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

Luke W. Miratrix† Jasjeet S. Sekhon‡ Bin Yu§

September 23, 2011 (11:47)

Abstract

Experimenters often use post-stratification to adjust estimates. Post-stratification is akin to blocking, except that the number of treated units in each strata is a random variable because stratification occurs after treatment assignment. We analyze both post-stratification and blocking under the Neyman model and compare the efficiency of these designs. We derive the variances for a post-stratified estimator and a simple difference-in-means estimator under different randomization schemes. Post-stratification is nearly as efficient as blocking: the difference in their variances is on the order of $1/n^2$, provided treatment proportion is not too close to 0 or 1. Post-stratification is therefore a reasonable alternative to blocking when the latter is not feasible. However, in finite samples, post-stratification can increase variance if the number of strata is large and the strata are poorly chosen. To examine why the estimators' variances are different, we extend our results by conditioning on the observed number of treated units in each strata. Conditioning also provides more accurate variance estimates because it takes into account how close (or far) a realized random sample is from a comparable blocked experiment. We then show that the practical substance of our results remain under an infinite population sampling model. Finally, we provide an analysis of an actual experiment to illustrate our analytical results.

Keywords: Neyman model; SATE and PATE estimation; regression adjustment; randomized trials

*Winston Lin provided much useful feedback, several references, and excellent editing. Luke Miratrix is grateful for the support of a Graduate Research Fellowship from the National Science Foundation.

†Ph.D. candidate, Department of Statistics, UC Berkeley. Email: `luke@stat.berkeley.edu`. Mail: 367 Evans Hall #3860, UC Berkeley, Berkeley, CA 94720

‡Associate Professor, Travers Department of Political Science, Department of Statistics and Director, Center for Causal Inference and Program Evaluation, Institute of Governmental Studies, UC Berkeley, `http://sekhon.berkeley.edu`

§Professor, Department of Statistics and EECS, UC Berkeley, Berkeley, CA 94720

1 Introduction

Arguably one of the most important tools for determining the causal effect of some action is the randomized experiment, where a researcher randomly divides units into groups and applies different treatments to each group. For example, for testing the efficacy of a drug, researchers would give the drug to one random portion of a patient population and compare their outcomes to a seperate, control population that was given a placebo or alternate, baseline drug. Randomized experiments are the "gold standard" for causal inference because, assuming proper implementation of the experiment, if a difference in outcomes is found the only possible explanations are a significant treatment effect or random chance. Math gives a handle on the chance, which allows for principled inference about the treatment effect. In the most basic analysis, a simple difference in means is used to estimate the overall sample average treatment effect (SATE), defined as the average difference in the units' outcomes if all were treated as compared to the average outcomes if they were not. This framework and estimator were analyzed by Neyman in 1923 (but see the English translation, Splawa-Neyman et al. (1990)) under what is now called the Neyman model or potential outcomes framework. Under the Neyman model, one need make almost no assumptions not guaranteed by the randomization itself.

For an experiment each observation usually comes at a cost, however, so it is desirable to find more informative ways than the simple difference-in-means estimator, in terms of a smaller variance, to measure treatment effect. Blocking, which is when experimenters first stratify their units and then randomize treatment within pre-defined blocks, can greatly reduce variance compared to the simple difference estimator if the strata differ from each other (Imai et al., 2008). However, because blocking must be conducted before randomization, it is often not feasible due to practical considerations or lack of foresight. Sometimes randomization may even be entirely out of the researcher's control, such as with so-called natural experiments. When blocking was not done, researchers often adjust for covariates after randomization. For example, Pocock et al. (2002) studied a sample of clinical trials and found that 72% of the articles used covariate adjustment. Keele et al. (2009) analyzed the experimental results in three major political science journals and found that 74%-95% of the articles relied on adjustment. Post-stratification is one simple form of adjustment where the researcher stratifies experimental units with a pretreatment variable after the experiment is complete, estimates treatment effects within the strata, and then uses a weighted average of these strata estimates for the overall average treatment effect estimate. This is the estimator we focus on.

In this paper, we use the Neyman model to compare post-stratification both to blocking and to using no adjustment. Neyman's framework does not require assumptions of a constant treatment effect or of identically or independently distributed disturbances, assumptions typically made when considering adjustment to experimental data without this framework (e.g., McHugh and Matts, 1983). This avenue for a robust analysis, revitalized in the 1970s (Rubin, 1974), has recently had much appeal. See, for example, work on general experiments (Keele et al., 2009), matched pairs (Imai, 2008), or matched pairs of clusters (Imai et al., 2009).[1] Our estimator is equivalent to one from a fully saturated OLS regression. Freedman (2008a,b) analyzes the regression adjusted estimator under the Neyman model without treatment by strata interactions and finds that the asymptotic variance might be larger than if no correction were made. Lin (2010) extends Freedman's results and shows that when a treatment by covariate interaction is included in the regression, adjustment cannot increase the asymptotic variance. We analyze the exact, finite sample properties of this saturated estimator. Imbens (2011) analyzes estimating the treatment effect of a larger population, assuming the given sample being experimented on is randomly drawn from it. However, because in

[1] See Sekhon (2009) for a historical review of the Neyman model's renaissance.

most randomized trials the sample is not taken at random from the larger population of interest, we focus on estimating the treatment effect of the sample.

We derive the variances for post-stratification and simple difference-in-means estimators under many possible randomization schemes. We show that the difference between the variance of the post-stratified estimator and that of a blocked experiment is on the order of $1/n^2$ with a constant primarily dependent on the proportion of units treated. Post-stratification is quite comparable to blocking. Like blocking, post-stratification can greatly reduce variance over using a simple difference-in-means estimate. However, in small samples post-stratification can substantially hurt precision, especially if the number of strata is large and the stratification variable poorly chosen.

After randomization, researchers can observe the proportion of units actually treated in each strata. We extend our results by deriving variance formula for the post-stratified and simple difference estimators conditioned on these observed proportions. These conditional formula help explain why the variances of the estimators can differ markedly with a prognostic covariate: the difference comes from the potential for bias in the simple-difference estimator when there is large imbalance (i.e., when these proportions are far from what is expected). Interestingly, if the stratification variable is not predictive of outcomes the conditional MSE of the simple-difference estimator usually remains the same or even goes down with greater imbalance, but the conditional MSE of the adjusted estimator increases. Adjusting for a poorly chosen covariate has real cost in finite samples.

The rest of the paper is organized as follows: In the next section, we set up the Neyman model, describe the estimators, and then derive the estimators' variances. We then, in in Section 3, show that post-stratification and blocking have similar characteristics in many circumstances. In Section 4, we present our formula for the estimators' variances conditioned on the observed number of treated units in the strata and discuss their implications. We then align our results with those of Imbens (2011) in Section 5 by extending our findings to the super-population model and discussing the similarities and differences of the two viewpoints. In Section 6, we apply our method to the real data example of a large, randomized medical trial to assess post-stratification's efficacy in a real-world example. We also make a hypothetical example from this data set to illustrate how an imbalanced randomization outcome can induce bias which the post-stratified estimator can adjust for. Section 7 concludes.

2 The Estimators and Their Variances

We consider the Neyman model with two treatments and n units. For example consider a randomized clinical trial with n people, half given a drug and the other half given a placebo. Let $y_i(1) \in \mathbb{R}$ be unit i's outcome if it were treated, and $y_i(0)$ its outcome if it were not. These are the *potential outcomes* of unit i. For each unit, we can only observe either $y_i(1)$ or $y_i(0)$ depending on whether we treat it or not. We make the assumption that treatment assignment for any particular unit has no impact on the potential outcomes of any other unit (this is typically called the stable-unit treatment value assumption or SUTVA). In the drug example this means the decision to give the drug to one patient would have no impact on the outcome of any other patient. The treatment effect t_i for unit i is then the difference in potential outcomes, $t_i \equiv y_i(1) - y_i(0)$, which is deterministic.

Although these t_i are the quantities of interest, we cannot in general estimate them because we cannot observe both potential outcomes of any unit i and because the t_i generally differ by unit. The average across a population of units, however, is estimable. Neyman (Splawa-Neyman et al., 1990) considered the overall Sample Average Treatment Effect, or SATE:

$$\tau \equiv \frac{1}{n}\sum_{i=1}^{n}[y_i(1) - y_i(0)]$$

To conduct an experiment, randomize units into treatment and observe outcomes. Many choices of randomization are possible. The observed outcome is going to be one of the two potential outcomes, and which one depends on the treatment given. Random assignment gives a treatment assignment vector $T = (T_1, \ldots, T_n)$ with $T_i \in \{0, 1\}$ being an indicator variable of whether unit i was treated or not. T's distribution depends on how the randomization was conducted. After the experiment is complete, we obtain the observed outcomes Y, with $Y_i = T_i y_i(1) + (1 - T_i) y_i(0)$. The observed outcomes are random—but only due to the randomization used. The $y_i(\ell)$, and t_i are all fixed. Neyman considered a *balanced complete randomization*:

Definition 2.1 (Complete Randomization of n Units)**.** Given a fixed $p \in (0, 1)$ such that $0 < pn < n$ is an integer, a *Complete Randomization* is a simple random sample of pn units selected for treatment, with the remainder left as controls. If $p = 0.5$ (and n is even) the randomization is *balanced* in that there are the same number of treated units as control units.

The classic unadjusted estimator $\hat{\tau}_{sd}$ is the observed *simple difference* in the means of the treatment and control groups:

$$\begin{aligned}
\hat{\tau}_{sd} &= \frac{1}{W(1)} \sum_{i=1}^{n} T_i Y_i - \frac{1}{W(0)} \sum_{i} (1 - T_i) Y_i \\
&= \sum_{i=1}^{n} \frac{T_i}{W(1)} y_i(1) - \sum_{i=1}^{n} \frac{(1 - T_i)}{W(0)} y_i(0),
\end{aligned}$$

where $W(1) = \sum_i T_i$ is the total number of treated units, $W(0)$ is total control, and $W(1) + W(0) = n$. For Neyman's balanced complete randomization, $W(1) = W(0) = n/2$. For other randomizations schemes the $W(\ell)$ are potentially random.

Neyman showed that the variance of $\hat{\tau}_{sd}$ is (given the potential outcomes y)

$$\text{Var}[\hat{\tau}_{sd}] = \frac{2}{n} \mathbb{E}\left[s_1^2 + s_0^2\right] - \frac{1}{n} S^2 \tag{1}$$

where s_ℓ^2 are the sample variances of the observed outcomes for each group, S^2 is the variance of the n treatment effects t_i, and the expectation is over all possible assignments under balanced complete randomization. We extend this work by considering an estimator that (ideally) exploits some pretreatment covariate b using post-stratification in order to reduce variance.

2.1 Stratification and the Post-Stratified Adjusted Estimator of SATE

Stratification is when an experimenter divides the experimental units into K strata according to some categorical covariate b with $b_i \in \mathcal{B} \equiv \{1, \ldots, K\}$, $i = 1, \ldots, n$. Each stratum k contains $n_k = \#\{j : b_i = k\}$ units. For example, in a cancer drug trial we might have the strata being different stages of cancer. If the strata are associated with outcome, an experimenter can adjust a treatment effect estimate to remove the impact of random variability in the proportion of units treated. This is the idea behind post-stratification. The b_i are observed for all units and are not affected by treatment. The strata defined by the levels of b have stratum-specific SATE$_k$:

$$\tau_k \equiv \frac{1}{n_k} \sum_{i: b_i = k} [y_i(1) - y_i(0)] \qquad k = 1, \ldots, K.$$

The overall SATE can then be expressed as a weighted average of these SATE$_k$s:

$$\tau = \sum_{k \in \mathcal{B}} \frac{n_k}{n} \tau_k. \tag{2}$$

We can view the strata as K mini-experiments. Let $W_k(1) = \sum_{i:b_i=k} T_i$ be the number of treated units in stratum k, and $W_k(0)$ be the number of control units. We can use the simple difference-in-mean estimators for each stratum to estimate the SATE$_k$s:

$$\hat{\tau}_k \quad = \quad \sum_{j:b_i=k} \frac{T_i}{W_k(1)} y_i(1) - \sum_{j:b_i=k} \frac{(1-T_i)}{W_k(0)} y_i(0), \tag{3}$$

A *Post-Stratification Adjusted* estimate is an appropriately weighted estimate of these strata-level estimates:

$$\hat{\tau}_{ps} \quad \equiv \quad \sum_{k \in \mathcal{B}} \frac{n_k}{n} \hat{\tau}_k. \tag{4}$$

These weights echo the weighted sum of SATE$_k$s in Equation 2. Because b and n are known and fixed, the weights are also known and fixed. We derive the variance of $\hat{\tau}_{ps}$ in this paper.

Technically, this estimator is undefined if $W_k(1) = 0$ or $W_k(0) = 0$ for any $k \in 1, \ldots, K$. Similarly, τ_{sd} is undefined if $W(1) = 0$ or $W(0) = 0$. We therefore calculate all means and variances conditioned on $\mathcal{D}$, the event that $\hat{\tau}_{ps}$ is defined, i.e., that each stratum has at least one unit assigned to treatment and one to control. This is fairly natural: if the number of units in each stratum is not too small the probability of $\mathcal{D}$ is close to 1 and the conditioned estimator is similar to an appropriately defined unconditioned estimator. See Section 2.2.

Different experimental designs and randomizations give different distributions on the treatment assignment vector T and all resulting estimators. Some distributions on T would cause bias. We disallow those. Define the *Treatment Assignment Pattern* for stratum k as the ordered vector $(T_i : i \in \{1, \ldots, n : b_i = k\})$. We assume that the randomization used has *Assignment Symmetry*:

Definition 2.2 (Assignment Symmetry)**.** A randomization is *Assignment Symmetric* if the following two properties hold:

1. *Equiprobable Treatment Assignment Patterns*
   All $\binom{n_k}{W_k(1)}$ ways to treat $W_k(1)$ units in stratum k are equiprobable, given $W_k(1)$.
2. *Independent Treatment Assignment Patterns*
   For all strata j, k, with $j \neq k$, the treatment assignment pattern in stratum j is independent of the treatment assignment pattern in stratum k, given $W_j(1)$ and $W_k(1)$.

Complete randomization and Bernoulli assignment (where independent p-coin flips determine treatment for each unit) satisfy assignment symmetry. So does blocking, where strata are randomized independently. Furthermore, given a distribution on T that satisfies Assignment Symmetry (Definition 2.2), conditioning on $\mathcal{D}$ also maintains Assignment Symmetry (as do many other reasonable conditionings, such as having at least x units in both treatment and control, and so on). See the supplementary material for a more formal argument. In our technical results, we assume that (1) the randomization is Assignment Symmetric and (2) we are conditioning on $\mathcal{D}$, the set of possible assignments where $\hat{\tau}_{ps}$ is defined.

The post-stratification adjusted estimator and the simple-difference estimator are used when the initial random assignment ignores the stratification variable b. In a blocked experiment, the estimator used is $\hat{\tau}_{ps}$, but the randomization is done within the strata defined by b. All three of these options are unbiased. We are interested in their relative variances. We express the variances of these estimators with respect to the population's (unknown) means, variances and covariances of potential outcomes divided into between-strata variation and within-stratum variation. The within-stratum variances and covariances are:

$$\sigma_k^2(\ell) = \frac{1}{n_k - 1} \sum_{i:b_i=k} \left[y_i(\ell) - \bar{y}_k(\ell)\right]^2 \qquad \ell = 0, 1$$

and

$$\gamma_k(1,0) = \frac{1}{n_k - 1} \sum_{i:b_i=k} \left[y_i(1) - \bar{y}_k(1)\right]\left[y_i(0) - \bar{y}_k(0)\right],$$

where $\bar{y}_k(\ell)$ denotes the mean of $y_i(\ell)$ for all units in stratum k. Like many authors, we use $n_k - 1$ rather than n_k for convenience and cleaner formula. The $(1,0)$ in $\gamma_k(1,0)$ indicates that this framework could be extended to multiple treatments. The population-wide $\sigma^2(\ell)$ and $\gamma(1,0)$ are analogously defined. The population-wide $\sigma^2(\ell)$ and $\gamma(1,0)$ are weighted sums of the component pieces. The between-stratum variance and covariance are weighted variances and covariances of the strata means:

$$\bar{\sigma}^2(\ell) = \frac{1}{n-1} \sum_{k=1}^{K} n_k \left[\bar{y}_k(\ell) - \bar{y}(\ell)\right]^2 \qquad \ell = 0, 1$$

and

$$\bar{\gamma}(1,0) = \frac{1}{n-1} \sum_{k=1}^{K} n_k \left[\bar{y}_k(1) - \bar{y}(1)\right]\left[\bar{y}_k(0) - \bar{y}(0)\right].$$

We also refer to the *correlation of potential outcomes* r, where $r \equiv \gamma(1,0)/\sigma(0)\sigma(1)$ and the strata-level correlations, $r_k \equiv \gamma_k(1,0)/\sigma_k(0)\sigma_k(1)$. An overall constant treatment effect gives $r = 1$, $\sigma(0) = \sigma(1)$, $r_k = 1$ for all k and $\sigma_k(0) = \sigma_k(1)$ for all k.

We are ready to state our main results:

Theorem 2.1. *The strata-level estimators $\hat{\tau}_k$ are unbiased, i.e.*

$$\mathbb{E}[\hat{\tau}_k] = \tau_k \qquad k = 1, \ldots, K$$

and their variances are

$$Var[\hat{\tau}_k] \;=\; \frac{1}{n_k}\left[\beta_{1k}\sigma_k^2(1) + \beta_{0k}\sigma_k^2(0) + 2\gamma_k(1,0)\right] \tag{5}$$

with $\beta_{1k} = \mathbb{E}[W_k(0)/W_k(1)|\mathcal{D}]$, the expected ratio of units in control to units treated in stratum k, and $\beta_{0k} = \mathbb{E}[W_k(1)/W_k(0)|\mathcal{D}]$, the reverse.

See Appendix A for a proof. The $\hat{\tau}_k$s are independent due to Assignment Symmetry, so the mean and variance of $\hat{\tau}_{ps}$ are weighted sums of the strata means and variances. Thus we have

Theorem 2.2. *The post-stratification adjusted $\hat{\tau}_{ps}$ is unbiased:*

$$\mathbb{E}[\hat{\tau}_{ps}|\mathcal{D}] = \mathbb{E}\left[\sum_k \frac{n_k}{n}\hat{\tau}_k\right] = \sum_k \frac{n_k}{n}\,\mathbb{E}[\hat{\tau}_k] = \sum_k \frac{n_k}{n}\tau_k = \tau.$$

Its variance is

$$Var[\hat{\tau}_{ps}|\mathcal{D}] \;=\; \frac{1}{n}\sum_k \frac{n_k}{n}\left[\beta_{1k}\sigma_k^2(1) + \beta_{0k}\sigma_k^2(0) + 2\gamma_k(1,0)\right]. \tag{6}$$

Corollary 2.3. *The unadjusted simple-difference estimator $\hat{\tau}_{sd}$ is unbiased, i.e.* $\mathbb{E}[\hat{\tau}_{sd}] = \tau$. *Its variance is*

$$Var[\hat{\tau}_{sd}|\mathcal{D}] \quad = \quad \frac{1}{n}\left[\beta_1\sigma^2(1) + \beta_0\sigma^2(0) + 2\gamma(1,0)\right], \tag{7}$$

where $\beta_1 \equiv \mathbb{E}[W(0)/W(1)|\mathcal{D}]$ *and* $\beta_0 \equiv \mathbb{E}[W(1)/W(0)|\mathcal{D}]$. *In terms of strata-level variances, its variance is*

$$\begin{aligned} Var[\hat{\tau}_{sd}|\mathcal{D}] \quad = \quad & \frac{1}{n}\left[\beta_1\bar{\sigma}^2(1) + \beta_0\bar{\sigma}^2(0) + 2\bar{\gamma}(1,0)\right] + \\ & \frac{1}{n}\sum_k \frac{n_k - 1}{n-1}\left[\beta_1\sigma_k^2(1) + \beta_0\sigma_k^2(0) + 2\gamma_k(1,0)\right]. \end{aligned} \tag{8}$$

For completely randomized experiments with np units treated, $\beta_1 = (1-p)/p$ and $\beta_0 = p/(1-p)$. For a balanced completely randomized experiment, Equation 7 is the result presented in Splawa-Neyman et al. (1990)—see Equation 1; the expectation of the sample variance is the overall variance. Then $\beta_\ell = 1$ and

$$\begin{aligned} \text{Var}[\hat{\tau}_{sd}] &= \frac{2}{n}\left(\sigma^2(1) + \sigma^2(0)\right) - \frac{1}{n}\text{Var}[y_i(1) - y_i(0)] \\ &= \frac{2}{n}\left(\sigma^2(1) + \sigma^2(0)\right) - \frac{1}{n}\left(\sigma^2(1) + \sigma^2(0) - 2\gamma(1,0)\right) \\ &= \frac{1}{n}\left(\sigma^2(1) + \sigma^2(0) + 2\gamma(1,0)\right). \end{aligned}$$

Remarks. β_{1k} is the expectation of $W_k(0)/W_k(1)$, the ratio of control units to treated units in stratum k. For large n, this ratio is close to the ratio $\mathbb{E}[W_k(0)] \,/\, \mathbb{E}[W_k(1)]$ since the $W_k(\ell)$ do not vary much relative to their size. For small n, however, they do vary more, which tends to result in the β_{1k} being noticeably larger than $\mathbb{E}[W_k(0)] \,/\, \mathbb{E}[W_k(1)]$. This is at root of how the overall variance of post-stratification differs from blocking. This is discussed more formally later on, and in Appendix B.

For $\ell = 0, 1$ the $\beta_{\ell k}$'s are usually larger than β_ℓ, being expectations of different variables with different distributions. For example in a balanced completely randomized experiment $\beta_1 = 1$ but $\beta_{1k} > 1$ for $k = 1, \ldots, K$ since $W_k(1)$ is random.

All the β's depend on both the randomization and the conditioning on $\mathcal{D}$, and thus the variances from both Equation 8 and Equation 6 can change (markedly) under different randomization scenarios. As a simple illustration, consider a complete randomization of a 40 unit sample with a constant treatment effect and four strata of equal size. Let all the $\sigma_k(\ell) = 1$ and all $r_k = 1$. If $p = 0.5$, then $\beta_1 = \beta_0 = 1$ and the variance is about 0.15. If $p = 2/3$ then $\beta_1 = 1/2$ and $\beta_0 = 2$. Equation 8 holds in both cases, but the variance in the second case will be about 10% larger due to the larger β_0. There are fewer control units, so the estimate of the control outcome is more uncertain. The gain in certainty for the treatment units does not compensate enough. For $p = 0.5$, $\beta_{1k} = \beta_{0k} \approx 1.21$. The post-stratified variance is about 0.11. For $p = 2/3$, $\beta_{1k} \approx 2.44$ and $\beta_{0k} \approx 0.61$. The average is about 1.52. The variance is about 14% larger than the $p = 0.5$ case. Generally speaking, the relative variances of different experimental setups are represented in the β's.

Comparing the Estimators. Both $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ are unbiased, so their MSEs are the same as their variances. To compare $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ take the difference of their variances:

$$\begin{aligned}\text{Var}[\hat{\tau}_{sd}] - \text{Var}[\hat{\tau}_{ps}] = &\left\{ \frac{1}{n} \left(\beta_1 \bar{\sigma}^2(1) + \beta_0 \bar{\sigma}^2(0) + 2\bar{\gamma}(1,0) \right) \right\} - \\ &\left\{ \frac{1}{n} \sum_{k=1}^{K} \left[\left(\frac{n_k}{n} \beta_{1k} - \frac{n_k - 1}{n-1} \beta_1 \right) \sigma_k^2(1) + \left(\frac{n_k}{n} \beta_{0k} - \frac{n_k - 1}{n-1} \beta_0 \right) \sigma_k^2(0) \right] + \right. \\ &\left. \frac{2}{n^2} \sum_{k=1}^{K} \frac{n - n_k}{n-1} \gamma_k(1,0) \right\}. \end{aligned} \tag{9}$$

Equation 9 breaks down into two parts as indicated by the curly brackets. The first part, $\beta_1 \bar{\sigma}^2(1) + \beta_0 \bar{\sigma}^2(0) + 2\bar{\gamma}(1,0)$, is the between-strata variation. It measures how much the mean potential outcomes vary across strata and captures how well the stratification variable separates out different units, on average. The larger the separation, the more to gain by post-stratification. The second part, consisting of the bottom two lines of Equation 9, represents the cost paid by post-stratification due to, primarily, the chance of random imbalance in treatment. This second part is non-positive and is a penalty except in some cases where the proportion of units treated is extremely close to 0 or 1 or is radically different across strata.

If the between-strata variation is larger than the cost paid then Equation 9 is positive and it is good to post-stratify. If Equation 9 is negative then it is bad to post-stratify. It can be positive or negative depending on the parameters of the population. In particular, if there is no between-strata difference in the mean potential outcomes, then the terms on the first line of Equation 9 are 0, and post-stratification hurts. Post-stratification is not necessarily a good idea when compared to doing no adjustment at all.

To assess the magnitude of the penalty paid compared to the gain, multiply Equation 9 by n. The first term, representing the between-strata variation, is now a constant, and the scaled gain converges to it as n grows:

Theorem 2.4. *Take an experiment with n units randomized under either complete randomization or Bernoulli assignment. Let p be the expected proportion of units treated. Without loss of generality, assume $p \geq 0.5$. Let f be the ratio of the smallest stratum to n. Let $\sigma^2_{max} = \max_{k,\ell} \sigma_k^2(\ell)$ be the largest variance of all the strata. Then*

$$\left| n\left(Var[\hat{\tau}_{sd}] - Var[\hat{\tau}_{ps}]\right) - \beta_1 \bar{\sigma}^2(1) - \beta_0 \bar{\sigma}^2(0) - 2\bar{\gamma}(1,0) \right| \leq \left(\frac{8}{f(1-p)^2} + \frac{2p}{1-p} \right) \sigma^2_{max} \frac{1}{n} + O(\frac{1}{n^2}).$$

See Appendix B for the derivation. Theorem 2.4 shows us that the second part of Equation 9, the harm, diminishes quickly.

If the number of strata K grows with n, the story can change. The second and third lines of Equation 9 are sums over K elements. The larger the number of strata K, the more terms in the sums and the greater the potential penalty for stratification, unless the $\sigma_k^2(\ell)$'s shrink in proportion as K grows. For an unrelated covariate, they will not tend to do so. To illustrate, we made a sequence of experiments increasing in size with a continuous covariate z unrelated to outcome. For each experiment with n units, we made b by cutting up our continuous z into $K = n/10$ chunks. Post-stratification was about 15% worse, in this case, than the simple-difference estimator regardless of n.

Overall, post-stratifying on variables not heavily related to outcome is unlikely to be worthwhile and can be harmful. Post-stratifying on variables that do relate to outcome will likely result in large between-strata variation and thus a large reduction in variance as compared to a simple-difference estimator. More strata is not necessarily better, however. Simulations suggest that there is often a law of diminishing returns. For example, we made a simulated experiment with $n = 200$ units with a continuous covariate z related to outcome. We then made b by cutting z up into K chunks for $K = 1, \ldots, 20$. As K increased from 1 there was a sharp drop in variance and then, as the cost due to post-stratification increased, the variance leveled off and then climbed. In this case, $K = 5$ was ideal. We did a similar simulation for a covariate z unrelated to outcome. Now, regardless of K, the $\sigma_k^2(\ell)$ were all about the same and the between-strata variation was fairly low. As K grew, the overall variance climbed. In many cases a few moderate-sized strata give a dramatic reduction in variance, but having more strata beyond that has little impact, and can even lead to an increase in $\hat{\tau}_{ps}$'s variance.

Estimation. Equation 6 and Equation 8 are the actual variances of the estimators. In practice, the variance of an estimator, i.e., the standard error, would have to itself be estimated. Unfortunately, however, it is usually not possible to consistently estimate the standard errors of difference-in-means estimators due to so-called identifiability issues as these standard errors depend on $\gamma(1, 0)$, the typically un-estimable correlation of the potential outcomes of the units being experimented on (see Splawa-Neyman et al. (1990)). One approach to consistently estimate these standard errors is to impose structure to render this correlation estimable or known; Reichardt and Gollob (1999), for example, demonstrate that quite strong assumptions have to be made to obtain an unbiased estimator for the variance of $\hat{\tau}_{sd}$. It is straightforward, however, to make a non-trivial conservative estimate of this variance by assuming the correlation is maximal. Sometimes there can be nice tricks—see, for example, Abadie and Imbens (2007), who estimate these parameters for matched-pairs by looking at pairs of pairs matched on covariates—but generally bounding the standard error is the best one can do.

This paper compares the actual variances of the estimators. Estimating these variances is an area for future work, involving these identifiability issues and degrees-of-freedom issues as well. It is quite possible that, in small samples, the increased uncertainty in estimating the many variances composing the standard error of the post-stratification estimator would overwhelm any potential gains.

That being said, all terms except the $\gamma_k(1, 0)$ in Equation 9 are estimable with standard sample variance, covariance, and mean formula. In particular, $\bar{\gamma}(1, 0)$ is estimable. By then making the conservative assumption that the $\gamma_k(1, 0)$ are maximal (i.e., that $r_k = 1$ for all k so $\gamma_k(1, 0) = \sigma(1)\sigma(0)$), we can estimate a lower-bound on the gain. Furthermore, by then dividing by a similar upper bound on the standard error of the simple-difference estimator, we can give a lower-bound on the percentage reduction in variance due to post-stratification. We illustrate this when we analyze an experiment in Section 6.

2.2 Not Conditioning on $\mathcal{D}$ Changes Little

Our results are conditioned on $\mathcal{D}$, the set of assignments such that $W_k(\ell) \neq 0$ for all $k = 1, \ldots K$ and $\ell = 0, 1$. This, it turns out, results in variances only slightly different from not conditioning on $\mathcal{D}$.

Define the estimator $\hat{\tau}_{ps}$ so that $\hat{\tau}_{ps} = 0$ if $\neg\mathcal{D}$ occurs, i.e. $W_k(\ell) = 0$ for some k, ℓ. Other choices of how to define the estimator when $\neg\mathcal{D}$ occurs are possible, including letting $\hat{\tau}_{ps} = \hat{\tau}_{sd}$—the point is that this choice does not much matter. In this case $\mathbb{E}[\hat{\tau}_{ps}] = \tau \mathbf{P}\mathcal{D}$. The estimate of the treatment is shrunk by $\mathbf{P}\mathcal{D}$ towards 0. It is biased by $\tau\mathbf{P}\neg\mathcal{D}$. The variance is

$$\text{Var}[\hat{\tau}_{ps}] = \text{Var}[\hat{\tau}_{ps}|\mathcal{D}]\,\mathbf{P}\mathcal{D} + \tau^2\mathbf{P}\neg\mathcal{D}\mathbf{P}\mathcal{D}$$

and the MSE is

$$MSE\left[\hat{\tau}_{ps}\right] = \mathbb{E}\left[(\hat{\tau}_{ps} - \tau)^2\right] = \text{Var}[\hat{\tau}_{ps}|\mathcal{D}]\,\mathbf{P}\mathcal{D} + \tau^2\mathbf{P}\neg\mathcal{D}.$$

Not conditioning on $\mathcal{D}$ introduces a bias term and some extra variance terms. All these terms are small if $\mathbf{P}\mathcal{D}$ is near 1, which it is: $1-\mathbf{P}\mathcal{D}$ is $O(ne^{-n})$ (see Appendix B for proof). Not conditioning on $\mathcal{D}$, then, gives substantively the same conclusions as conditioning on $\mathcal{D}$, but the formula are a bit more unwieldy. Conditioning on the set of randomizations where $\hat{\tau}_{ps}$ is defined is more natural.

3 Comparing Blocking to Post-Stratification

Let the *assignment split* W of a random assignment be the number of treated units in the strata:

$$W \equiv (W_1(1), \ldots, W_K(1))$$

A *randomized block trial* ensures that W is constant because we randomize within strata, ensuring a pre-specified number of units are treated in each. This randomization is Assignment Symmetric (Def 2.2) and, further, the probability of being defined, $\mathcal{D}$, is 1. For blocking, the standard estimate of the treatment effect has the same expression as $\hat{\tau}_{ps}$, but the $W_k(\ell)$s are all fixed. If all blocks have the same proportion treated (i.e., $W_k(1)/n_k = W(1)/n$ for all k), this coincides with $\hat{\tau}_{sd}$.

Because W is constant

$$\beta_{1k} = \mathbb{E}\left[\frac{W_k(0)}{W_k(1)}\right] = \frac{W_k(0)}{W_k(1)} = \frac{1-p_k}{p_k}, \tag{10}$$

where p_k is the proportion of units assigned to treatment in stratum k. Similarly, $\beta_{0k} = p_k/(1-p_k)$. Letting the subscript "blk" denote this randomization, plug Equation 10 into Equation 6 to get the variance of a blocked experiment:

$$\text{Var}_{blk}\left[\hat{\tau}_{ps}\right] = \frac{1}{n}\sum_k \frac{n_k}{n}\left(\frac{1-p_k}{p_k}\sigma_k^2(1) + \frac{p_k}{1-p_k}\sigma_k^2(0) + 2\gamma_k(1,0)\right). \tag{11}$$

Post-stratification is similar to blocking, and the post-stratified estimator's variance tends to be close to that of a blocked experiment. Taking the difference between Equation 6 and Equation 11 gives

$$\text{Var}[\hat{\tau}_{ps}|\mathcal{D}] - \text{Var}_{blk}\left[\hat{\tau}_{ps}\right] = \frac{1}{n}\sum_k \frac{n_k}{n}\left[\left(\beta_{1k} - \frac{1-p_k}{p_k}\right)\sigma_k^2(1) + \left(\beta_{0k} - \frac{p_k}{1-p_k}\right)\sigma_k^2(0)\right]. \tag{12}$$

The $\gamma_k(1,0)$ cancelled; Equation 12 is identifiable and therefore estimable.

Randomization without regard to b can have block imbalance due to ill luck: W is random. The resulting cost in variance of post-stratification over blocking is represented by the $\beta_{1k} - (1-p_k)/p_k$ terms in Equation 12. This cost is small, as shown by Theorem 3.1:

Theorem 3.1. *Take a post-stratified estimator for a completely randomized or Bernoulli assigned experiment. Use the assumptions and definitions of Theorem 2.4. Assume the common case for blocking of $p_k = p$ for $k = 1, \ldots, K$. Then*

$$n\Big(Var[\hat{\tau}_{ps}|\mathcal{D}] - Var_{blk}\left[\hat{\tau}_{ps}\right] \Big) \leq \frac{8}{(1-p)^2} \frac{1}{f} \sigma^2_{max} \frac{1}{n} + O(e^{-fn}).$$

See Appendix B for the derivation.

Theorem 3.1 bounds how much worse post-stratification can be to blocking. The scaled difference is on the order of $1/n$. The differences in variance are order $1/n^2$. 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 increased uncertainty in the stratum-level estimates. Many strata are generally not helpful and can be harmful if b is not prognostic.

A note on blocking. Plug Equation 10 into the gain equation (Equation 9) to immediately see under what circumstances blocking has a larger variance than the simple difference estimator for a completely randomized experiment:

$$\begin{aligned} \text{Var}[\hat{\tau}_{sd}] - \text{Var}_{blk}\left[\hat{\tau}_{ps}\right] \;\; &= \;\; \frac{1}{n}\left(\frac{1-p}{p}\bar{\sigma}^2(1) + \frac{p}{1-p}\bar{\sigma}^2(0) + 2\bar{\gamma}(1,0) \right) - \\ & \quad\; \frac{1}{n^2}\sum_k \frac{n-n_k}{n-1}\left(\frac{1-p}{p}\sigma_k^2(1) + \frac{p}{1-p}\sigma_k^2(0) + 2\gamma_k(1,0) \right). \end{aligned} \tag{13}$$

If $p = 0.5$, this is identical to the results in the appendix of Imai et al. (2008). In the worst case where there is no between-strata variation, the first term of Equation 13 is 0 and so the overall difference is $O(K/n^2)$. The penalty for blocking is small, even for moderate-sized experiments, assuming the number of strata does not grow with n. If the first term is not zero, then it will dominate for large enough n; i.e. blocking will give a more precise estimate. For more general randomizations, Equation 9 still holds but the β's differ. The difference in variances is still $O(1/n^2)$.

4 Conditioning on the Assignment Split W

By conditioning on the assignment split W we can break down the expressions for variance to better understand when $\hat{\tau}_{ps}$ outperforms $\hat{\tau}_{sd}$. For $\hat{\tau}_{**}$ with $** = ps$ or sd we have

$$\text{Var}[\hat{\tau}_{**}] = \text{MSE}[\hat{\tau}_{**}] = \mathbb{E}_W[\text{MSE}[\hat{\tau}_{**}|W]] = \sum_{w\in\mathcal{W}} \text{MSE}[\hat{\tau}_{**}|W = w]\,\mathbf{P}\{W = w\}$$

with $\mathcal{W}$ being the set of all allowed splits where $\hat{\tau}_{ps}$ is defined. The overall MSE is a weighted average of the conditional MSE, with the weights being the probability of the given possible splits W. This will give us insight into when $\text{Var}[\hat{\tau}_{sd}]$ is large.

Conditioning on the split W maintains Assignment Symmetry and sets $\beta_{\ell k} = W_k(1-\ell)/W_k(\ell)$ for $k \in 1, \ldots, K$ and $\beta_\ell = W(1-\ell)/W(\ell)$. For $\hat{\tau}_{ps}$ we immediately obtain

$$\text{Var}[\tau_{ps}|W] = \frac{1}{n}\sum_k \frac{n_k}{n}\left(\frac{W_k(0)}{W_k(1)}\sigma_k^2(1) + \frac{W_k(1)}{W_k(0)}\sigma_k^2(0) + 2\gamma_k(1,0) \right). \tag{14}$$

Under conditioning $\hat{\tau}_{ps}$ is still unbiased and so the conditional MSE is the conditional variance. $\hat{\tau}_{sd}$, however, can now be *biased* with a conditional MSE larger than the conditional variance if the extra bias term is nonzero. Theorem 4.1 show the bias and conditional variance of $\hat{\tau}_{sd}$:

Theorem 4.1. *The bias of $\hat{\tau}_{sd}$ conditioned on W is*

$$\mathbb{E}[\hat{\tau}_{sd}|W] - \tau = \sum_{k\in\mathcal{B}} \left[\left(\frac{W_k(1)}{W(1)} - \frac{n_k}{n} \right) \bar{y}_k(1) - \left(\frac{W_k(0)}{W(0)} - \frac{n_k}{n} \right) \bar{y}_k(0) \right],$$

which is not 0 in general, even with a constant treatment effect. $\hat{\tau}_{sd}$'s variance conditioned on W is

$$Var[\hat{\tau}_{sd}|W] = \sum_{k\in\mathcal{B}} \frac{W_{1k}W_{0k}}{n_k} \left(\frac{1}{W_1^2}\sigma_k^2(1) + \frac{1}{W_0^2}\sigma_k^2(0) - \frac{2}{W_1 W_0}\gamma_k(1,0) \right).$$

See Appendix A for a sketch of these two derivations. They come from an argument similar to the proof for the variance of $\hat{\tau}_{ps}$, but with additional weighting terms.

The conditional MSE of $\hat{\tau}_{sd}$ has no nice formula that we are aware, and is simply the sum of the variance and the squared bias:

$$MSE[\hat{\tau}_{sd}|W] = \text{Var}[\hat{\tau}_{sd}|W] + (\mathbb{E}[\hat{\tau}_{sd}|W] - \tau)^2 \tag{15}$$

In a typical blocked experiment, W would be fixed at W^{blk} where $W_k^{blk} = n_k p$ for $k = 1, \ldots, K$. For complete randomization, $\mathbb{E}[W] = W^{blk}$. We can now gain insight into the difference between the simple-difference and post-stratified estimators. If W equals W^{blk}, then the conditional variance formula for both estimators reduce to that of blocking, i.e., Equation 14 and Equation 15 reduce to Equation 11. But the more W deviates from W^{blk}—i.e., the more *imbalanced* the assignment is—the larger the post-stratified variance formula will tend to be. To see this note how for each strata the overall variance is a weighted sum of $W_k(0)/W_k(1)$ and $W_k(1)/W_k(0)$. The more unbalanced these terms, the larger the sum. The simple-difference estimator, on the other hand, tends to have smaller variance as W deviates further from W^{blk} due to the greater restrictions on the potential random assignments.

However, if b is prognostic then, generally, the bias of the simple-difference estimator increases with greater imbalance. This bias can radically inflate the MSE. But if b is not prognostic then there is little or no bias. Overall, then, as imbalance increases, the variance (and MSE) of $\hat{\tau}_{ps}$, on the other hand, moderately increases. The variance of $\hat{\tau}_{sd}$ moderately decreases but the bias increases, giving a MSE that can grow quite large.

Because the overall MSE of these estimators is a weighted average of the conditional MSEs, and because under perfect balance the conditional MSEs are the same, we know that the variance of $\hat{\tau}_{sd}$ being larger than $\hat{\tau}_{ps}$ comes from the impact of potentially having bad imbalance with resulting large bias.

The split W is directly observable and gives hints to the experimenter as to the success, or failure, of the randomization. Unbalanced splits tell us we have less certainty while balanced splits are comforting. For example, take a hypothetical balanced completely randomized experiment with $n = 32$ subjects, half men and half women. Consider the case where only one man ends up in treatment as compared to 8 men. In the former case, a single man gives the entire estimate for average treatment outcome for men and a single woman gives the entire estimate for average control outcome for women. This seems *very* unreliable. In the latter case, each of the four mean outcomes are estimated with 8 subjects, which seems more reliable. Our estimates of uncertainty should take this observed split into account. We take the observed W into account by using the conditional MSE rather than overall MSE when estimating uncertainty. The conditional MSE estimates how close ones actual experimental estimate is likely to be from the SATE. The overall MSE estimates how close such estimates will generally be to the SATE. The idea of using all observed information is not new. When sampling to find the mean of a population, Holt and Smith (1979) argue that, for estimators adjusted using post-stratification, variance estimates should be conditioned on the distribution of units in the strata as this gives a more relevant estimate of uncertainty. Pocock et al. (2002) extends Senn (1989) and examines conditioning on the imbalance of a continuous covariate in ANCOVA. They show that not correcting for imbalance (as measured as a standardized difference in means) gives one inconsistent control on the error rate when testing for an overall treatment effect.

5 Extension to an Infinite-Population Model

The presented results apply to estimating the treatment effect for a specific sample of units, but there is often a larger population of interest. One approach is to consider the sample to be a random draw from this larger population, which introduces an additional component of randomness capturing how the SATE varies about the Population Average Treatment Effect, or PATE. For example, see Imbens (2011). But if the sample has not been so drawn, using this PATE model might not be appropriate. The SATE perspective should instead be used, with additional work to then generalize the results. See Hartman et al. (2011); Imai et al. (2008). Regardless, under the PATE approach, the variances of all the estimators increase, but the substance of this paper's findings remain.

Let f_k, $k = 1, \ldots, K$, be the proportion of the population in stratum k. The PATE can then be broken down by strata:

$$\tau^* = \sum_{k=1}^{K} f_k \tau_k^*$$

with τ_k^* being the population average treatment effect in stratum k. Let the sample $\mathcal{S}$ be a stratified draw from this population holding the proportion of units in the sample to f_k (i.e. $n_k/n = f_k$ for $k = 1, \ldots, K$). (See below for different types of draws from the population.) τ, the SATE, is random, depending on $\mathcal{S}$. Due to the size of the population, the sampling is close to being with replacement. Alternatively, the sample could be generated by independent draws from a collection of K distributions, one for each stratum. Let $\sigma_k^2(\ell)^*$, $\gamma_k^2(1,0)^*$, etc., be population parameters. Then the PATE-level MSE of $\hat{\tau}_{ps}$ is

$$\text{Var}[\hat{\tau}_{ps}] = \frac{1}{n} \sum_k f_k \left[(\beta_{1k} + 1)\sigma_k^2(1)^* + (\beta_{0k} + 1)\sigma_k^2(0)^* \right]. \tag{16}$$

See Appendix A for the derivation. Imbens (2011) has a similar formula for the two-strata case. Compare to Equation 6: All the correlation of potential outcomes terms $\gamma_k(1,0)$ vanish when moving to PATE. This is due to a perfect trade-off: the more they are correlated, the harder to estimate the SATE τ for the sample, but the easier it is to draw a sample with a SATE τ close to the overall PATE τ^*.

The simple-difference estimator. For the simple-difference estimator, use Equation 16 with $K = 1$ to get

$$\text{Var}[\hat{\tau}_{sd}] = \frac{1}{n} \left[(\beta_1 + 1)\sigma^2(1)^* + (\beta_0 + 1)\sigma^2(0)^* \right]. \tag{17}$$

Now let $\bar{\sigma}^2(\ell)^*$ be a weighted sum of the squared differences of the strata means to the overall mean:

$$\bar{\sigma}^2(\ell)^* = \sum_{k=1}^{K} f_k \left(\bar{y}_k^*(\ell) - \bar{y}^*(\ell) \right)^2.$$

The population variances then decompose into $\bar{\sigma}^2(\ell)^*$ and strata-level terms:

$$\sigma^2(\ell)^* = \bar{\sigma}^2(\ell)^* + \sum_{k=1}^{K} f_k \sigma_k^2(\ell)^*.$$

Plug this decomposition into Equation 17 to get

$$\text{Var}[\hat{\tau}_{sd}] = \frac{1}{n}\left[(\beta_1+1)\left(\bar{\sigma}^2(1)^* + \sum_{k=1}^{K} f_k\sigma_k^2(1)^*\right) + (\beta_0+1)\left(\bar{\sigma}^2(0)^* + \sum_{k=1}^{K} f_k\sigma_k^2(0)^*\right)\right]$$

Variance gain from post-stratification. For comparing the simple-difference to the post-stratified estimator at the PATE level, take the difference of Equation 17 and Equation 16 to get

$$\begin{aligned}\text{Var}[\hat{\tau}_{sd}] - \text{Var}[\hat{\tau}_{ps}] = &\frac{1}{n}(\beta_1+1)\bar{\sigma}^2(1)^* + \frac{1}{n}(\beta_0+1)\bar{\sigma}^2(0)^* \\ &- \frac{1}{n}\sum_{k=1}^{K} f_k\left[(\beta_{1k}-\beta_1)\sigma_k^2(1)^* + (\beta_{0k}-\beta_0)\sigma_k^2(0)^*\right].\end{aligned}$$

Similar to the SATE view, we again have a gain component (the first line) and a cost (the second line). For Binomial assignment and complete randomization, $\beta_\ell \le \beta_{\ell k}$ for all k, making the cost nonnegative. There are no longer terms for the correlation of potential outcomes, and therefore this gain formula is directly estimable. The cost is generally smaller than for the SATE model due to the missing $\gamma_k(1,0)$ terms.

The variance of blocking under PATE. For equal-proportion blocking, $W_k(1) = pn_k$ and $W_k(0) = (1-p)n_k$. Using this and $\beta_{\ell k}+1 = \mathbb{E}[n_k/W_k(\ell)]$, the PATE-level MSE for a blocked experiment is then

$$\text{Var}[\hat{\tau}_{ps}] = \frac{1}{n}\sum_k \frac{n_k}{n}\left[\frac{1}{p}\sigma_k^2(1)^* + \frac{1}{1-p}\sigma_k^2(0)^*\right]$$

For comparing complete randomization (with pn units assigned to treatment) to blocked experiments, plug in the β's. The $\beta_\ell - \beta_{\ell k}$ terms all cancel, leaving

$$\text{Var}[\hat{\tau}_{sd}] - \text{Var}[\hat{\tau}_{ps}] = \frac{1}{n}\frac{1}{p}\bar{\sigma}^2(1)^* + \frac{1}{n}\frac{1}{1-p}\bar{\sigma}^2(0)^* \ge 0$$

Unlike from the SATE perspective, blocking can never hurt from the PATE perspective.

Not conditioning on the n_k. Allowing the n_k to vary introduces some complexity, but the gain formula remain unchanged. If the population proportions are known, but the sample is a completely random draw from the population, the natural post-stratified estimate of the PATE would use the population weights f_k. These weights can be carried through and no problems result. Another approach is to estimate the f_k with n_k/n in the sample. In this latter case, we first condition on the seen vector $N \equiv n_1, \ldots n_k$ and define a τ^N based on N. Conditioned on N, both $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ are unbiased for estimating τ^N, and we can use the above formula with n_k/n instead of f_k. Now use the tower-property of expectations and variances. This results in an extra variance of a multinomial to capture how τ^N varies about τ as N varies. The variances of both the estimators will each be inflated by this extra term, which therefore cancels when looking at the difference.

6 PAC Data Illustration

We apply our methods to evaluating Pulmonary Artery Catheterization (PAC), an invasive and controversial cardiac monitoring device, that was, until recently, widely used in the management of critically ill patients (Dalen, 2001; Finfer and Delaney, 2006). Controversy arose regarding the use of PAC when a non-random study using propensity score matching found that PAC insertion for critically ill patients was associated with increased costs and mortality (Connors et al., 1996). Other observational studies came to similar conclusions leading to reduced PAC use (Chittock et al., 2004). However, an RCT (PAC-Man) found no difference in mortality between PAC and no-PAC groups (Harvey et al., 2005), which substantiated the concern that the observational results were subject to selection bias (Sakr et al., 2005).

PAC-Man has 1013 subjects, half treated. The outcome variable investigated here is "qualys" or quality-adjusted life years. Higher values indicate, generally, longer life and higher quality of life. Death at time of PAC insertion or shortly after receives a value of 0. Living two years in full health would be a 2. There is a lot of fluctuation in these data. There is a large point mass at 0 and a long tail up to 20 years.

Unfortunately, the RCT itself had observed covariate imbalance in predicted probability of death, a powerful predictor of the outcome, which calls into question the reliability of the simple-difference estimate of the treatment effect. More low-risk patients were assigned to receive treatment, which could induce a perceived treatment effect even if none were present. Post-stratification could help with this potential bias and decrease the variance of the estimate of treatment effect. To estimate the treatment effect using post-stratification we first divide the continuous probability of death covariate into K K-tiles. We then estimate the treatment effect within the resulting strata and average appropriately.

This analysis is simplified for the purposes of illustration. We are only looking at one of the outcomes and have dropped several potentially important covariates for the sake of clarity. Statistics on the strata for $K = 4$ are listed on Table 1. A higher proportion of subjects in the first two groups were treated than one would expect given the randomization. Imbalance in the first group, with its high average outcome, could heavily influence the overall treatment effect estimate of $\hat{\tau}_{sd}$.

Strata	# Tx	# Co	$SD_k(1)$	$SD_k(0)$	$\hat{y}_k(1)$	$\hat{y}_k(0)$	$\hat{\tau}_k$
Low Risk	136	118	5.80	5.68	5.57	5.41	0.15
Moderate Risk	142	111	3.42	4.17	1.69	2.70	-1.01
High Risk	106	147	3.60	3.75	1.97	2.36	-0.39
Extreme Risk	122	131	3.41	3.10	1.37	1.19	0.18
Overall	506	507	4.56	4.48	2.72	2.84	-0.13

Table 1: Strata-Level Statistics for the PAC Illustration

We estimate the minimum gain in precision due to post-stratification by calculating point estimates of all the within-strata and between-strata variances and plugging these values into Equation 9. We are not taking the variability of these estimates into account. By assuming the strata r_k are maximal, i.e., $r_k = 1$ for all k, we estimate a lower bound on the reduction in variance due to post-stratification. We show this for several different stratifications. For $K = 4$, we estimate the reduction of variance to be no less than 12%. More strata appear somewhat superior, but gains level off rather quickly. See Table 2.

The estimate of treatment effect changes markedly under post-stratification. The estimates $\hat{\tau}_{ps}$ hover around -0.28 for $K = 4$ and higher, as compared to the -0.13 from the simple-difference estimator. The post-stratified estimator appears to be correcting for a significant bias from random imbalance in treatment assignment. For $K = 2$, $\hat{\tau}_{ps} = -0.34$. Here we may be over-correcting for the imbalance in assignment.

If we assume the actual treatment effect is $\hat{\tau}_{ps}$, then we can estimate the MSE for both the simple-difference and post-stratified estimator conditioned on the imbalance by plugging point estimates into Equation 15 and Equation 14. These results are the last columns of Table 2; the percentage gain in this case is higher primarily due to the correction of the bias term from the imbalance. When conditioning on the imbalance W, the estimated *variance* of the post-stratified estimator is slightly higher than that of the simple-difference estimator, but the overall MSE is estimated to be significantly lower. This is due to the bias correction. Because the true variances and the r_k for strata are unknown, these gains are estimates only. They do, however, illustrate the potential value of post-stratification. Measuring the uncertainty of these estimates is an area of future work.

			Uncond. Variance			MSE Conditioned on W				
K	$\hat{\tau}_{ps}$	$\hat{\tau}_{sd}$	$\hat{\tau}_{ps}$	$\hat{\tau}_{sd}$	%	MSE$\hat{\tau}_{ps}$	var$\hat{\tau}_{sd}$	bias$\hat{\tau}_{sd}$	MSE$\hat{\tau}_{sd}$	%
2	-0.34	-0.13	0.077	0.081	5%	0.077	0.076	0.118	0.207	63%
4	-0.27	-0.13	0.071	0.081	12%	0.072	0.070	0.089	0.137	48%
10	-0.25	-0.13	0.070	0.081	13%	0.071	0.069	0.083	0.119	41%
15	-0.24	-0.13	0.070	0.081	14%	0.070	0.067	0.081	0.115	39%
30	-0.28	-0.13	0.069	0.081	15%	0.068	0.064	0.086	0.148	54%

Table 2: Estimated Standard Errors for PAC. Table shows both conditioned and unconditioned estimates for different numbers of strata.

Matched Pairs Estimation. We can also estimate the gains by building a fake set of potential outcomes by matching treated units to control units on observed covariates. We match as described in Sekhon and Grieve (2011). We then consider each matched pair a single unit with two potential outcomes. We use this synthetic set to calculate the variances of the estimators using the formula from Section 2 and Section 4.

Matching treatment to controls and controls to treatment gives a synthetic dataset with 1013 observations with all potential outcomes "known." The correlation of potential outcomes is 0.21 across all strata. The unconditional variance for the simple-difference and post-stratified estimators are 0.048 and 0.038, respectively. The percent reduction in variance due to post-stratification is 24.4%.

We can use this data set to further explore the impact of conditioning. Assume the treatment probability is $p = 0.5$ and repeatedly randomly assign a treatment vector and compute the resulting conditional variance. Also compute the "imbalance score" for the treatment vector with a chi-squared statistic:

$$\text{Imbalance} \equiv \sum_k \frac{(W_k(1) - pn_k)^2}{pn_k}$$

This procedure produces Figure 1. As imbalance increases, the MSE (variance) of $\hat{\tau}_{ps}$ steadily, but slowly, increases as well. The MSE of $\hat{\tau}_{ps}$ is fairly resistant to large imbalance. This is not the case for $\hat{\tau}_{sd}$, however. 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. Also, for a given imbalance, the simple-difference estimator can vary widely depending on whether strata-level bias terms are canceling out or not. This variability is not apparent for the post-

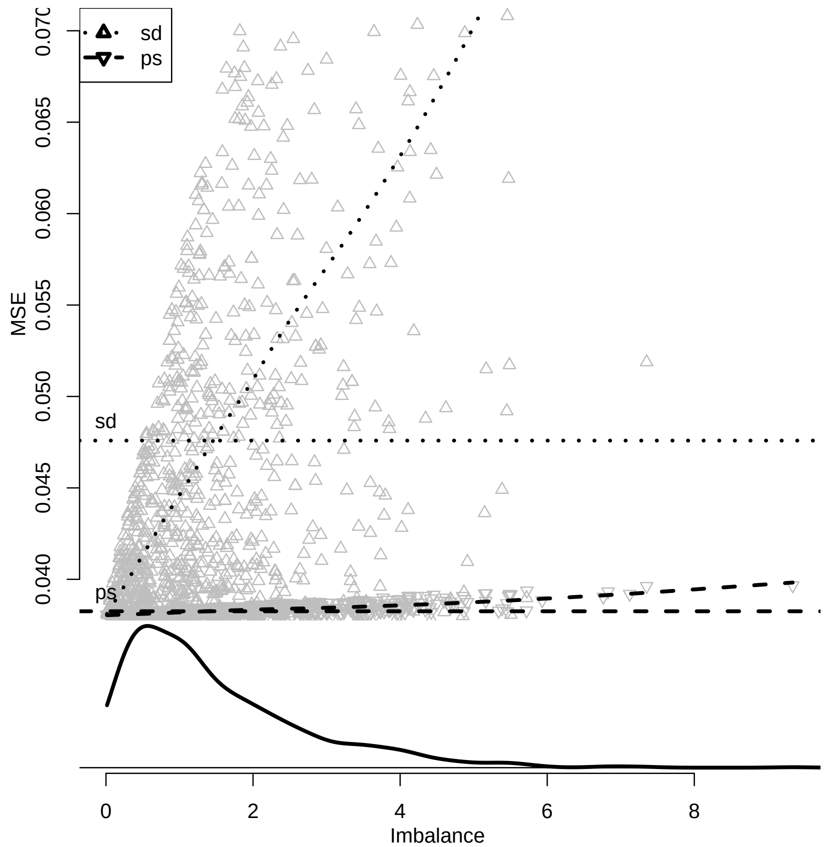

Figure 1: PAC MSE Conditioned on Imbalance. Uses constructed matched PAC dataset. Points indicate the conditional MSE of $\hat{\tau}_{ps}$ and $\hat{\tau}_{sd}$ given various specific splits of W. x-axis is the imbalance score for the split. Curved dashed lines interpolate point clouds. Horizontal dashed lines mark unconditional variances for the two estimators. The curve at bottom is the density of the imbalance statistic.

stratified estimator, where only the number of units treated drives the variance; the post-stratified points cluster closely to their trend line.

The curve at the bottom shows the density of the realized imbalance score: there is a good chance of a fairly even split with low imbalance. In these cases, the variance of $\hat{\tau}_{sd}$ is smaller than the unconditional formula would suggest. If the randomization turns out to be "typical" the unconditional variance formula would be conservative. If the imbalance is large, however, the unconditional variance may be overly optimistic. This chance of large imbalance with large bias is why the unconditioned MSE of $\hat{\tau}_{sd}$ is larger than that of $\hat{\tau}_{ps}$.

The observed imbalance for the actual assignment was about 2.37. The conditional MSE is 0.083 for $\hat{\tau}_{sd}$ and 0.039 for $\hat{\tau}_{ps}$. The conditional MSE for the simple-difference estimator is 73% larger than the unconditional MSE due to the bias induced by the imbalance. We would be overly optimistic if we were to use $\text{Var}[\hat{\tau}_{sd}]$ as a measure of certainty, given the observed, quite imbalanced, split W. For the post-stratified estimator the conditional variance is only about 1% higher than the unconditional; the degree of imbalance is not meaningfully impacting the precision. Generally, with post-stratification, the choice of using an unconditional or conditional formula is less of a concern.

Discussion. The PAC RCT has a strong predictor of outcome. Using it to post-stratify substantially increases the precision of the treatment effect estimate. Furthermore, post-stratification mitigates the bias induced by an unlucky randomization. Especially when concerned about imbalance, it is important to calculate conditional standard errors—not doing so could give overly optimistic estimates of treatment effect. This is especially true when using the simple-difference estimator.

7 Conclusions

Post-stratification is a viable approach to experimental design in circumstances where blocking is not feasible. If the stratification variable is determined beforehand, post-stratification is nearly as efficient as a randomized block trial would have been: the difference in variances between post-stratification and blocking is a small $O(1/n^2)$. However, the more strata, the larger the potential penalty for post-stratification. There is no guarantee of gains.

Conditioning on the observed distribution of treatment across strata allows for a more appropriate assessment of precision. Most often the observed balance will be good, even in moderate-sized experiments, and the conditional variance of both the post-stratified and simple-difference estimator will be smaller than estimated by the unconditional formula. However, as balance degrades having selected a truly prognostic covariate increases in importance: for a covariate unrelated to outcome, it is better in these cases to use a simple-difference estimator than a post-stratified one. For a prognostic covariate, the reverse is true.

When viewing a post-stratified or a blocked estimate as an estimate of the PATE, under the assumption that the sample is a random, independent, draw from a larger population, the potential for decreased precision is reduced. However, in most cases the sample in a randomized trial is not such a random draw. We therefore advocate for viewing the estimators as estimating the SATE, not the PATE.

Problems arise when stratification is determined after treatment assignment. The results of this paper assume that the stratification is based on a fixed and defined covariate b. However, in practice covariate selection is often done after-the-fact in part because, as is pointed out by Pocock et al. (2002), it is often quite difficult to know which of a set of covariates are significantly prognostic *a priori*. But variable selection invites fishing expeditions, which undermine the credibility of any findings. Doing variable selection in a principled manner is still notoriously difficult, and is often poorly implemented; Pocock et al. (2002), for example, found that many clinical trial analyses select variables inappropriately. Tsiatis et al. (2007) summarize the controversy in the literature and, in an attempt to move away from strong modeling, and to allow for the implicit multiple-testing in selecting an optimal subset of variables from all possible, propose a semiparametic approach as a solution.

Beach and Meier (1989) suggests that, at minimum, all potential covariates for an experiment be listed in the original protocol. Call these z. In our framework, variable-selection is then to *build* a stratification b from z *and* T after having randomized units into treatment and control. b (now B) is random as it depends on T. Questions immediately arise: how does one define the variance of the estimator? Can substantial bias be introduced by the strata-building process? The key to these questions likely depends on appropriately conditioning on the final, observed, strata and the process of constructing B. This is an important area of future work.

8 Appendix A

Theorem 2.1. The proof of Theorem 2.1 is based on iterated expectations and a lot of unpleasant algebra. The following shows the hilights. See the supplementary material for a version with more detail. We first set up a few simple expectations. Under Assignment Symmetry,

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

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

$$\mathbb{E}\left[\frac{T_i^2}{W_k^2(1)}\right] = \mathbb{E}\,\mathbb{E}\left[\frac{T_i}{W_k^2(1)}|W_k(1)\right] = \frac{1}{n_k}\,\mathbb{E}\left[\frac{1}{W_k(1)}\right] = \frac{\beta_{1k}+1}{n_k^2}. \tag{18}$$

These derivations are easier if we use $\alpha_{1k} \equiv \mathbb{E}[1/W_k(1)]$, but the β's are more interpretable and lead to nicer final formula. There are analogous formula for the control unit terms and cross terms. We use these relationships to compute means and variances for the strata-level estimators.

Unbiasedness. The strata-level estimators are unbiased:

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

Variance. $\mathrm{Var}[\hat{\tau}_k] = \mathbb{E}\left[\hat{\tau}_k^2\right] - \tau_k^2$. Expand τ_k^2 into three parts $(a)' - (b)' + (c)'$:

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

Similarly, expand the square of $\mathbb{E}\left[\hat{\tau}_k^2\right]$ to get $(a) - (b) + (c)$. Simplify these parts. For example, algebra and relationships such as shown in Equation 18 give

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

Part (b) and (c) are similar.

The variance is then $\text{Var}[\hat{\tau}_k] = (a) - (a') - (b) + (b') + (c) - (c')$, a sum of several ugly differences. Algebra, and recognizing formulas for the sample variances and covariances, gives:

$$(a) - (a') = \frac{\beta_{1k}}{n_k}\sigma_k^2(1)$$
$$(b) - (b') = -\frac{2}{n_k}\gamma_k(1,0)$$

and

$$(c) - (c') = \frac{\beta_{0k}}{n_k}\sigma_k^2(0)$$

Sum these differences to get Equation 5.

Theorem 4.1. Calculate the MSE of $\hat{\tau}_{sd}$ conditioned on the split W with a slight modification to the above derivation. Define a new estimator that is a weighted difference in means:

$$\hat{\alpha}_k \equiv A_k \sum_{i:b_i=k} \frac{T_i}{W_k(1)} y_i(1) - B_k \sum_{i:b_i=k} \frac{1-T_i}{W_k(0)} y_i(0)$$

with A_k, B_k constant. $\hat{\alpha}_k$ is an unbiased estimator of the difference in means weighted by A_k and B_k:

$$\mathbb{E}[\hat{\alpha}_k] = \mathbb{E}\left[A_k \sum_{k:b_i=b} \frac{T_k}{W_k(1)} y_k(1) - B_k \sum_{k:b_i=b} \frac{T_k}{W_k(0)} y_k(0)\right] = A_k \bar{y}_k(1) - B_k \bar{y}_k(0).$$

Now follow the derivation of the variance of $\hat{\tau}_k$ propagating A_k and B_k through. These are constant and they come out, giving

$$\text{Var}[\hat{\alpha}_k] = \frac{1}{n_k}\left[A_k^2 \beta_{1k}\sigma_k^2(1) + B_k^2 \beta_{1k}\sigma_k^2(0) - 2A_k B_k \gamma_k(1,0)\right].$$

Expand $\hat{\tau}_{sd}$ into strata terms:

$$\hat{\tau}_{sd} = \sum_{k=1}^{K} \frac{W_{1k}}{W_1} \sum_{i:b_i=k} \frac{T_i}{W_{1k}} y_i(1) - \frac{W_{0k}}{W_0} \sum_{i:b_i=k} \frac{1-T_i}{W_{0k}} y_i(0) = \sum_{k=1}^{K} \hat{\alpha}_k$$

with $A_k = W_{1k}/W_1$ and $B_k = W_{0k}/W_0$. Conditioning on W makes the A_k and the B_k constants. Assignment symmetry ensures the strata are independent, so the $\hat{\alpha}_k$ are as well, and the variances then add:

$$\text{Var}[\hat{\tau}_{sd}|W] = \sum_{k=1}^{K} \text{Var}[\hat{\alpha}_k].$$

The bias is $\mathbb{E}[\hat{\tau}_{sd}|W] - \tau$ with

$$\mathbb{E}[\hat{\tau}_{sd}|W] = \sum_{k=1}^{K} \mathbb{E}[\hat{\alpha}_k|W] = \sum_{k=1}^{K} A\bar{y}_k(1) - B\bar{y}_k(0).$$

Expand τ as in Equation 2 and rearrange terms.

Extending to PATE. First, decompose the variance:

$$\text{Var}[\hat{\tau}_{ps}|\mathcal{D}] = \mathbb{E}[\text{Var}[\hat{\tau}_{ps}|\mathcal{S},\mathcal{D}]\,|\mathcal{D}] + \text{Var}[\mathbb{E}[\hat{\tau}_{ps}|\mathcal{S},\mathcal{D}]\,|\mathcal{D}]$$

The first term is simply the expectation of Equation 6, the SATE variance formula. Since $\mathcal{S}$ is random, so are the $\sigma_k^2(\ell)$, etc. The expectation of these quantities over $\mathcal{S}$ gives the population parameters as they are unbiased estimators. The β's are all constant, and $\mathcal{D}$ is independent of $\mathcal{S}$. Therefore:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}[\text{Var}[\hat{\tau}_{ps}|\mathcal{S},\mathcal{D}]] &= \mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_k \frac{n_k}{n}\left[\beta_{1k}\sigma_k^2(1) + \beta_{0k}\sigma_k^2(0) + 2\gamma_k(1,0)\right]|\mathcal{D}\right] \\
&= \frac{1}{n}\sum_k \frac{n_k}{n}\left[\beta_{1k}\sigma_k^2(1)^* + \beta_{0k}\sigma_k^2(0)^* + 2\gamma_k(1,0)^*\right]. \qquad (19)
\end{aligned}$$

The second term is

$$\begin{aligned}
\text{Var}[\mathbb{E}[\hat{\tau}_{ps}|\mathcal{S},\mathcal{D}]] &= \text{Var}[\tau] \\
&= \text{Var}\left[\sum_{k=1}^K \frac{n_k}{n}\tau_k\right] \\
&= \frac{n_k^2}{n^2}\sum_{k=1}^K \text{Var}[\bar{y}_{k1} - \bar{y}_{k0}] \\
&= \frac{n_k^2}{n^2}\sum_{k=1}^K \frac{1}{n_k}\left[\sigma_k^2(1)^* + \sigma_k^2(0)^* - 2\gamma_k(1,0)^*\right]. \qquad (20)
\end{aligned}$$

Sum Equation 19 and Equation 20 to get the PATE-level MSE.

9 Appendix B

$\beta_{\ell k}$ can be approximated by $\mathbb{E}[W_k(1-\ell)]\,/\,\mathbb{E}[W_k(\ell)]$. For example, in the complete randomization case $\beta_{1k} \approx (1-p)/p$. Generally, the β's are larger than their approximations. They can be less, but only by a small amount. For complete randomization and Bernoulli assignment, the difference between the β's and their approximations is bounded by the following theorem:

Theorem 9.1. *Take an experiment with n units randomized under either complete randomization or Bernoulli assignment. Let p be the expected proportion of units treated. Let $\mathcal{D}$ be the event that $\hat{\tau}_{ps}$ is defined. Without loss of generality assume $0.5 \leq p < 1$. Let n_{min} be the smallest strata size. Then $\beta_{1k} - (1-p)/p$ is bounded above by:*

$$\begin{aligned}
\beta_{1k} - \frac{1-p}{p} &\leq \frac{4}{p^2}\frac{1}{n_k} + \max\left[\left(\frac{n_k}{2} - \frac{4}{p^2 n_k}\right)e^{-\frac{p^2}{2}n_k}, 0\right] + n_k(1-p)^{n_k+1} + 2n_k K\,(p)^{n_{min}} \\
&= \frac{4}{p^2}\frac{1}{n_k} + O(e^{-n_k}).
\end{aligned}$$

Furthermore, it is tightly bounded blow:

$$\beta_{1k} - \frac{1-p}{p} \geq -\frac{2}{p}(1-p)^{n_k} - 2n_k K(p)^{n_{min}} = -O(e^{-n_k}).$$

Similar results apply for the β_{0k} and β_ℓ.

Proof. Start without conditioning on $\mathcal{D}$. $W_{1k} = \sum T_i$ with $T_i \in \{0, 1\}$. For Bernoulli assignment, the T_i are i.i.d Bernoulli variables with probability p of being 1. For completely randomized experiments, the W_{1k} are distributed according to a hypergeometric distribution, i.e., as the number of white balls drawn in n_k draws without replacement from an urn of n balls with np white balls. Regardless, $\mathbb{E}[W_{1k}] = n_k p$.

Define $Y_{n_k} \equiv (n_k/W_{1k}) \times \mathbf{1}_{\{W_{1k}>0\}}$. Due to the indicator function, $Y_{n_k} \leq n_k$. Given $\mathcal{D}$, the event that *all* strata-level estimators are well-defined, $Y_{n_k} = n_k/W_{1k}$ so

$$\beta_1 - \frac{1-p}{p} = \mathbb{E}\left[\frac{W_{0k}}{W_{1k}}|\mathcal{D}\right] - \frac{1-p}{p} = \mathbb{E}\left[\frac{n_k}{W_{1k}}|\mathcal{D}\right] - \frac{1}{p} = \mathbb{E}[Y_{n_k}|\mathcal{D}] - \frac{1}{p}.$$

We first show the probability of $\neg\mathcal{D}$ is very small, which will allow for approximating the expectation of the conditioned Y_{n_k} with the unconditioned. If n_{min} is the size of the smallest strata, then

$$\begin{aligned}
\mathbf{P}\neg\mathcal{D} &\leq \sum_{k=1}^{K} \mathbf{P}\{W_{1k} = 0 \text{ or } W_{0k} = 0\} \\
&\leq 2K \max_{\ell=0,1;k=1,\ldots,K} \mathbf{P}\{W_{\ell k} = 0\} \\
&\leq 2K\,(p)^{n_{min}}.
\end{aligned}$$

Expand the expected value of Y as

$$\mathbb{E}[Y_{n_k}] = \mathbb{E}[Y_{n_k}|\mathcal{D}]\,\mathbf{P}\mathcal{D} + \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\,\mathbf{P}\neg\mathcal{D}.$$

Use this and the bound $Y_{n_k} \leq n_k$ to get

$$\begin{aligned}
\Big|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}]\Big| &= \Big|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}|\mathcal{D}]\,\mathbf{P}\mathcal{D} - \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\,\mathbf{P}\neg\mathcal{D}\Big| \\
&= \Big|\mathbb{E}[Y_{n_k}|\mathcal{D}]\,(1 - \mathbf{P}\mathcal{D}) - \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\,\mathbf{P}\neg\mathcal{D}\Big| \\
&= \Big|\mathbb{E}[Y_{n_k}|\mathcal{D}] - \mathbb{E}[Y_{n_k}|\neg\mathcal{D}]\Big|\mathbf{P}\neg\mathcal{D} \\
&\leq n\mathbf{P}\neg\mathcal{D} = 2nK\,(p)^{n_{min}}
\end{aligned} \tag{21}$$

This shows that $\mathbb{E}[Y_{n_k}]$ is quite close to $\mathbb{E}[Y_{n_k}]$, i.e.

$$\mathbb{E}[Y_{n_k}] - 2nK\,(p)^{n_{min}} \leq \beta_1 - \frac{1-p}{p} \leq \mathbb{E}[Y_{n_k}] + 2nK\,(p)^{n_{min}}.$$

Now we need the following lemma to get a handle on $\mathbb{E}[Y_{n_k}]$:

Lemma 9.2. *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*

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

See the supplementary material for proof. The lemma uses results from Hoeffding (1963). Use Lemma 9.2 on $\mathbb{E}[Y_{n_k}]$. This gives our stated bounds.

□

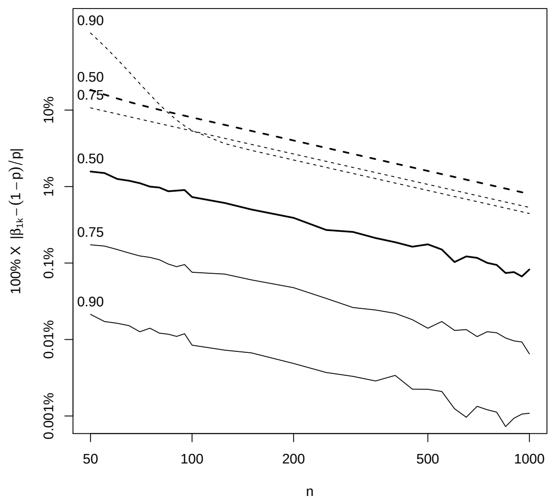

Figure 2: log-log plot comparing actual difference between β values to their approximations and the bounds provided by Lemma 9.2. Four probabilities of assignment shown, $p = 0.5, 0.75, and 0.9$. Actual differences computed with monte carlo.

Remark on Lemma 9.2. Numerical calculation shows the constants of the $1/n$ term are overly large, but the rate of $1/n$ appears to be correct. Figure 2 show a log-log plot of the actual percent increase of the β's over $1/p$ for several values of p and n along with the calculated bounds. When the exponential term becomes negligible, the bound appears to be about 8, 15, and 42 times bigger for $p = 0.5, 0.75$, and 0.9 respectively, i.e., the constants on the $1/n$ term are overstated by this much. The log-log slope is -1 suggesting the $1/n$ relationship.

Proof of Theorem 2.4. Assume the conditions of Theorem 9.1 and consider Equation 9. First,

$$\begin{aligned}
\left|\frac{n_k}{n}\beta_{1k} - \frac{n_k-1}{n-1}\beta_1\right| &\leq \left|\frac{n_k}{n}\beta_{1k} - \frac{n_k}{n}\beta_1\right| + \left|\frac{n_k-1}{n-1}\beta_1 - \frac{n_k}{n}\beta_1\right| \\
&= \frac{n_k}{n}\left|\beta_{1k} - \beta_1\right| + \beta_1\left|\frac{n_k-1}{n-1} - \frac{n_k}{n}\right| \\
&\leq \frac{n_k}{n}\frac{4}{p^2}\frac{1}{fn} + \frac{1-p}{p}\frac{1}{n} + O(\frac{1}{n^2})
\end{aligned}$$

Because the lower bound is so tight, we don't need to double the bound from Theorem 9.1 for bounding the difference $|\beta_{1k} - \beta_1|$. Plug these into the sums of Equation 9 and replace all variances and covariances with the maximum variance and covariance. The bound follows.

Proof of Theorem 3.1. This is handled the same way as for Theorem 2.4, but is more direct.

References

Abadie, A. and G. W. Imbens (2007). Estimation of the conditional variance in paired experiments. matched pairs experiments — estimating the variance of the estimator more appropriately/closely.

Beach, M. L. and P. Meier (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials 10*, 1615–1735.

Chittock, D., V. Dhingra, J. Ronco, J. Russell, D. Forrest, M. Tweeddale, and J. Fenwick (2004). Severity of illness and risk of death associated with pulmonary artery catheter use. *Critical Care Medicine 32*, 911–915.

Connors, A., T. Speroff, N. Dawson, C. Thomas, F. Harrell, D. Wagner, N. Desbiens, L. Goldman, A. Wu, R. Califf, W. Fulkerson, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. Knaus (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association 276*, 889–897.

Dalen, J. (2001). The pulmonary artery catheter—friend, foe, or accomplice? *Journal of the American Medical Association 286*, 348–350.

Finfer, S. and A. Delaney (2006). Pulmonary artery catheters as currently used, do not benefit patients. *British Medical Journal 333*, 930–1.

Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *The annals of applied statistics 2*(1), 176–196.

Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics 40*, 180–193.

Hartman, E., R. D. Grieve, and J. S. Sekhon (2011). From sate to patt: The essential role of placebo tests for combining experimental with observational studies.

Harvey, S., D. Harrison, M. Singer, J. Ashcroft, C. Jones, D. Elbourne, W. Brampton, D. Williams, D. Young, and K. Rowan (2005). An assessment of the clinical effectiveness of pulmonary artery catherters in patient management in intensive care (pac-man): a randomized controlled trial. *Lancet 366*, 472–77.

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

Holt, D. and T. M. F. Smith (1979). Post stratification. *J. R. Statistic Society 142*(Part 1), 33–46.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine 27*, 4857–4873.

Imai, K., G. King, and C. Nall (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science 24*(1), 29–53.

Imai, K., G. King, and E. A. Stuart (2008). Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Statistic Society 171*(Part 2), 481–502.

Imbens, G. W. (2011). Experimental design for unit and cluster randomized trials.

Keele, L., C. McConnaughy, and I. White (2009). Adjusting experimental data. In *Experiments in Political Science*.

Lin, W. (2010). Agnostic notes on regression adjustment to experimental data. Working Paper.

McHugh, R. and J. Matts (1983). Post-stratification in the randomized clinical trial. *Biometrics 39*, 217–225.

Pocock, S. J., S. E. Assmann, L. E. Enos, and L. E. Kasten (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: curent practice and problems. *Statistics in Medicine 21*, 2917–2930.

Reichardt, C. S. and H. F. Gollob (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychological Methods 4*(1), 117–128.

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

Sakr, Y., J. Vincent, K. Reinhart, D. Payen, C. Wiedermann, D. Zandstra, and C. Sprung (2005). Sepsis occurrence in acutely ill patients investigators. use of the pulmonary artery catheter is not associated with worse outcome in the ICU. *Chest 128*(4), 2722–31.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science 12*, 487–508.

Sekhon, J. S. and R. D. Grieve (2011). A matching method for improving covariate in cost-effectiveness analysis. *Health Economics*. Forthcoming.

Senn, S, J. (1989). Covariate imbalance andrandom allocation in clinical trials. *Statistics in Medicine 8*, 467–475.

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed ([1923] 1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472. ArticleType: research-article / Full publication date: Nov., 1990 / Copyright 1990 Institute of Mathematical Statistics.

Tsiatis, A. A., M. Davidian, M. Zhang, and X. Lu (2007). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine 27*, 4658–4677.