}{p-\alpha_n}\right)^2\left(1 - \mathbf{P}\left\{p - \frac{W}{n} \geq \alpha_n\right\}\right) + \frac{n}{2}\mathbf{P}\left\{p - \frac{W}{n} \geq \alpha_n\right\} \\ &\leq \frac{1}{n}\left(\frac{1}{p-\alpha_n}\right)^2 + \max\left[\left(\frac{n}{2} - \frac{1}{n}\frac{1}{(p-\alpha_n)^2}\right)\exp(-2n\alpha_n^2), 0\right]. \end{aligned}$$

The $\max(\cdot,\cdot)$ comes from the choice of α_n possibly making $n/2 - 1/n(p-\alpha_n)^2 < 0$ which would invert the Hoeffding bound. We instead conservatively set this quantity to 0.

To evaluate I, consider the Binomial case first. Express the expectations as a sum and re-index the sum and add in the first two terms to get the sum of the distribution of a $(n+1, p)$ binomial variable:

$$\begin{aligned} I \equiv \mathbb{E}\left[\frac{n}{W+1}; W > 0\right] &= \sum_{k=1}^{n} \frac{n}{k+1}\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\ &= \frac{n}{n+1}\frac{1}{p}\left(\sum_{k=0}^{n+1}\frac{(n+1)!}{k!(n+1-k)!}p^k(1-p)^{n+1-k} - (1-p)^{n+1} - (n+1)p(1-p)^n\right) \\ &= \frac{n}{n+1}\frac{1}{p}\Big(1 - (1-p)^{n+1} - (n+1)p(1-p)^n\Big) \\ &= \frac{1}{p} - \frac{1}{n+1}\frac{1}{p} - \frac{n}{n+1}\frac{1}{p}(np+1)(1-p)^n. \end{aligned}$$

This is exact for the Binomial case. To extend to complete randomization, we use a further result from Hoeffding. Hoeffding showed that, for a continuous, convex function $f(x)$, $\mathbb{E}_{srs}[f(W)] \leq \mathbb{E}_{bin}[f(W)]$. Let $f(x)$ be $n/(x+1)$. $f(x)$ is continuous, convex for $x \geq 0$. Furthermore for Binomial W

$$\mathbb{E}\left[\frac{n}{W+1}; W > 0\right] + n(1-p)^n = \mathbb{E}[f(W)]$$

as $n/(W+1) = f(W)$ for all W. So

$$\mathbb{E}_{srs}[\frac{n}{W+1}; W > 0] \leq \mathbb{E}_{srs}[f(W)] \leq \mathbb{E}_{bin}[f(W)] = \mathbb{E}_{bin}[\frac{n}{W+1}; W > 0] + n(1-p)^n$$

Thus we gain an extra (small) $n(1-p)^n$ term to bound I, but this term is more than offset by the negative term $n/(n+1) \times (np+1)/p \times (1-p)^n$ and so we drop both.

To get the overall bound, sum the bounds for I and D. □

Remarks: As a side note, Serfling (1974) improves Hoeffding's bound for sampling without replacement, implying that the rate of the βs convergence is faster under complete randomization than for Bernoulli.

1 Toy Examples of Gain and Loss

In this section we provide a few small examples to demonstrate the potential for gain or loss due to post-stratification. Each of the following scenarios specify a particular collection of potential outcomes, and Table 1 shows the resulting variances of the unadjusted estimator and post-stratified estimator (both conditioned on $\mathcal{D}$). In all cases we assume complete randomization with $100p\%$ units treated, with p as stated on the table. Table 1 also shows the variance if the randomization were done via blocking.

We plug the parameters defined by the stated population into the variance formulas presented in the main paper. We numerically compute the βs by conducting the described randomization $50,000$ times and computing the mean βs for those randomizations where all strata estimators were defined (i.e., we condition on $\mathcal{D}$). The results on Table 1 are exact up to the uncertainty in computing the βs. Bernoulli randomization gives near-identical results (since the β's are near identical). Directly estimating variance with a monte-carlo of point estimates also gives identical results up to sampling error, further validating the formula as correct.

						variances			% gain/loss		
	n	K	p	$\mathbf{P}\mathcal{D}$	τ	$\hat{\tau}_{ps}$	$\hat{\tau}_{sd}$	blk	blk:ps	sw:ps	sw:blk
I.A	40	4	0.50	99.9%	1.00	1.01	1.36	0.92	-7%	26%	33%
I.B	40	4	0.50	99.9%	1.00	1.01	0.85	0.92	-11%	-20%	-8%
I.C	40	4	0.30	93.7%	1.00	1.28	1.62	1.09	-12%	21%	33%
I.D	40	4	0.50	99.9%	1.00	1.01	2.24	0.92	-4%	55%	59%
I.E	40	4	0.50	99.9%	1.00	1.01	0.91	0.92	-10%	-11%	-1%
II.A	100	4	0.50	99.9%	1.91	0.39	0.63	0.37	-3%	39%	42%
II.B	100	4	0.30	97.2%	1.91	0.52	0.80	0.47	-5%	35%	41%
III.A	200	2	0.50	100%	0.94	0.24	0.30	0.24	0%	21%	21%
III.B	200	5	0.50	100%	0.94	0.21	0.30	0.21	-2%	28%	30%
III.C	200	10	0.50	100%	0.94	0.22	0.30	0.21	-4%	25%	29%
III.D	200	20	0.50	97%	0.94	0.24	0.30	0.21	-10%	20%	30%
III.E	200	25	0.50	84.4%	0.94	0.25	0.30	0.21	-13%	18%	30%
IV.A	200	2	0.50	100%	0.94	0.30	0.30	0.30	-1%	0%	0%
IV.B	200	5	0.50	100%	0.94	0.31	0.30	0.30	-2%	-2%	0%
IV.C	200	10	0.50	100%	0.94	0.32	0.30	0.30	-5%	-7%	-2%
IV.D	200	20	0.50	97%	0.94	0.35	0.30	0.31	-14%	-17%	-3%
IV.E	200	25	0.50	84.4%	0.94	0.37	0.30	0.31	-19%	-23%	-4%
V.A	100	4	0.50	99.9%	1.80	0.32	0.55	0.30	-4%	42%	45%
V.B	200	4	0.50	100%	1.80	0.15	0.28	0.15	-2%	45%	47%
V.C	400	4	0.50	100%	1.80	0.07	0.14	0.07	-1%	47%	47%
V.D	800	4	0.50	100%	1.80	0.04	0.07	0.04	0%	47%	48%

Table 1. Variances of Estimators for Several Scenarios. K is the number of strata. $\mathbf{P}\mathcal{D}$ is the probability of $\hat{\tau}_{ps}$ being defined, estimated by simulation. τ is actual SATE. The percentages are calculated as $100\% \times \Delta/\text{Var}[\hat{\tau}_{sd}]$ with Δ being the specified difference between variances.

The families of scenarios are as follows:

I) We first consider a simple experiment with four strata, A, B, C, and D, with 10 units each.

   A) In the first scenario, the units in strata B, C and D are replicates of A shifted up by $+2$, $+4$, and $+6$, respectively. There is a constant treatment effect of $+1$. There is substantial between-strata variation, and therefore post-stratification is beneficial. The left plot in Figure 1 displays the relationship between potential outcomes and strata. This is the idealized constant-treatment effect situation where stratification separates units of different types.

   B) As Scenario I.A, but now the units in B, C, and D are simple replicates of A's units not shifted. There is no difference between strata and so we see the full price paid by spurious post-stratification. It is easy in this small experiment for a random imbalance to occur. An imbalance overweights some of the units, making it easier to reach extreme values for estimated treatment effect. This results in a larger variance. In this scenario blocking is also a poor choice, incurring a small cost.

   C) As Scenario I.A, but with probability of treatment $p = 0.3$. The small proportion treated makes it easier to have very few units estimating the average treatment effect in a stratum (or overall). All estimators' variances increase, but post-stratification still comes out ahead of simple difference.

   D) Now we have differing treatment effects of $-3, 0, +2$, and $+5$ for the four strata. We no longer have an overall constant treatment effect: different strata respond to treatment differently. Here the between-strata correlation of potential outcomes is near 1.00. This makes post-stratification work very well.

   E) A reverse of Scenario 1.D, we now have differing treatment effects of $+5, +2, 0, -3$ for the four strata. The trend of the group means is opposite to the trend of outcomes within groups, which causes problems. The between-strata correlation of potential outcomes is -0.95. See right plot on Figure 1. The between-strata term is small due to this negative correlation. Negative correlations are good for randomization because it means that if a randomly high unit is put into treatment, a randomly high unit will probably be put into control as well to compensate. Post-stratification does not take advantage of this, and thus does more poorly than the unadjusted estimator, which does. Blocking also does not fair well in this case for the same reasons.

   In all the above, which are for a small-sized experiment, post-stratification is somewhat close to blocking.

II) A slightly more complex experiment with unequal strata sizes. A has 60 units, B has 15, C has 15 and D has 10. We drew the $y_k(0)$ for A from a $N(5, 5)$ population, B from a $N(3, 10)$, C from a $N(7, 15)$ and D from a $N(3, 15)$, where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The treatment effects for all units, drawn from a $unif(-1, 5)$ distribution, were added to the control outcomes. The presented

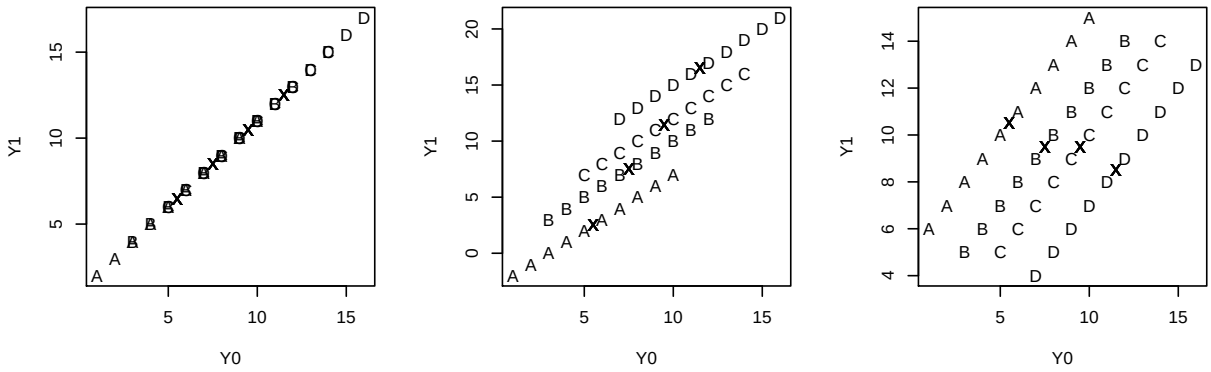

Figure 1. From left to right, the potential outcomes for scenarios I.A, I.D, and I.E. Strata membership on the plots are denoted A through D. "X" denotes strata means.

results are the variances of the estimators under different randomizations of a *single sample* drawn from this described population.

A) Equal treatment proportions of $p = 0.5$. Post-stratification helps. It is also close to blocking.

B) $p = 0.3$. The efficacy declines slightly due to the increased chance of imbalance. Blocking does not suffer as much.

III) A set of experiments with a continuous covariate z evenly spaced on the interval $[0, 100]$ which we then partition into K strata of equal sizes. We vary K to see the impact of finer stratification. The control outcome for unit i is distributed as $y_i(0) \sim N(\sqrt{z_i}, 9)$ and the treatment outcome as $y_i(1) \sim N(y_i(0) + 1, 1)$. About 5 strata seems ideal although even two strata is far better than doing nothing. Too many strata and we see less benefit, plus a large increase in the chance of an undefined estimator.

IV) As III, but now z is useless. We generate this set by permuting the observed z from III, breaking any connection between the covariate and the outcomes. $\hat{\tau}_{sd}$ is completely unaffected. As the number of strata increase, things worsen for post-stratification due to the increased chance of an accidental imbalance giving a single unit a great deal of weight. Blocking also suffers, but not by nearly as much.

V) In this set of experiments, the set-up being the same as for Experiment II, we first generated an initial set of data, and then replicated the units within the strata to increase n. The number of strata is thus held constant and the treatment effect, covariances and variances for subsequent experiments remain essentially unchanged. As n grows, the percentage increase in variance of $\hat{\tau}_{ps}$ over blocking converges to 0 at rate $1/n$, and thus the percentage gain over $\hat{\tau}_{sd}$ converges to a fixed relative improvement in precision over the unadjusted estimate.

Discussion. Generally speaking, post-stratification is similar to blocking in terms of efficiency. The more strata, however, the worse this comparison becomes due to the increased chance of severe imbalance with consequential increase high uncertainty in the stratum-level estimates. Post-stratification's overall efficacy depends on how much larger the between-stratum variation is compared to the penalty paid by giving some observations greater weight due to random assignment imbalance. Having many strata is generally not helpful and can be harmful if b is not prognostic. A moderate number of strata seems to offer protection from this: compare $K = 5$ for scenarios III and IV.

1.1 Examining Conditional Variance

To illustrate how the variance of the estimators conditioned on the split W varies, we repeatedly conduct a randomization for a specific sample and calculate the conditional MSE for both estimators given the generated split as shown in the latter half of Section 7 of the main paper. These simulations demonstrate that if b is indeed prognostic, then the MSE of $\hat{\tau}_{ps}$ is far lower than that of $\hat{\tau}_{sd}$, and this difference increases with degree of imbalance. However, if b is not prognostic, then the reverse trend is evident. The post-stratified estimator does worse in the very circumstance when people might use it: to adjust for a seen imbalance in the randomization. It is not necessarily beneficial to adjust—the variable used for adjustment must be selected with care.

The left side of Figure 2 shows 5000 such calculations for Scenario III.B, presented above. With low imbalance, the variance of $\hat{\tau}_{ps}$ is even smaller than the unconditional formula would suggest. But as imbalance deteriorates, the variance of $\hat{\tau}_{ps}$ increases.

Compared to $\hat{\tau}_{ps}$, the simple-difference estimator $\hat{\tau}_{sd}$ is vulnerable to poor splits. Generally, high imbalance means high conditional MSE. This is due to the bias term which can get exceedingly large if there is imbalance between different heterogeneous strata. We see a similar trend to the analogous PAC-Man example in the primary paper.

If b is not prognostic, however, the story changes. The experimental units in Scenario IV.B, shown on right of Figure 2, are the same as for Scenario III.B, but the elements of the covariate vector b have been shuffled to break b's prognostic ability. Because the units are the same, the unconditional variance of $\hat{\tau}_{sd}$ is the same as well. Because b is no longer prognostic, post-stratification does not help, as illustrated by the elevated unconditional and conditional trend lines. The post-stratified estimator still worsens with greater imbalance as it did before because the cost of imbalance comes from the number of observations in the treatment and control groups, something unrelated to b. The simple-difference estimator, however, often can even improve with large imbalance. This is due to imbalance ensuring a greater comparability of treatment and control units—if it were known that b was not connected to the potential outcomes then it would actually be most ideal to treat all of some strata and none of the others.

In other scenarios (not shown) these trends are repeated. Furthermore, when there are few strata, the imbalance tends to be low (e.g., Scenarios I and II, or III with small K) with a heavily right skewed distribution of conditional variance—most of the time there is a good balance and low conditional variance, but there is a low chance of a bad split and a high conditional variance. In such circumstances, there is very a good chance that

the conditional variance of a post-stratified experiment is even closer to its corresponding blocked experiment than one would initially expect from Equation 12 in the main document. Also in such circumstances the pattern of the MSE of $\hat{\tau}_{sd}$ worsening for prognostic b and improving for unrelated b as imbalance increases is even more apparent.

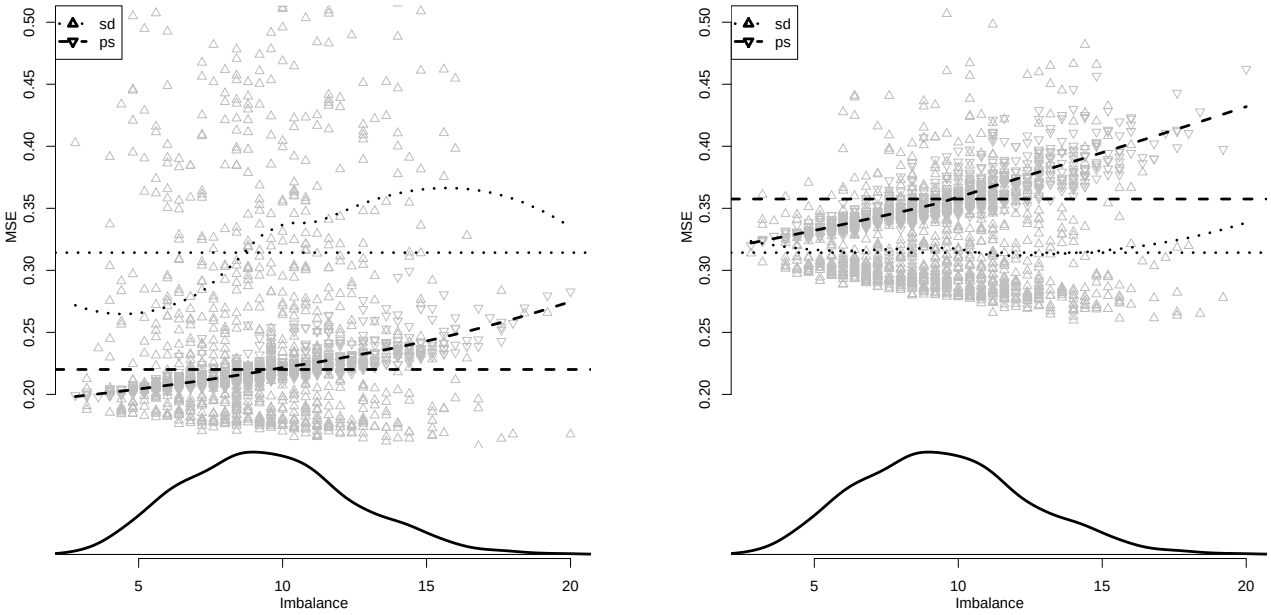

Figure 2. Conditional Variance of Scenario III.D (left) and Scenario IV.D (right). Points indicate the conditional MSE of $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ given various specific splits of W. x-axis is the imbalance score for the split. Curved dashed lines interpolate point clouds. Horizontal dashed lines mark unconditional variances for the two estimators. The curves at bottom are the densities of the imbalance statistic.

References

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*(301), 13–30.

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics 2*(1), 39–48.