Appraising covariate balance after assignment to treatment by groups*
Technical Report # 436

Ben B. Hansen
Statistics Department
University of Michigan
ben.hansen@umich.edu

7 April 2006

Abstract

In a randomized experiment, treatment and control groups should be roughly the same — balanced — in their distribution of pre-treatment variables. But how nearly so? Can descriptive comparisons meaningfully be paired with significance tests? Should there be several such tests, one for each pre-treatment variable, or should there be a single, omnibus test? Is there a large-sample test that is reliable in samples of moderate size, notwithstanding recent evidence to the contrary about one natural large-sample procedure, or is simulation needed for reliable appraisals of balance? What new concerns are introduced by random assignment of clusters? Which tests of balance are optimal?

To address these questions, Fisher's randomization inference is applied the question of balance. The procedures that result are not new, although certain arguments for them are. Its application suggests the reversal of published conclusions about two studies, one clinical and the other a field experiment in political participation.

Key words: cluster; contiguity; community intervention; group randomization; matching; randomization inference

*The author thanks Jake Bowers, Alan Gerber, Donald Green, Nancy Reid and Paul Rosenbaum for helpful discussions. Portions of this work were presented at the 2005 meetings of the Political Methodology Society and in seminars at the Department of Medicine, Case Western Reserve University, and at the Department of Biostatistics, Yale University. The author is grateful for comments received at these venues.

1 Introduction

In a controlled, randomized experiment, treatment and control groups should be roughly the same — balanced — in their distribution of pre-treatment variables. But how nearly so? Reports of clinical trials are urged to present tables of treatment and control group means of x-variables (Campbell et al. 2004), and often do. These greatly assist qualitative assessments of similarity and difference between the groups, but in themselves they are silent as to whether, given the design, the discrepancies between the groups are large or small. Can the descriptive comparisons meaningfully be paired with significance tests? If so, must there be several, one for each variable, or can there be a single omnibus test? Would the omnibus test always require a simulation experiment, as proposed at some places in the literature on random assignment by group (Raab and Butcher 2001)? Is there a large-sample test that is reliable in samples of moderate size, notwithstanding recent evidence to the contrary about one natural procedure (Gerber and Green 2005)? At the level of foundations, some authors note that to posit a data-generating model for x-variables is somewhat antithetic to the removal of parametric assumptions (Ho et al. 2005). Does testing for balance require such a model, as these authors also claim, or are there tests that narrowly probe data's conformity to the experimental ideal? At the level of technical detail, tests based differences of group means require precise instructions for combining differences across strata or blocks, with the optimal approach appearing to depend on within- and between-block variation in x — population variation, not sample variation (Kalton 1968). Doesn't the fine-tuning of these instructions require assumptions about, or estimation of, variability in the superpopulation, introducing sources of uncertainty that are generally ignored at the stage of inference (Yudkin and Moher 2001)? Without speculative superpopulations of x-values, how are alternatives to the null hypothesis to be conceived? What tests are optimal against these alternatives?

The most familiar randomized comparisons of human subjects, perhaps, are drug and vaccine studies. Generally these are randomized at the level of individuals. But interventions upon neighborhoods, classrooms, clinics, and families are increasingly the objects of study, and are increasingly studied experimentally; and even non-experimental interventions at the group level may analyzed using a combination of poststratification and analogies with hypothetical experiments. Might it be safe to ignore the group structure (as outcome analyses of cluster-randomized data often do [MacLennan et al. 2003; Isaakidis and Ioannidis 2003], in some conflict with the recommendations of methodologists [Gail et al. 1996; Murray 1998; Donner and Klar 2000]) if interest focuses on individual-level outcomes, if correlations within group are low, or if the groups are small? Or do methods appropriate to individual-level assignment readily generalize to assignment by group? If so, then how is this best done?

1.1 Example: a clinical trial with randomization at the clinic level

In order to study the benefit of up-to-date, best practices in monitoring and treatment of coronary heart disease, the ASSIST trial randomized 14 of 21 participating clinics to receive new systems for the regular review of heart disease patients (Yudkin and Moher 2001). For external validity of outcome analyses, treatment and control groups would have to be well balanced, at baseline, on the proportion of patients adequately assessed, and on other outcome variables. As is evident from Table 1 this aim is somewhat complicated by the fact that clinics varied greatly in size and in patient characteristics. Seemingly sizable differences between treatment and control groups' proportions of adequately assessed patients may still compare favorably with differences that would have obtained in alternate random assignments. Viewed in isolation, such a difference would appear to threaten external validity, although the appearance would be misleading. A principled means of distinguishing threatening and unthreatening cases is needed.

	Numbers of coronary heart disease patients...				
		adequately	treated with		
Practice	in total	assessed	aspirin	hypotensives	lipid-reducers
3	38	6	30	17	6
6	58	19	38	31	16
9	91	23	60	56	22
12	114	46	86	60	35
15	127	58	103	86	30
18	138	68	106	86	57
21	244	93	181	93	63

Table 1: Sizes of a subset of the 21 clinics participating in the ASSIST trial of register and recall systems for heart disease patients, along with baseline measurements of primary and secondary outcome variables. Despite the great variation in practice sizes, and in practice benchmarks the intervention sought to improve, a balanced allocation of practices to treatment conditions was sought. Adapted from Yudkin and Moher (2001, Table II).

A related need is for metrics with which to appraise the likely benefit, in terms of balance, of randomizing within blocks of relative uniformity on baseline measures.

1.2 Example: a field experiment on political participation

A second case in point is A. Gerber and D. Green's Vote '98 campaign, a voter turnout intervention in which get-out-the-vote (GOTV) appeals were randomly assigned. Gerber and Green's original report assumed the targeting of appeals to be independent of subjects' covariates, finding that in-person appeals effectively stimulated voting whereas solicitations delivered over the telephone, by professional calling firms, had little or no effect (2000). Criticizing this analysis, Imai observes that data Gerber and Green made available alongside their publication did not accord with the assumption of independence (2005). So poorly balanced are the groups, writes Imai, that the hypothesis of independence can be rejected at the 10^{-4} level (2005, Table 6). Had experimental protocol broken down, effectively spoiling the random assignment? Imai deduces that it must have, dismissing the original analysis and instead mounting another upon very different assumptions. Contrary to Gerber and Green, Imai's revision attaches significant benefits to paid GOTV calls.

In a pointed response, Gerber and Green (2005) shift doubt from the implementation of their experiment to Imai's methodology — particularly, the method by which he checks for balance. Their counter-attack has three fronts. First, they explain that the original experiment's randomization was performed at the household level, rather than the individual level. Since Imai's independence check assumed independence at the subject level, no theory supports tests associated with it. Second, they present results from a replication of the telephone GOTV experiment on a much larger scale, now randomizing individuals rather than households. The replication results were consistent with those of the the original study. Third, they present simulation evidence that would cast doubt on Imai's recommended balance tests even had randomization been as he assumed. Those tests carried an asymptotic justification, for which the Vote '98 sample appears to have been too small — even though it comprised some 30,000 subjects, in more than 20,000 households!

The manifold nature of this argument makes methodological lessons difficult to draw. If the conclusion that the Vote '98 treatment assignment lacked balance is mistaken, then did the mistake lie in the conflation of household and individual level randomization, in the use of an inappropriate statistical test, or both?

1.3 Structure of the paper

This and Section 2 introduce the paper. Section 3 describes the Fisherian model for comparative studies and its consequences for the difference of group means and variations on it, arguing for one such variation as a general measure of balance on a covariate. Section **??** adapts this measure to studies with group assignment to treatment, and to testing for balance on several variables simultaneously. Section 5 develops theoretical arguments for the optimality of this approach, and for the setting of a tuning constant, which Section 6 illustrates uses of the methodology at design and at analysis stages. Section 7 concludes.

2 Two ways not to check for balance

This section examines appealing but ad-hoc adaptations of two standard techniques, the method of standardized differences and goodness-of-fit testing with logistic regression, to the problem of testing for balance after random assignment of groups. To illustrate, I use the rich and publicly available Vote '98 dataset (Gerber and Green 2005). It describes some 31,000 voters, falling in about 23,000 households; to complement this unusually large randomized experiment with a smaller one, we consider a simple random subsample of voters falling in 100 households. Telephone reminders to vote were attempted to roughly a fifth of the subjects, and it is around the putative randomness of this treatment assignment that Gerber, Green, and Imai's debate centers; we study the association of this treatment assignment z with covariates $\mathbf{x}$, which include age, ward of residence, registration status at the time of the previous election, whether a subject had voted in that election, and whether he had declared himself a member of a major political party.

2.1 Standardized differences of measurement units

Let us contrast *observation* or *measurement* units, here voters, with *clusters* or *assignment units*, here households containing one or two voters. The *standardized difference of measurement units* in a numeric variable x is a scaled difference of the average of x-values among measurement units in the treatment group and the corresponding average for controls. To facilitate interpretation, the difference is scaled by the reciprocal of one s.d. of measurement x's, so that $100 \times$ (standardized difference) can be read as a percent fraction of an s.d.'s difference.

Setting aside this scaling for the purpose of mounting a statistical test, one has differences $\bar{x}_t - \bar{x}_c$, or, in vector notation, $\mathbf{z}^t\mathbf{x}/\mathbf{z}^t\mathbf{1} - (1-\mathbf{z})^t\mathbf{x}/(1-\mathbf{z})^t\mathbf{1}$, where $\mathbf{z} \in \{0,1\}^n$ indicates assignment to the treatment group. See this as a random variable, conditioning on the measurement units' x values and on the numbers of measurement units $m_t = \mathbf{Z}^t\mathbf{1}$ and $m_c = (\mathbf{1}-\mathbf{Z})^t\mathbf{1}$ in the treatment and control groups gives the random sum

$$\frac{\mathbf{Z}^t\mathbf{x}}{m_t} - \frac{(\mathbf{1}-\mathbf{Z})^t\mathbf{x}}{m_c} = k\mathbf{Z}^t\mathbf{x} - \mathbf{1}^t\mathbf{x}/m_c, \tag{1}$$

$k = m_c^{-1} + m_t^{-1}$. Were treatment-group measurement units a simple random subsample of the sample as a whole, this difference would have mean zero and variance equal to $(m_t m_c/m)s^2(\mathbf{x})$, for $s^2(\mathbf{x}) = (m-1)^{-1}\sum_i (x_i - \bar{x})^2$. Consider instead the case where a simple random sample of clusters of measurement units, not of measurement units themselves, are selected for treatment, but the analysis pretends the opposite. Gerber and Green (2000) make this simplification — perhaps because the Vote '98 data have as little clustering as one might hope to find in a group randomized study, with never more than two subjects, and often only just one, to a cluster.

With this simplification, differences $\bar{x}_t - \bar{x}_c$ are readily converted to z-scores. Table 2 appraises accuracy of the resulting approximate p-values, comparing them to simulation p-values and to p-values attached to the statistic $\mathbf{z}^t\mathbf{x}$ by an approximation to be discussed below. The simulation mimics the structure of the experiment's actual design, forming simulated treatment groups from random samples of 5,275 of the 23,450 households, calculating differences d_x^* in means of measurement unit x-values in the simulated treatment and control groups, and comparing these differences to the treatment-control group difference, d_x, observed in the actual sample. The comparisons are summarized with two-sided mid-p values, averages of the proportion of simulated differences d_x^* of greater magnitude than d_x and the proportion of d_x^*'s of magnitude at least as large as that of d_x. The treatment group was reshuffled 10^6 times, so the mid-p values are accurate to within .001.

I had expressed the nominal "Ward" variable as 29 indicator variables, one for each ward, and the age measurement in terms of cubic B-splines with knots at quintiles of the age distribution, yielding six new x-variables. The table displays the four of the 29 ward indicators, and the four of the six spline basis variables, for which the approximate p-values ignoring groups were most and least discrepant from actual p-values in the subsample and the full sample.

	100 households ($m = 133$)			All households ($m = 31$K)		
	Accounting for groups?			Accounting for groups?		
x	No	Yes	Actual	No	Yes	Actual
`num.voters.in.hh`	.12	.24	.21	.85	.82	.82
`voted.prev.elec`	.40	.22	.22	.23	.39	.39
`maj.pty.member`	.45	.14	.16	.24	.18	.18
`age.Bspline.2`	.68	.59	.60	.06	.31	.31
`age.Bspline.4`	.82	.39	.40	.68	.68	.68
`age.Bspline.5`	.72	.24	.24	.39	.22	.22
`age.Bspline.6`	.19	.62	.62	.56	.89	.89
`Ward.2`	1.00	1.00	.50	.81	.87	.87
`Ward.5`	.89	.98	.89	.44	.47	.48
`Ward.10`	.58	.54	.65	.95	.97	.97
`Ward.11`	.75	.92	.87	.27	.42	.42

Table 2: Effect of accounting for assignment by groups on approximations to p-values, in the full Vote '98 sample and in a subsample of 100 households.

The approximation ignoring the clustered nature of the randomization is not particularly good. Its p-values differ erratically from the actual p-values, at some points incorrectly suggesting departures from balance and elsewhere exaggerating it. Increasing the sample size from 133 to 31 thousand appears to improve the approximation somewhat, but not nearly as much as it does the approximation that accounts for clustering. It is noteworthy that so

striking a discrepancy arises even with only half the experimental subjects assigned as part of a cluster, and with no clusters larger than two.

2.2 The p-value from logistic regression of z on x's

With or without treatment assignment by clusters, and with or without analytic adjustments to account for clusters, the method of standardized differences has the limitation that it produces a long list of non-independent p-values, one for each x-variate studied. In order either to decide prospectively whether a possible treatment assignment is balanced enough for use, or to retrospectively appraise the need for post-stratification, just a few p-values, ideally one, would be more convenient.

Logistic regression seems well suited to the task. Regress treatment assignment, z, on covariates x and a constant, then on the constant alone, then compare the two fits using a standard asymptotic likelihood-ratio test. Should the asymptotics of this deviance test apply, it will reject (at the .05 level) no more than about 5% of treatment assignments, presumably the ones in which, by coincidence, covariate balance failed to obtain. The problem with this procedure is that its sample-size requirements are more stringent than one might think, are difficult to ascertain, and are typically incompatible with checking for balance thoroughly.

	Size of test			
	Asymptotic			
Method	.001	.01	.05	.10
	Actual			
Logistic regression-based	.0281	.0620	.16	.24
Standardized differences	.0000	.0003	.018	.064

Table 3: Small-sample ($n = 21$) Type I error rates of two types of test, one based on logistic regression and another, to be described in Section 4, based on standardized differences. The actual size of the logistic regression tests well exceed their nominal levels, while the alternate test is somewhat conservative but holds to advertised levels. Based on 10^6 simulated assignments to treatment of 14 of the 21 ASSIST clinics.

Table 3 shows the small-sample performance of the logistic regression deviance test, presenting the actual sizes of asymptotic level .001, .01, .05, and .10 tests as applied to assignments of 14 of Yudkin and Maher's 21 clinics to treatment. The test's Type I error rates are markedly too high. Perhaps poor performance of asymptotic tests is to be expected, given the small sample size; but it is noteworthy that another asymptotic test, Section 4's method of combined standardized differences, succeeds in maintaining sizes no greater than advertised levels of significance.

Figure 1 illustrates the limited accuracy of the logistic regression approach in samples of moderate size. It compares asymptotic and actual null distributions of p-values from the logistic regression deviance test, effecting the actual distribution by simulation. One

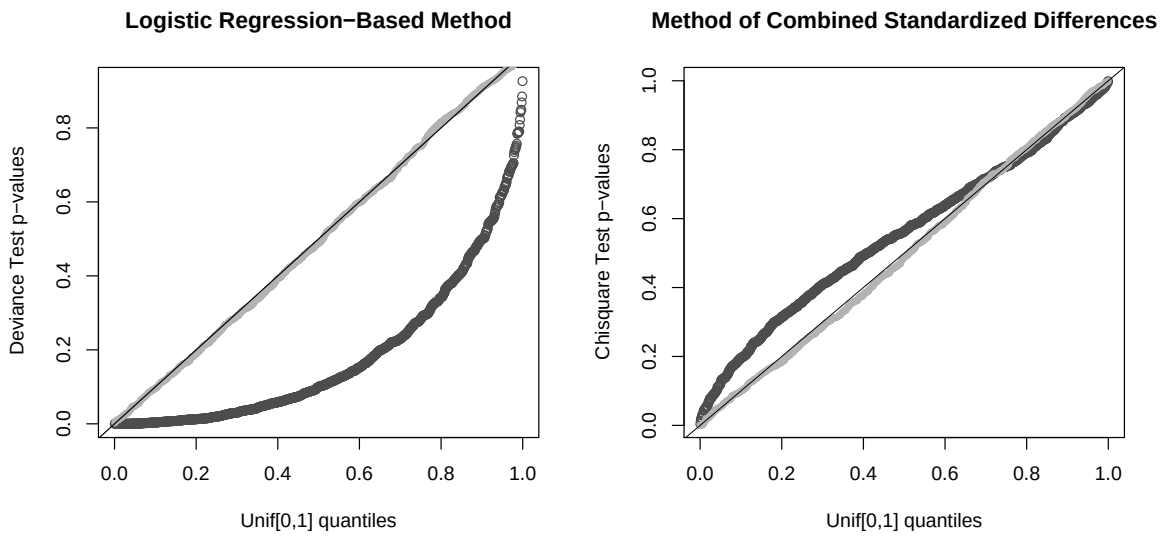

Figure 1: Theoretical and actual p-values of omnibus tests of covariate balance. With 100 assignment units and 38 degrees of freedom (dark trace), logistic regression's p-values are markedly too small, whereas p-values from the method to be described in § 4 err towards conservatism, and to a lesser degree. With the same 38 degrees of freedom and with the full 23,000 assignment units (lighter trace), both methods perform well.

thousand simulation replicates are shown, both for the 100-household subsample and for the full sample. The covariates $x_{(1)}, \ldots, x_{(k)}$ are those described in Section 2.1, with x-values for two-person households determined by summing x-values of individuals in each household.

While p-values based on the asymptotic approximation appear accurate for the full sample, with its 23 thousand-someodd households, those for the subsample are quite exaggerated. In it, the nominal .05-level test has an actual size of about .37. Would an alert applied statistician have identified the 100-household subsample as too small for the likelihood ratio test? Perhaps; it has only $2\frac{1}{2}$ times as many observations as x-variables, once the Age and Ward variables have been expanded as in Table 2. But how large a ratio of observations to covariates would be sufficient? Principles to settle the question are not forthcoming, particularly for regression with binary rather than continuous outcomes, and intuition may be a poor guide. To explore the difference in information carried by binary and continuous outcomes, Brazzale, Davison, and Reid (2006, §4.2; see also Davison 2003, ex. 10.17) construct artificial data sets from a real one with a binary independent variable, some retaining the binary outcome structure but increasing the apparent information in the data set by replicating observations, and others imputing continuous outcomes according to a logistic distribution. The results are striking; one observation with continuous response carries about as much information as eight observations with binary response, and deviance tests are found to be unreliable even

with eleven times as many observations as x-variables.

For contrast, the right panel of Figure 1 offers an analogous comparison between asymptotically approximate and actual p-values of a test statistic to be introduced in Section 4. Even with relatively few observations as compared to x-variables, its size never exceeds its nominal level (if it errs somewhat toward conservatism).

Let us suppose, somewhat implausibly, a statistician whose feel for sample size is precisely calibrated to the demands of the asymptotic deviance test for logistic regression. Such a statistician would insist on large sample sizes relative to the number of x-variables, much larger than would those of use whose intuitions are calibrated to regression with continuous outcomes. Confronted with too few observations relative to the number of variables, she'd sooner whittle down the number of variables, surely, than collect more observations for the sole purpose of making herself more comfortable using a favored balance-checking technique. But this arbitrarily reduces the scope of her check; wouldn't it be better to check all the x-variables that seemed important, selecting a technique appropriate to the data, rather than the reverse?

3 Randomization tests of balance, with and without clusters

A common form of frequentism, often traced to Neyman (1990), posits that subjects arrive in a study through random sampling from a broader population, and takes as its goal to articulate how much the sample and population may differ. In Fisher's model of a comparative study, in contrast, no background population is supposed; but randomization is supposed to govern division of the sample into comparison groups, and inference asks only whether chance alone could explain differences between the groups. One need not side with Fisher over Neyman in general, nor even prefer frequentism to Bayesianism, to adopt Fisher's conceptualization when appraising balance.

3.1 Simple randomized experiments

To illustrate, consider the question of whether in the Vote '98 experiment, subjects assigned to receive a telephone reminder, or not, had voted in the previous election in similar proportions. Since past voting is predictive of future voting, differences to the advantage of either group would have prejudged the outcome of the experiment.

Let the index $i = 1, \ldots, n$ run over assignment units, so that z_i indicates the treatment assignment of the ith cluster of observation units. Interpret x_i as the total of x-values for observation units in cluster i, in this case the number of subjects in the household who voted in the previous election, and let m_i be the size of that cluster, here 1 or 2, in observation units. $\mathbf{z}$, $\mathbf{x}$, and $\mathbf{m}$ are n-vectors recording these data for each assignment unit. The observed difference of the proportions of treatment and control group subjects who had cast votes in 1996 can be written as a function $d_p(\mathbf{z}, \mathbf{x})$ of the treatment-group indicator vector $\mathbf{z}$ and indicators $\mathbf{x}$ of voting in the previous election. In symbols, $d_p(\mathbf{z}, \mathbf{x}) = \mathbf{z}^t\mathbf{x}/\mathbf{z}^t\mathbf{m} - (\mathbf{1} - \mathbf{z})^t\mathbf{x}/(\mathbf{1} - \mathbf{z})^t\mathbf{m}$;

for general measurement variables v, $d_p(\mathbf{z}, \mathbf{v})$ is the difference of treatment and control group means. Let A be the set of treatment assignments from which the actual assignment $\mathbf{z}$ was randomly selected; for each member $\mathbf{z}^*$ of A, it is straightforward to compute the amount $d_p(\mathbf{z}^*, \mathbf{x})$ by which treatments and controls would have differed had assignment $\mathbf{z}^*$ been selected. A (two-sided) randomization p-value attaching to the hypothesis of balance on x is

$$\begin{aligned} & \frac{\#\{\mathbf{z}^* \in A : |d_p(\mathbf{z}^*, \mathbf{x})| > |d(\mathbf{z}, \mathbf{x})|\} + \frac{1}{2}\#\{\mathbf{z}^* \in A : |d_p(\mathbf{z}^*, \mathbf{x})| = |d(\mathbf{z}, \mathbf{x})|\}}{\#A} \\ = \quad & \mathbf{P}(|d_p(\mathbf{Z}, \mathbf{x})| > |d(\mathbf{z}, \mathbf{x})|) + \frac{1}{2}\mathbf{P}(|d_p(\mathbf{Z}, \mathbf{x})| = |d(\mathbf{z}, \mathbf{x})|), \end{aligned} \tag{2}$$

where $\mathbf{Z}$ is a random vector that is uniformly distributed on possible treatment assignments A. (Weighting by one-half those $\mathbf{z}^* \in A$ for which $|d_p(\mathbf{z}^*, \mathbf{x})| = |d(\mathbf{z}, \mathbf{x})|$ makes this a mid-p value, the null distribution of which will be more nearly uniform on $[0, 1]$ than would a p-value without this weighting. Agresti and Gottard discuss merits of the mid-p value.) This appraisal of balance on x does involve probability; but note carefully that only treatment assignment, not the covariate, is modeled as stochastic.

Such p-values can in principle be calculated by enumeration; in practice, one approximates them by simulation. Under favorable designs, fast and accurate Normal approximation are also available. Consider first the case in which

(A) the assignment scheme allocates a fixed and predetermined number n_t of the n clusters to treatment, and

(B) each cluster contains the same number $\bar{m}$ of measurement units.

Then the ratios $\mathbf{Z}^t\mathbf{x}/\mathbf{Z}^t\mathbf{m}$ and $(\mathbf{1}-\mathbf{Z})^t\mathbf{x}/(\mathbf{1}-\mathbf{z})^t\mathbf{m}$ of which $d_p(\mathbf{Z}, \mathbf{x})$ is a difference have constants, respectively $\bar{m}n_t$ and $\bar{m}n_c$, as denominators, so that, as in (1), $d_p(\mathbf{Z}, \mathbf{x})$ has an equivalent of the form $k\mathbf{Z}^t\mathbf{x} - \mathbf{1}^t\mathbf{x}/\bar{m}n_c$. Then it is necessary only to approximate the distribution of $\mathbf{Z}^t\mathbf{x}$, an easier task than approximating the distribution of its ratio with another random variable. Indeed, if $\{i \in \{1, \ldots, n\} : Z_i = 1\}$ is a simple random sample of size n_t, then $\mathbf{Z}^t\mathbf{x}$ is simply the sample sum of a simple random sample of n_t from n cluster totals $x_1, \ldots, x_n$. Common results for simple random sampling give that $\mathbf{E}(\mathbf{Z}^t\mathbf{x}) = n_t\bar{x} = \frac{n_t}{n}\sum x_i$; that $\mathrm{Var}(\mathbf{Z}^t\mathbf{x}) = n_t\left(1 - \frac{n_t}{n}\right)s^2(\mathbf{x})$, where $s^2(\mathbf{x}) = \left(\sum_1^n (x_i - \bar{x})^2\right)/(n-1)$; and that if n_t grows to infinity while n_t/n approaches a constant in $(0, 1)$, and mean squares and cubes of $|\mathbf{x}|$ remain bounded, then the limiting distribution of $\mathbf{Z}^t\mathbf{x}$ is Normal (Hájek 1960; Erdős and Rényi 1959). Over and above this finite population central limit theorem, Höglund's Berry-Esseen principle for simple random sampling (1978) limits the error of the Normal approximation in finite samples, suggesting that it should govern $\mathbf{Z}^t\mathbf{y}$ similarly well for covariates y other than x, and that it should be quite good even in samples of moderate size.

Cases in which (A) or (B) fails might appear to frustrate this argument. For instance, suppose treatment were assigned, in violation of (A), by n independent Bernoulli(p) trials. Then there would be some random fluctuation in treatment and control group sizes $\mathbf{Z}^t\mathbf{m}$ and $(\mathbf{1}-\mathbf{Z})^t\mathbf{m}$, and the denominators of the ratios of which $d_p(\mathbf{Z},\mathbf{x})$ is a difference would no longer be constants, so that the argument by which Höglund's Berry-Esseen principle bounded the error of the Normal approximation would no longer be available.

This particular frustration is circumvented by referring observed differences $d_p(\mathbf{z},\mathbf{x})$ to conditional, rather than marginal, distributions of $d_p(\mathbf{Z},\mathbf{x})$. For conditional on $\mathbf{Z}^t\mathbf{1} = \mathbf{z}^t\mathbf{1} = n_t$, condition (A) is restored, and provided (B) also holds the distribution of $d_p(\mathbf{Z},\mathbf{x})$ is close to Normal, with mean and variance as previously indicated. What of departures from (B), *i.e.* clusters that vary in size? Here the representation of $d_p(\mathbf{Z},\mathbf{x})$ as a linear transformation of $\mathbf{Z}^t\mathbf{x}$ need not apply, even after conditioning on the number of clusters selected for treatment, since then the number of treatment-group subjects $\mathbf{Z}^t\mathbf{m}$ may vary between possible assignments. A modification to the standardized difference $d_p(\cdot,\cdot)$ circumvents the problem. Writing m_t for the expected number of measurement units in the treatment group, set

$$\begin{aligned} d(\mathbf{z},\mathbf{x}) &:= \frac{\mathbf{Z}^t\mathbf{x}}{m_t} - \frac{(\mathbf{1}-\mathbf{Z})^t\mathbf{x}}{m-m_t} \qquad (m_t := \mathbf{E}(\mathbf{Z}^t\mathbf{m}),\ m = \mathbf{1}^t\mathbf{m}) \\ &= \bar{m}^{-1}\left[k\mathbf{Z}^t\mathbf{x} - \mathbf{1}^t\mathbf{x}/(n-n_t)\right] \qquad (k^{-1} := n_t(1-n_t/n)) \end{aligned}$$

This is equivalent to a difference of weighted means of cluster averages, rather than cluster totals, of x, provided the clusters are weighted in proportion to the number of measurement units they contain; Kerry and Bland (1998) recommend it for outcome analysis in cluster randomized trials.

In designs with size variation among assignment units, $d(\mathbf{z},\mathbf{x})$ and $d_p(\mathbf{z},\mathbf{x}) = \mathbf{Z}^t\mathbf{x}/\mathbf{Z}^t\mathbf{m} - (\mathbf{1}-\mathbf{Z})^t\mathbf{x}/(m-\mathbf{Z}^t\mathbf{m})$ may differ. The differences will tend to be small, particularly if $\mathbf{m}$, now regarded as a covariate, is well balanced; and of course this balance is expediently measured using $d(\mathbf{z},\mathbf{m})$ and its associated p-value. These considerations recommend $d(\mathbf{z},\mathbf{x})$ as a basic measure of balance on a covariate x.

3.2 Simple randomization within blocks or matched sets

The approach extends to the case of block-randomized designs, and to designs that result from poststratification or matching. Let there be blocks or poststrata $b = 1,\ldots,B$ containing $n_1,\ldots,n_B$ clusters within which simple random samples of $n_{t1},\ldots,n_{tB}$ clusters are selected into the treatment group, for each $b = 1,\ldots,B$; let $\mathbf{Z} = (\mathbf{Z}_1^t,\ldots,\mathbf{Z}_b^t,\ldots,\mathbf{Z}_B^t)^t = (Z_{11},\ldots,Z_{1n_1},\ldots,Z_{b1},\ldots,Z_{bn_b},\ldots,Z_{B1},\ldots,Z_{Bn_B})$ be a vector random variable of which the experimental assignment was a realization, and let $\mathbf{m} = (\mathbf{m}_1^t,\ldots,\mathbf{m}_b^t)^t$ record sizes of clusters in terms of observation units. For each $b = 1,\ldots,B$ let $m_{tb} = \mathbf{E}(\mathbf{Z}_b^t\mathbf{m}_b) = \bar{m}_b n_{tb}$ be

the expected number of observation units in the treatment group. Let $\mathbf{x} = (\mathbf{x}_1^t, \ldots, \mathbf{x}_B^t)^t$ and $\mathbf{v} = (\mathbf{v}_1^t, \ldots, \mathbf{v}_B^t)^t$ be single covariates, either cluster-level measurements or cluster sums of individual measurements.

Measures of balance should be made separately within each block and then aggregated. Within a block b, the (modified) difference of treatment and control group means on $\mathbf{x}$ is simply $\mathbf{z}_b^t\mathbf{x}_b/m_{tb} - (\mathbf{1} - \mathbf{z}_b)^t\mathbf{x}_b/(m - m_{tb})$; one can combine the differences by taking a weighted average of them. Weights may be proportional to the number of subjects in each block, proportional to the number of treatment-group subjects in each block, or selected so as to be optimal under some model; this latter approached is developed below in Section 5. For now, fix positive weights $w_1, \ldots, w_B$ such that $\sum_i w_i = 1$. The w-weighted average of within-block differences on the mean of $\mathbf{x}$ is the same as the difference of weighted average of means on $\mathbf{x}$, where treatment and control subjects in block b are weighted in proportion to w_b/m_{tb} and $w_b/(m - m_{tb})$, respectively.

Considered as a random variable, the block-adjusted, modified difference of treatment and control group means is

$$\begin{aligned} d(\mathbf{Z}, \mathbf{x}) &= \sum_{b=1}^{B} w_b \left[\mathbf{Z}_b^t\mathbf{x}_b/m_{tb} - (\mathbf{1} - \mathbf{Z}_b)^t\mathbf{x}_b/(m - m_{tb})\right] \quad (3) \\ &= \sum_{b=1}^{B} w_b k_b \bar{m}_b^{-1}\mathbf{Z}_b^t\mathbf{x}_b - \sum_{b=1}^{B} w_b \bar{m}_b^{-1}\mathbf{1}^t\mathbf{x}_b/(n_b - n_{tb}), \quad (4) \end{aligned}$$

where $k_b = n_{tb}^{-1} + (n_b - n_{tb})^{-1} = [n_{tb}(1 - n_{tb}/n_b)]^{-1}$. Within block b, $\mathbf{Z}_b^t\mathbf{x}_b$ is the sample total of a simple random sample of size n_{tb} from $(x_{b1}, \ldots, x_{bn})$. It follows that it has mean $n_{tb}/n_b\mathbf{1}^t\mathbf{x}_b = n_{tb}\bar{x}_b$; that its variance is $n_{tb}(1-n_{tb}/n_b)s^2(\mathbf{x}_b)$; and that its covariance with $\mathbf{Z}_b^t\mathbf{v}_b$ is $n_{tb}(1-n_{tb}/n_b)s(\mathbf{x}_b; \mathbf{v}_b)$, for $s(\mathbf{x}_b; \mathbf{v}_b) = (\mathbf{x}_b - \bar{x}_b\mathbf{1})^t(\mathbf{v}_b - \bar{v}_b\mathbf{1})/(n_b - 1)$ and $s^2(\mathbf{x}_b) = s(\mathbf{x}_b, \mathbf{x}_b)$. In virtue of the design, for blocks $b' \neq b$ the treatment group totals $\mathbf{Z}_b^t\mathbf{x}_b$ and $\mathbf{Z}_b^t\mathbf{v}_b$ are independent of $\mathbf{Z}_{b'}^t\mathbf{x}_{b'}$ and $\mathbf{Z}_{b'}^t\mathbf{v}_{b'}$. Together, these facts entail the following description of the first and second moments of $d(\mathbf{Z}, \mathbf{x})$ and $d(\mathbf{Z}, \mathbf{v})$.

Proposition 3.1 *Suppose that within blocks $b = 1, \ldots, B$, simple random samples of n_{tb} from n_b clusters are selected for treatment, with the rest assigned to control. Let $\mathbf{Z}$ indicate*

sample membership and let $\mathbf{x}$ *and* $\mathbf{v}$ *denote covariates. For* $d(\cdot,\cdot)$ *as in (3), one has*

$$\begin{aligned}
\mathbf{E}d(\mathbf{Z},\mathbf{x}) &= \mathbf{E}d(\mathbf{Z},\mathbf{v}) = 0; \\
\mathrm{Var}(d(\mathbf{Z},\mathbf{x})) &= \sum_{b=1}^{B} w_b^2 \frac{k_b}{\bar{m}_b} \frac{s^2(\mathbf{x}_b)}{\bar{m}_b};\ \textit{and} \\
\mathrm{Cov}(d(\mathbf{Z},\mathbf{x}), d(\mathbf{Z},\mathbf{v})) &= \sum_{b=1}^{B} w_b^2 \frac{k_b}{\bar{m}_b} \frac{s(\mathbf{x}_b;\mathbf{v}_b)}{\bar{m}_b}.
\end{aligned}$$

When $d(\mathbf{Z},\mathbf{x})$ can be assumed Normal, Proposition 3.1 permits analysis of its distribution. In fact, relevant central limit theorems do entail its convergence to the Normal distribution as the size of the sample increases, and they suggest that the convergence should be fast and uniform across covariates $\mathbf{x}$, $\mathbf{v}$, ... There are two cases. In the first case, the number of strata tends to infinity, so that each stratum makes an independent contribution to the sum that is $d(\mathbf{Z},\mathbf{x})$. In this case ordinary central limit theory entails that its distribution tends to Normal. Indeed, the ordinary Berry-Esseen lemma limits the difference between the distribution function of $d(\mathbf{Z},\mathbf{x})$ and an appropriate Normal distribution in terms of its variance and its third central moment (Feller 1971, ch. 16), both of which are calculable precisely from the design and from the configuration of $\mathbf{x}$. In the second case, the number of strata is bounded but the size of at least one stratum tends to infinity. Assume that in each growing stratum the proportions of clusters assigned to treatment and to control tend to non-zero constants. Then the contribution $k_b\bar{m}_b^{-1}\mathbf{Z}_b^t\mathbf{x}_b$ from any growing stratum b is a rescaled sum of a simple random sample from $(x_{b1}, x_{b2}, \ldots, x_{bn_b})$ and is governed by the central limit theorem and Berry-Esseen principle for simple random sampling (see § 3.1). Contributions from small strata that do not grow are asymptotically negligible (assuming their weights shrink to a negligible fraction of the weights of growing strata), and it follows that the overall sum of stratum contributions tends to normality.

Although any weighting of blocks is possible, some are more likely to reveal imbalances than others. Section 5 shows weights $w_b^* \propto \bar{m}_b/k_b$ to be optimal in an important sense; it so happens that differences $d(\cdot,\cdot)$ weighted in this way have first and second moments with particularly simple expressions.

Corollary 3.1 *Suppose that within blocks* $b = 1, \ldots, B$*, simple random samples of* n_{tb} *from* n_b *clusters are selected for treatment, with the rest assigned to control. Let* $\mathbf{Z}$ *indicate sample membership and let* $\mathbf{x}$ *and* $\mathbf{v}$ *denote covariates. For* $d(\cdot,\cdot)$ *as in (3), with* $w_b \equiv w_b^* \propto \bar{m}_b/k_b =$

$\bar{m}_b n_{tb}(1 - n_{tb}/n_b)$, *one has*

$$
\begin{aligned}
d(\mathbf{z}, \mathbf{x}) &= \left(\sum \bar{m}_b/k_b\right)^{-1} \left[\sum_{b=1}^{B} \mathbf{Z}_b^t \mathbf{x}_b - \sum_{b=1}^{B} n_{tb}(\mathbf{1}^t \mathbf{x}_b/n_b)\right] && (5)\\
\mathbf{E}d(\mathbf{Z}, \mathbf{x}) &= \mathbf{E}d(\mathbf{Z}, \mathbf{v}) = 0; \\
\mathrm{Var}(d(\mathbf{Z}, \mathbf{x})) &= \left(\sum \bar{m}_b/k_b\right)^{-2} \sum_{b=1}^{B} \frac{\bar{m}_b}{k_b} \frac{s^2(\mathbf{x}_b)}{\bar{m}_b};\ \textit{and} && (6)\\
\mathrm{Cov}(d(\mathbf{Z}, \mathbf{x}), d(\mathbf{Z}, \mathbf{v})) &= \left(\sum \bar{m}_b/k_b\right)^{-2} \sum_{b=1}^{B} \frac{\bar{m}_b}{k_b} \frac{s(\mathbf{x}_b; \mathbf{v}_b)}{\bar{m}_b}.
\end{aligned}
$$

3.3 Accommodating independent assignment by conditioning

Proposition 3.1 assumes simple random sampling of treatment groups within blocks. Were assignments within block b made by independent Bernoulli(p_b) trials, the induced first and second moments of $d(\mathbf{Z}, \mathbf{x})$ — understood as a w_b-weighted sum of terms

$$\frac{\mathbf{Z}_b^t \mathbf{x}_b}{\bar{m} \mathbf{Z}_b^t \mathbf{1}} - \frac{(\mathbf{1} - \mathbf{Z}_b)^t \mathbf{x}_b}{\bar{m}(n_b - \mathbf{Z}_b^t \mathbf{1})},$$

since n_{tb} would no longer be a fixture of the design — would be formally and numerically similar to those of the proposition, as a simple argument shows. $B_b := \mathbf{Z}_b^t \mathbf{1}$ is $\mathrm{Bin}(n_b, p_b)$, independently of $B_{b'} \sim \mathrm{Bin}(n_{b'}, p_{b'})$, $b' \neq b$, and conditionally on $B_b = n_{bt}$ the distribution of $\mathbf{Z}_b^t \mathbf{x}_b$ is that of a sample sum of a simple random sample of size n_t from $\{x_{b1}, \ldots, x_{bn_b}\}$. In general, conditioning on $B_1, \ldots, B_B$ gives $d(\mathbf{Z}, \mathbf{x})$ and $d(\mathbf{Z}, \mathbf{v})$ distributions of the type described in Proposition 3.1. Since they have mean zero under the conditional distribution, their unconditional means vanish as well; and furthermore the unconditional variance of $d(\mathbf{Z}, \mathbf{v})$ is an average of the conditional variances. The conditional variances, $\mathrm{Var}(\mathbf{Z}_b^t \mathbf{x}_b | B_b = n_{tb}) = n_{tb}(1 - n_{tb}/n_b) s^2(\mathbf{x}_b)$, average over $B_b \sim \mathrm{Bin}(n_b, p_b)$ to $n_b p_b (1 - p_b) \sigma^2(\mathbf{x}_b)$, where $\sigma^2(\mathbf{x}) := [(n-1)/n] s^2(\mathbf{x})$. These expressions can be expected to give quite similar results, unless B_b far exceeds or falls short of its expectation. In sum, for typical configurations of $(\mathbf{Z}_1^t \mathbf{1}, \ldots, \mathbf{Z}_B^t \mathbf{1})$, conditional and unconditional distributions of $d(\mathbf{Z}, \mathbf{x})$ will be nearly alike, to second order.

Their differences are that, first, the unconditional distribution may differ from normality more than the conditional one. Each block's contribution to $d(\mathbf{Z}, \mathbf{x})$ differs in distribution from normality to an extent bounded by Höglund's theorem, if $\mathbf{Z}^t \mathbf{1}$ is fixed, but if $\mathbf{Z}^t \mathbf{1}$ is permitted to vary then the difference is limited neither by Höglund's theorem nor by ordinary Berry-Esseen principles.

Second, conditional assessments of $d(\mathbf{Z}, \mathbf{x})$ are immune from disruption by the occurrence of an atypical configuration of $(\mathbf{Z}_1^t \mathbf{1}, \ldots, \mathbf{Z}_B^t \mathbf{1})$. A conditionality argument shows that such

an immunity is desirable. To pose the question of balance as a null hypothesis about the value of a parameter, consider the broader model in which $\mathbf{P}(Z_{bi} = 1)$ is not a constant for all $i = 1, \ldots, n_b$, but instead $\text{logit}\,(\mathbf{P}(Z_{bi})) = \psi_b + \psi_x(x_{bi})$. The null holds that $\psi_x \equiv 0$. The likelihood of the full model, with independent sampling of Z_{bi}'s and possibly nonzero ψ_x, is

$$\begin{aligned} &\prod_b \exp\left\{ (\sum_{i=1}^{n_b} Z_{bi})\psi_b + \sum_{i=1}^{n_b} Z_{bi}\psi_x(x_{bi}) - \sum_{i=1}^{n_b} \log[1 + \exp(\psi_b + \psi_x(x_{bi}))] \right\} \qquad (7) \\ &= \prod_b \exp\left\{ (\sum_{i=1}^{n_b} Z_{bi})\psi_b + \sum_{i=2}^{n_b} (Z_{bi} - Z_{b1}/(n_b - 1))\psi_x(x_{bi}) - \sum_{i=1}^{n_b} \log[1 + \exp(\psi_b + \psi_x(x_{bi}))] \right\}, \end{aligned}$$

which admits an alternate parametrization in terms of the function $\psi_x(\cdot)$ and moment parameters $\eta_b = \mathbf{E}(\sum_i Z_{bi} | \psi_b, \psi_x)$, $b = 1, \ldots, B$. Thus $\mathbf{Z}_b^t \mathbf{1}$ reflects largely on $(\eta_1, \ldots, \eta_B)$, not $\psi_x(\cdot)$, since model parameters $(\eta_1, \ldots, \eta_B)$ and ψ_x vary freely of one another, and the distribution of $\mathbf{Z}_b$ conditional on B_b depends only on $\psi_x(\cdot)$, not on $(\eta_1, \ldots, \eta_B)$ (Barndorff-Nielsen and Cox 1994, p.130, p.40 *ff.*). The statistic $(\mathbf{Z}_1^t \mathbf{1}, \ldots, \mathbf{Z}_B^t \mathbf{1})$ is an approximate ancillary for inference about the functions ψ_x, $b = 1, \ldots, B$, a consideration favoring tests conditional on it over tests based on the marginal distribution. Whether treatment is assigned by independent coin tosses or by sampling without replacement, one is led to the without-replacement model, and to Proposition 3.1 and its associated Normal approximation, for tests of balance.

3.4 An example

The p-values that Table 2 contrasts with mistaken ones are calculated from test statistics $d(\mathbf{z}, \mathbf{x})$ and distributional approximations developed in this section. Although it wasn't mentioned in § 2.1, the Vote '98 experiment involved blocks, since it combined telephone voting reminders with two other GOTV interventions in a factorial design; Table 2 uses the method of Section 3.2, specifically that described in Corollary 3.1, to combine balance measures across the four blocks.

The first row of Table 2 gives results for the test as to whether $\mathbf{z}^t\mathbf{m}$ differed substantially from $\mathbf{E}(\mathbf{Z}^t\mathbf{m})$, in a subsample of 100 clusters and in the full sample of some 23,0000. The z-scores $d(\mathbf{z}, \mathbf{m})/\sqrt{V(d)}$ (not shown in the table) were 1.186 and 0.226 for the sub- and full samples, respectively, which by Normal tables give approximate p-values of .236 and .821. This suggests $\mathbf{z}^t\mathbf{m}$ was relatively quite close to its null expectation, a suggestion that gains further support from simulations, which find the mid-p values to be .211 and .821, respectively. Having confirmed balance on cluster sizes, the next row of the table asks about voting in the previous election. It isn't precisely the same in treatment and control groups, either for the subsample or for the full sample, as indicated by normalized differences of $d(\mathbf{z}, \mathbf{x})/\sqrt{V(d)} = 1.228$ and $-.853$, respectively; but the p-values, .224 and .391, indicate that voting in the previous election is as similar in the two groups as could be expected, and

the Normal approximation locates them with some accuracy, .220 and .394.

4 Simultaneously testing balance on multiple x's

Ordinarily there will be several, perhaps many, x-variables along which balance ought to be checked. Write

$$d^2(\mathbf{z};\mathbf{x}_1,\ldots,\mathbf{x}_k) := [d(\mathbf{z},\mathbf{x}_1),\ldots,d(\mathbf{z},\mathbf{x}_k)] \left\{ \mathrm{Cov}\left(\begin{bmatrix} d(\mathbf{Z},\mathbf{x}_1) \\ \vdots \\ d(\mathbf{Z},\mathbf{x}_k) \end{bmatrix} \right) \right\}^{-} \begin{bmatrix} d(\mathbf{z},\mathbf{x}_1) \\ \vdots \\ d(\mathbf{z},\mathbf{x}_k) \end{bmatrix}, \quad (8)$$

where $\mathrm{Cov}(d(\mathbf{Z},\mathbf{x}_i), d(\mathbf{Z},\mathbf{x}_j))$ is as in Proposition 3.1 and M^- denotes a generalized inverse of M. This test has the desirable properties that: (i), it culminates in a single test statistic and p-value; (ii), it inherits the desirable Normal approximability of the individual appraisals $d(\mathbf{z},\mathbf{x}_1)$,..., $d(\mathbf{z},\mathbf{x}_k)$; and, (iii), it appraises balance not only on $\mathbf{x}_1,\ldots,\mathbf{x}_k$, but also on all linear combinations of them.

Linearity of $d(\mathbf{z},\cdot)$ immediately establishes (iii). Arguments of Sections 3.1 and 3.2 entail that $d(\mathbf{Z},\beta_1\mathbf{x}_1+\cdots+\beta_k\mathbf{x}_k)$, suitably scaled, must be asymptotically $N(0,1)$ provided the $\mathbf{x}_i$'s are suitably regular, whatever be $\beta_1,\ldots,\beta_k$. It follows that the vector $[d(\mathbf{z},\mathbf{x}_1),\ldots,d(\mathbf{z},\mathbf{x}_k)]$ has the multivariate Normal distribution in large samples, showing (ii). Then $d^2(\mathbf{Z};\mathbf{x}_1,\ldots,\mathbf{x}_k)$ is scalar-valued with a large-sample χ^2 distribution on rank $(\mathrm{Cov}\,([d(\mathbf{z},\mathbf{x}_1),\ldots,d(\mathbf{z},\mathbf{x}_k)]))$ degrees of freedom.

The χ^2-approximation seems to work reasonably well even in small samples. Its distribution in one small simulation experiment is graphed in the right panel of Figure 1, while Table 3 summarizes its distribution in another; in both cases it tends somewhat toward conservatism. As a practical tool for the data analyst, it has the important advantage that it stably handles saturation with x-variables; one wouldn't bring about a spurious rejection of the hypothesis of balance by adding to the list of x-variables to be tested. One certainly would decrease the test's power to detect imbalance along individual $\mathbf{x}_i$'s included among covariates tested, but that is to be expected. (An example is given in § 6.2.) This is in important contrast with methods based on regression of $\mathbf{z}$ on $\mathbf{x}$'s; as the left panel of Figure 1 shows, natural tendencies toward overfitting inflate the Type I errors of such tests.

As an omnibus measure of balance, $d^2(\mathbf{z};\mathbf{x}_1,\ldots,\mathbf{x}_k)$ is similar in form and spirit to a statistic suggested by Raab and Butcher (2001), namely a weighted sum of squares of differences of means of cluster means: in present notation $\alpha_1 d(\mathbf{z},\mathbf{x}_1/\mathbf{m})^2+\cdots+\alpha_k d(\mathbf{z},\mathbf{x}_k/\mathbf{m})^2$, where $\alpha_1,\ldots,\alpha_k \geq 0$ sum to 1. The ability of the statistician to decide the relative weightings α of the variables might in some contexts be an advantage, but in others it may be burdensome. In all cases it lends some arbitrariness to the criterion. Also, the criterion directly measures only imbalances in $\mathbf{x}_1$, ..., $\mathbf{x}_k$. In contrast, $d^2(\mathbf{z};\mathbf{x}_1,\ldots,\mathbf{x}_k)$ measures imbalances in linear combinations of $\mathbf{x}_1,\ldots,\mathbf{x}_k$ as much as in these variables themselves, lets the data

drive the weighting scheme, upweighting discrepancies along variables with less variation in general, and has the advantage of ready appraisability against χ^2 tables.

5 Optimizing local power

This section develops and analyzes a statistical model of the *absence* of balance that is appropriate to randomization inference. Casting this model as an alternative to the null hypothesis of balance, tests based on d or d^2 are seen to have greatest power when weights $w_b^* \propto \bar{m}_b/k_b$ are used to combine differences across blocks or matched sets. A related analysis justifies the decision, encoded in the definition (3) of $d(\cdot,\cdot)$, to combine cluster *totals,* rather than cluster means, of individual measurements. Readers not seeking justification on these points may prefer to skip it, as the discussion is more technical than in other sections.

5.1 Optimal weights

Say balance is to be assessed against a canonical model (§ 3.2) with B blocks, perhaps after conditioning as in Section 3.3. What choice of weights $w_1, \ldots, w_B$ maximizes the power of the test for balance? Common results give the answer for models positing that $\mathbf{x}$ is sampled while $\mathbf{z}$ is held fixed. Kalton (1968), for instance, assumes random sampling from $2B$ superpopulations with means $\mu_{t1}, \mu_{c1}, \ldots, \mu_{tB}, \mu_{cB}$. He finds that in order to maximize power against alternatives to the effect that $\mu_{tb} \equiv \mu_{cb} + \delta$, $\delta \neq 0$, blocks' differences of means should be weighted in proportion to the inverse of the estimated variance of those differences. With the simplifying assumption that individuals in the $2B$ superpopulations have a common variance in x, and ignoring intra-cluster correlations and within-block fluctuations in cluster sizes, this leads to $w_b \propto \bar{m}_b/k_b$. To avoid these simplifications, weights might be set in proportion to reciprocals of estimated variances. But such a procedure would seem to add complexity, and to detract from the credibility of assessments of statistical significance, since the sample-to-sample fluctuation it imposes on the weighting scheme is difficult to account for at the stage of analysis (Yudkin and Moher 2001, p.347).

The randomization perspective leads to the same result, but by a cleaner route, avoiding the need to estimate or make assumptions about dispersion in superpopulations. In support of this claim, the present section analyzes the problem of distinguishing unbiased from biased sampling of Z's, rather than differences in superpopulations from which treatment and control x's are supposed to be drawn. The result that $w_b = w_b^* \propto \bar{m}_b/k_b$ is optimal from a randomization perspective is, to my knowledge, new, as is its justification for an analogue of "inverse-variance weighting" that removes estimating or making assumptions about superpopulation variances from the process of deciding weights. Our analysis is asymptotic, assuming increasing sample size; since any non-trivial test would have overwhelming power given a limitless stock of similarly informative observations, we mount an analysis of local power, in which the observations become less informative as sample size increases. This is modeled with x's that cluster increasingly around a single value as their number increases.

Let there be constants $\{x_{(l)bi}\}$, $\{m_{(l)bi}\}$, and random indicator variables $\{Z_{(l)bi}\}$, arranged in triangular arrays the lth rows of which contain $n_{(l)}$ entries, $\mathbf{x}_{(l)}^t = (\mathbf{x}_{(l)1}^t, \ldots, \mathbf{x}_{(l)B(l)}^t)$, $\mathbf{m}_{(l)}^t = (\mathbf{m}_{(l)1}^t, \ldots, \mathbf{m}_{(l)B(l)}^t)$ and $\mathbf{Z}_{(l)}^t = (\mathbf{Z}_{(l)1}^t, \ldots, \mathbf{Z}_{(l)B(l)}^t)$, respectively, where $\mathbf{x}_{(l)b} = (x_{(l)b1}, \ldots, x_{(l)bn_{(l)b}})^t$, $\mathbf{m}_{(l)b} = (m_{(l)b1}, \ldots, m_{(l)bn_{(l)b}})^t$ and $\mathbf{Z}_{(l)b} = (Z_{(l)b1}, \ldots, Z_{(l)bn_{(l)b}})^t$, for some whole numbers $B(l)$ and $n_{(l)1}, \ldots, n_{(l)B(l)}$. Within a given row l, $\mathbf{x}_{(l)b}, \mathbf{m}_{(l)b}$, and $\mathbf{Z}_{(l)b}$ describe cluster totals on a variable x, cluster sizes (averaging to $\bar{m}_{(l)b}$) and treatment assignments within block b, any $b \leq B(l)$. Suppose $1 \leq n_{(l)tb} < n_{(l)b}$ for all l, b, and assume of the random variables $\mathbf{Z}_{(l)b}$ that with probability 1, $\mathbf{Z}_{(l)b}^t \mathbf{1} = n_{(l)tb}$ for each l and $b \leq B(l)$; say vectors $\mathbf{z}_{(l)}$ that lack this property are *excluded*. The null hypothesis asserts that for each l and block $b \leq B(l)$, $\mathbf{P}(Z_{(l)bi} = 1)$ is the same for all indices i. Alternately put, its probability density $P(\mathbf{z}_{(l)})$ vanishes for excluded $\mathbf{z}_{(l)}$ and otherwise is proportional to (7) with $\psi_x \equiv 0$. For alternatives Q to this null, define (for non-excluded $\mathbf{z}_{(l)}$) a likelihood proportional to (7), with bias function $\psi_x(\cdot)$ the same for all l. Assume of this sequence of models that:

A1 $\{m_{(l)bi}\}$ is uniformly bounded, and $\{n_{(l)tb}/n_{(l)b}\}$ is uniformly bounded away from 0 and 1;

A2 weights $w_{(l)b}$ have the property that $w_{(l)b}/w^*_{(l)b}$ is uniformly bounded away from 0 and ∞, where $w^*_{(l)b} \propto \bar{m}_{(l)b} n_{(l)tb}(1 - n_{(l)tb}/n_{(l)b})$ and $\sum_b w^*_{(l)b} = 1$;

A3 for some c, $\sup_{b,i} |x_{(l)bi} - c| \downarrow 0$ and $\sum_{b \leq B(l)} \sum_i (x_{(l)bi} - c)^2$ is $O(1)$ as $l \to \infty$;

A4 ψ_x is differentiable at c, where c is the constant referred to in A3.

A1 limits the divergence of $w^*_{(l)b}$ and other common weighting schemes; should weights $w_{(l)b}$ be proportional to the number of subjects in a block, the number of treatment group subjects in a block, or the total of controls by block, then by A1, $w^*_{(l)b}/w_{(l)b}$ will be universally bounded away from 0 and ∞. In other words, given A1 condition A2 is not restrictive. A1 also ensures that $\sum_b \bar{m}_{(l)b}/k_{(l)b}$ is $O(n_{(l)})$. A3 ensures tightening dispersion of x's around c. In particular, combined with A1, A3 entails $\sum_b w^*_{(l)b} s^2_{(l)b}(\mathbf{x}_{(l)b})/\bar{m}_{(l)b}$ is $O(n^{-1}_{(l)b})$, or that with weighting by either $w^*_{(l)b}$ or $w_{(l)b}$, the weighted average of block mean differences $d(\mathbf{Z}_l, \mathbf{x}_l)$ has variance of order $O(n^{-2}_{(l)b})$ — see Proposition 3.1 and Corollary 3.1.

By narrowing attention, if necessary, to a subsequence of models, let there be positive constants K, s_{0x}, s_{wx} and v_{wx} such that as $l \to \infty$,

$$n_{(l)}^{-1} \sum_b \bar{m}_{(l)b} n_{(l)tb} \left(1 - \frac{n_{(l)tb}}{n_{(l)b}}\right) \to K \quad \text{and} \quad n_{(l)} \sum_b w^*_{(l)b} \frac{s^2(\mathbf{x}_{(l)b})}{\bar{m}_{(l)b}} \to s^2_{0x}; \tag{9}$$

$$n_{(l)} \sum_b w_{(l)b} \frac{s^2(\mathbf{x}_{(l)b})}{\bar{m}_{(l)b}} \to s^2_{wx} \quad \text{and} \quad n^2_{(l)} \mathrm{Var}_P \left(d(\mathbf{Z}_{(l)}, \mathbf{x})\right) \to v^2_{wx}, \tag{10}$$

where $d(\mathbf{Z}_{(l)}, \mathbf{x})$ in (10) is understood in the sense of (11).

Proposition 5.1 *Let*

$$d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}) = \sum_b w_{(l)b} \left[\frac{\mathbf{Z}_{(l)b}^t \mathbf{x}_{(l)b}}{\bar{m}_{(l)b} n_{(l)tb}} - \frac{(\mathbf{1} - \mathbf{Z}_{(l)b})^t \mathbf{x}_{(l)b}}{\bar{m}_{(l)b}(n_{(l)b} - n_{(l)tb})} \right]. \tag{11}$$

Assuming A1–A4, writing P and Q for distributions of $\mathbf{Z}_{(l)}$ under the null and alternative hypotheses, respectively, letting s_{wx}, v_{wx} be as in (10)*, and writing β for $\psi'(c)$,*

$$\mathbf{P}_Q(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}) > z^* \mathrm{Var}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}))^{1/2}) \to 1 - \Phi\left(z^* - \beta \frac{s_{wx}^2}{v_{wx}} \right). \tag{12}$$

Display (12) should be compared to

$$\mathbf{P}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}) > z^* \mathrm{Var}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}))^{1/2}) \to 1 - \Phi(z^*),$$

a statement of the asymptotic normality of $d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)})$ under the null hypothesis. For power against alternatives with $\beta > 0$, the acceptance region will be limited from above at $z_u \mathrm{Var}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}))^{1/2}$, $z_u > 0$; this power is optimized by calibrating the stratum weights $(w_{(l)b})$ so as to maximize $\mathrm{Var}_P\left(d(\mathbf{Z}_{(l)}, \mathbf{x})\right)^{-1/2} \left(\sum_b w_{(l)b} s^2(\mathbf{x}_{(l)b})/\bar{m}_{(l)b}\right)$, the limit of which is $\frac{s_{wx}^2}{v_{wx}}$. To effect this calibration, for the moment fix l and write s_b^2 for $s^2(\mathbf{x}_{(l)b})/\bar{m}_{(l)b}$, $b = 1, \ldots, B$. Recall that $k_b = [n_{tb}(1 - n_{tb}/n_b)]^{-1}$ (§ 3.2). Then

$$\begin{aligned} \frac{\left(\sum_b w_{(l)b} s^2(\mathbf{x}_{(l)b})/\bar{m}_{(l)b}\right)}{\mathrm{Var}_P\left(d(\mathbf{Z}_{(l)}, \mathbf{x})\right)^{1/2}} &= \frac{\left(\sum_b w_b s_b^2\right)}{\left(\sum_b w_b^2 [k_b/\bar{m}_b] s_b^2\right)^{1/2}} \\ &= \frac{\left(w_b \left[k_b/\bar{m}_b\right]^{1/2} s_b) : b \le B\right)^t}{\left\| \left(w_b \left[k_b/\bar{m}_b\right]^{1/2} s_b) : b \le B\right) \right\|_2} \\ &\quad \cdot \left(\left[k_b/\bar{m}_b\right]^{-1/2} s_b : b \le B \right), \end{aligned} \tag{13}$$

where $\| \cdot \|_2$ is the Euclidean norm, $\|\mathbf{x}\|_2 = (\sum_i x_i^2)^{1/2}$. Selecting $(w_b : b = 1, \ldots, B)$ so as to maximize this expression amounts to maximizing the correlation between B-dimensional vectors $(w_b [k_b/\bar{m}_b]^{1/2} s_b : b = 1, \ldots, B)$ and $([k_b/\bar{m}_b]^{-1/2} s_b : b = 1, \ldots, B)$, which is achieved by setting $w_b \propto [k_b/\bar{m}_b]^{-1}$ — that is, by $w_{(l)b} = w_{(l)b}^*$.

Similarly, for power against alternatives with $\beta < 0$ one limits the acceptance region from below at $z_l \mathrm{Var}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}))^{1/2} < 0$. The Q- and P-limits of $\mathbf{P}(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}) < z_l \mathrm{Var}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}))^{1/2})$ are $\Phi(z_l + (-\beta)\frac{s_{wx}^2}{v_{wx}})$ and $\Phi(z_l)$, respectively, and again power is maximized by $w_{(l)b} = w_{(l)b}^* \propto \bar{m}_{(l)b}/k_{(l)b}$.

We establish Proposition 5.1 using principles of contiguity (Le Cam 1960; Hájek and Šidák 1967), which describe the limiting Q-distribution of a test statistic $t(\mathbf{Z})$ in terms of the limiting joint distribution, *under* P, of $(t(\mathbf{Z}), \log \frac{dQ}{dP}(\mathbf{Z}))$. A technical lemma, Lemma 5.1, is

needed, after which contiguity results are invoked to establish Lemma 5.2 (from which the proposition is immediate). Both lemmas are proved in the appendix.

Lemma 5.1 *Under the hypotheses of Proposition 5.1,*

$$\log \frac{dQ}{dP}(\mathbf{Z}) \stackrel{P}{\Rightarrow} N(-\frac{1}{2}\beta^2 K s_{0x}^2, \beta^2 K s_{0x}^2)$$

(where "$\stackrel{P}{\Rightarrow}$" denotes convergence in distribution under P.)

Lemma 5.2 *Under the hypotheses of Proposition 5.1,*

$$\frac{d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)})}{\sqrt{\mathrm{Var}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}))}} \stackrel{Q}{\Rightarrow} N(\beta \frac{s_{wx}^2}{v_{wx}}, 1).$$

5.2 The summary of a cluster by its total of individual measurements

Section 3 recommends comparing means of cluster totals, rather than means of cluster means or of some other summary of a cluster's measurements, in order to better assess treatment and control groups' similarity in the composition of individuals, rather than clusters, that they contain. This choice is also better suited to distinguish unbiased assignment, $\mathbf{Z} \sim P$, from treatment assignment with bias, $\mathbf{Z} \sim Q$, as in Section 5.1.

Under the null model, clusters are assigned without regard to cluster or individual-level characteristics of units, but failures of this model might reflect bias along factors present at either of these levels. To represent this possibility, suppose for this section that the indices $(l)bi$ in (14) refer to individuals rather than clusters. The assignment to treatment by group amounts to an additional set of restrictions on $\mathbf{z}_{(l)}$, to the effect that individuals $(l)bi$ and $(l)bj$ belonging to the same group must have $z_{(l)bi} = z_{(l)bj}$. To formalize these restrictions, for each l and block b let the sequence $(bc_1), \ldots, (bc_{(l)bN_b})$ select the index of one individual from each cluster and let $\mathcal{C}_{(l)b}$ consist of all pairs of form (c_i, j) such that individuals (bc_i) and (bj) belong to the same cluster. Then the likelihood is (suppressing the index (n) and supposing the constant c of A3 and A4 to be 0)

$$Q(\mathbf{z}) \propto \begin{cases} \exp\{\mathbf{z}^t \psi_x(\mathbf{x})\} \approx \exp\{\beta \mathbf{z}^t \mathbf{x}\}, & \text{if } \forall b,\ \sum_i z_{bc_i} = n_{tb} \text{ and } \forall b,\ \forall (c_i, j) \in \mathcal{C}_b,\ z_{bc_i} = z_{bj}; \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

An individual $((l)bi)$'s likelihood contribution $\exp\{\beta z_{(l)bi} x_{(l)bi} - k_{(l)bi}\}$ represents the chance, given possibly biased sampling, of that individual's assignment to treatment in the absence of clustering; the probability of his assignment to treatment in the presence of clustering will depend also on the x's of other members of his cluster. Indeed, if Q_β denotes the

approximation to Q with $\beta \mathbf{z}^t \mathbf{x}$ instead of $\mathbf{z}^t \psi_x(\mathbf{x})$ in its likelihood, then (14) gives

$$Q_\beta((z_{bc_i} : b, i)) \propto \begin{cases} \exp\{\beta \sum_b \sum_i z_{bc_i} t_{bc_i}\}, & \text{if } \forall b, \sum_i z_{bc_i} = n_{tb} \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

where $t_{bc_i} = \sum_{(c_i,j) \in \mathcal{C}_{(l)b}} x_j$ records cluster totals of x. According to the Neyman-Pearson lemma, the most powerful test of null hypothesis P against alternative Q_β is based on the statistic $\sum_b \sum_i z_{bc_i} t_{bc_i}$ — in the notation of preceding sections, $\sum_b \mathbf{z}_b^t \mathbf{x}_b$, which by (5) is equivalent to basing the test on $d(\cdot, \cdot)$ with $w_b = w_b^*$, as recommended in Section 5.1.

6 Applications to study design and analysis

6.1 Whether to stratify, and which stratification is best

Randomization within well-chosen blocks may lead to imbalances on baseline measures of smaller absolute magnitude than unrestricted randomization, and smaller baseline imbalances are preferable for various reasons. Raab and Butcher (2001) sought to avoid imbalances large enough to create noticeable discrepancies between treatment effects estimated with and without covariance adjustment. Such differences might be troubling to the policymakers who were a central audience for their study, even if they fell within estimated standard errors. Yudkin and Moher (2001) worry that designs in which sizable imbalances are possible may sacrifice power.

To reduce this penalty, their ASSIST team elected to randomize clinics within three blocks, consisting of 6, 9, and 6 clinics, rather than to randomly assign treatment to 14 of 21 clinics outright. It remained to be decided which baseline variable to block on. Yudkin and Moher report deciding against blocking on clinic size after finding only weak correlations between clinics' sizes and baseline rates of adequate heart disease assessments; they feared that privileging size in the formation of blocks could have "resulted in imbalance in the main prognostic factor" (2001, p.345). While these correlations are certainly reasonable to consider, it might have been more direct to compare candidate blocking schemes on the basis of the variance in $d(\mathbf{Z}, \cdot)$'s they would entail, preferring those schemes that offer lesser mean-square imbalances on key prognostic variables.

Table 4 offers such a comparison. It emerges that, despite the weak relationship between clinic size and baseline rate of adequate assessment, blocking on size balances the rate of adequate assessment quite well, nearly as well as does blocking on the rate itself. Meanwhile, to balance the baseline variables, rates of treatment with various drugs, that at follow-up would be measured as secondary outcomes, it is much better to block on size. Perhaps the investigators were too quick to reject this option. In any case, the comparison of $\mathrm{Var}(d(\mathbf{Z}, \mathbf{x}))$, from (6), for various blocking schemes and covariates, $\mathbf{x}$, would more directly have informed their decision.

Stratification	Adequately assessed	Treated with aspirin	Treated with hypotensives	Treated with lipid-reducers
None	.46	.46	.46	.46
By rate of adequate assessment	.31	.42	.43	.36
By clinic size	.33	.24	.24	.31

Table 4: Standard deviations of $d(\mathbf{Z}, \mathbf{x})$ under various stratification schemes, expressed as fractions of an s.d. of $\mathbf{x}/\bar{m}$. Both stratification schemes offer distinctly better expected balance than no stratification at all, and stratification on clinic size seems preferable to stratification on clinics' baseline rates of adequate assessment.

6.2 Whether to post-stratify, and whether a given post-stratification suffices

Comparative studies typically present a small number of covariates that *must* be balanced in order for the study to be convincing, along with longer list of variables on which balance would be advantageous. In the ASSIST trial, the short list consists of baseline measures on variables to be used as outcomes; in the Vote '98 experiment, it comprises a "baseline" measure of the outcome, voting in the previous election, along with party membership and demographic data, age and neighborhood, that well predict voting. Were the treatment group significantly older than controls, or more likely to have voted in past elections, then one would suspect significant positive error in unadjusted estimates of the treatment effect — even in the presence of randomization, which limits typical errors, but not all errors.

When discovered after treatments have been applied, the most direct remedy for such an imbalance is post-stratification. If treatments are on the whole older than controls, then compare older treatments only to older controls, and also compare younger subjects only among themselves. There is the possibility that one could introduce imbalances on other variables by subclassifying on age; to assess this, one might apply $d^2(\mathbf{z}; \mathbf{x}_1, \ldots, \mathbf{x}_k)$, where $\mathbf{x}_1, \ldots, \mathbf{x}_k$ make up the short list, to the post-stratified design. Should subclassifying only on age fail to sufficiently reduce $d^2(\mathbf{z}; \mathbf{x}_1, \ldots, \mathbf{x}_k)$, or should there be a more complex pattern of misalignment to begin with, propensity-score methods are a reliable alternative (Rosenbaum and Rubin 1984). Indeed, with the option of propensity score subclassification, there's little reason to restrict one's attention entirely to the short list; one can reasonably hope to relieve gross imbalances on any of a longer list of covariates, as well as marginal imbalances on the most important ones.

Perhaps with this in mind, Imai (2005) suggests checking the Vote'98 data for imbalance twice, once focusing on short-list variables and a second time considering also second-order interactions of them. As discussed by Arceneaux et al. (2004), his logistic-regression based check gives misleading results. Despite this technical impediment, however, the spirit of the suggestion is sound; one might hope the check based on d^2 would perform more reliably. In

fact it does: in 10^6 simulated reassignments of telephone GOTV, the d^2 statistic combining imbalances on all first- and second-order interactions of x-variables exceeded nominal .001, .01, .05, and .10 levels of the $\chi^2(363)$ distribution in .09%, .9%, 4.8%, and 9.7% of trials, respectively. The treatment assignment actually used gives, for the long list, $d^2 = 360.6$, with theoretical and simulation p-values .526 and .527, respectively, and for the short list, $d^2 = 26.6$ on 38 d.f.'s, with p-values .918 and .918, respectively; it is well balanced.

7 Discussion and summary

Clinical trials methodologists note, with some alarm, how few cluster randomized trials explicitly make note of cluster-level assignment and account for it in the analysis (Divine et al. 1992; MacLennan et al. 2003; Isaakidis and Ioannidis 2003). We have seen the need for such an accounting even when indications for it might appear to be at their weakest, with clusters that are small, uniform in size, and numerous. We have also seen that one natural model-based test for balance along covariates, the test based on logistic regression, is prone to spuriously indicate lack of balance when there are too many covariates relative to observations, and that this condition obtains for surprisingly large ratios of observations to the number of covariates.

Tests of balance are important in observational studies that subclassify on the propensity score (Rosenbaum and Rubin 1984). When treatments are administered at the clinic or neighborhood level, it is appropriate to subclassify these clusters, rather than the elements contained in them. Whether or not treatments are given to clusters, a subclassification is adequate to remove bias on observed variables if it passes the balance test that would have been appropriate had treatment assignment been random. When observational studies match (Rosenbaum and Rubin 1985) or finely subclassify (Hansen 2004) on the propensity score, tests of balance that accommodate large numbers of subclasses are needed; the d and d^2 statistics presented here meet this requirement, although a likelihood ratio test after logistic regression would not have (Agresti 2002, § 6.3.4).

Cluster-level randomization is said to confront investigators with "a bewildering array of possible approaches to the data analysis" (Donner and Klar 1994). Randomization inference presents a less cluttered field of options, and has the additional advantages of adaptation specifically to comparative studies and of being non-parametric. With appropriate attention to the form of the test statistic, it is quite possible in the randomization framework to respect the study's design while training attention on differences among individuals. This aim also suggests conditioning strategies appropriate to the problem of assessing covariate balance, and with these strategies special bounds on the error of large-sample approximations become available. The result is a class of test statistics, for balance on a single covariate or for balance on any number of covariates, that can be expediently appraised using Normal approximations; the approximations are quite accurate in small and moderate samples. A

noncommittal analysis of one parametric model suggests values for tuning parameters that completely specify, indeed simplify, the form of the resulting nonparametric tests, ending with a simple proscription that is suitable for general use: assess balance along individual covariates $\mathbf{x}$ using $d(\mathbf{z}, \mathbf{x})$ with optimal weights, as in (5), then fold these comparisons into an overall χ^2-statistic $d^2(\mathbf{z}; \mathbf{x}_1, \ldots, \mathbf{x}_k)$, as in (8).

References

Agresti, A. (2002), *Categorical data analysis*, John Wiley & Sons.

Agresti, A. and Gottard, A. (2005), "Comment: Randomized Confidence Intervals and the Mid-p Approach," *Statistical Science*, 20, 367–371.

Arceneaux, K. T., Gerber, A. S., and Green, D. P. (2004), "Monte Carlo Simulation of the Biases in Misspecified Randomization Checks," Tech. rep., Yale University, Institution for Social and Policy Studies.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994), *Inference and asymptotics*, Chapman & Hall Ltd.

Brazzale, A. R., Davison, A. C., and Reid, N. (2006), *Applied Asymptotics*, Cambridge University Press.

Campbell, M. K., Elbourne, D. R., and Altman, D. G. (2004), "CONSORT statement: extension to cluster randomised trials," *British Medical Journal*, 328, 702–708.

Davison, A. (2003), *Statistical Models*, Cambridge University Press.

Divine, G., Brown, J., and Frazier, L. (1992), "The Unit of Analysis Error in Studies about Physicians' Patient Care Behaviour," *Journal of General Internal Medicine*, 71, 623–9.

Donner, A. and Klar, N. (1994), "Methods for comparing event rates in intervention studies when the unit of allocation is a cluster," *American Journal of Epidemiology*, 140, 279–289.

— (2000), *Design and Analysis of Cluster Randomization Trials in Health Research*, Edward Arnold Publishers Ltd.

Erdős, P. and Rényi, A. (1959), "On the central limit theorem for samples from a finite population," *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 4, 49–61.

Feller, W. (1971), *An introduction to probability theory and its applications. Vol. II.*, Second edition, New York: John Wiley & Sons Inc.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996), "On Design Considerations and Randomization-based Inference for Community Intervention Trials," *Statistics in Medicine*, 15, 1069–1092.

Gerber, A. S. and Green, D. P. (2000), "The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment," *American Political Science Review*, 94, 653–663.

— (2005), "Correction to Gerber and Green (2000), replication of disputed findings, and reply to Imai (2005)," *American Political Science Review*, 99, 301–313.

Hájek, J. (1960), "Limiting distributions in simple random sampling from a finite population," *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5, 361–374.

Hájek, J. and Šidák, Z. (1967), *Theory of rank tests*, New York: Academic Press.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2005), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Tech. rep., Department of Government, Harvard University.

Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, 58, 13–30.

Höglund, T. (1978), "Sampling from a finite population. A remainder term estimate," *Scandinavian Journal of Statistics*, 5, 69–71.

Imai, K. (2005), "Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments," *American Political Science Review*, 99, 283–300.

Isaakidis, P. and Ioannidis, J. P. A. (2003), "Evaluation of cluster randomized controlled trials in sub-Saharan Africa," *American Journal of Epidemiology*, 158, 921–6.

Kalton, G. (1968), "Standardization: A technique to control for extraneous variables," *Applied Statistics*, 17, 118–136.

Kerry, S. M. and Bland, J. M. (1998), "Analysis of a trial randomised in clusters," *British Medical Journal*, 316, 54.

Le Cam, L. (1960), "Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses," *Univ. California Publ. Statist.*, 3, 37–98.

MacLennan, G., Ramsay, C., Mollison, J., Campbell, M., Grimshaw, J., and Thomas, R. (2003), "Room for improvement in the reporting of cluster randomised trials in behaviour change research," *Controlled Clinical Trials*, 24, 69S–70S.

Murray, D. M. (1998), *Design and Analysis of Group-randomized Trials*, Oxford University Press.

Neyman, J. (1990), "On the application of probability theory to agricultural experiments. Essay on principles. Section 9," *Statistical Science*, 5, 463–480, reprint. Transl. by Dabrowska and Speed.

Raab, G. M. and Butcher, I. (2001), "Balance in Cluster Randomized Trials," *Statistics in Medicine*, 20, 351–365.

Rosenbaum, P. R. and Rubin, D. (1984), "Reducing Bias in Observational Studies using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

Yudkin, P. L. and Moher, M. (2001), "Putting theory into practice: a cluster randomized trial with a small number of clusters," *Statistics in Medicine*, 20, 341–349.

A Appendix

A.1 Proof of Lemma 5.1

Without loss of generality, the c named in A3 and A4 is 0. Then one has $\psi_x(x) = \psi'_x(0)x + o(|x|) = \beta x + xE(x)$, where, because of A3, $\max_{b,i} |E(x_{(l)bi})| \downarrow 0$ as $l \uparrow \infty$. Since Q is defined by (7) and P is defined by (7) without the ψ_x term, one can write

$$\begin{aligned} \log \frac{dQ}{dP}(\mathbf{Z}_{(l)b}) &= \beta \sum_b \mathbf{Z}^t_{(l)b}(\mathbf{x}_{(l)b} - \bar{x}_{(l)b}) + \sum_b \mathbf{Z}^t_{(l)b}(\mathbf{x}_{(l)b}E(\mathbf{x}_{(l)b}) - \overline{x_{(l)b}E(x_{(l)b})}) + \kappa_{lP} - \kappa_{lQ} \\ &=: X_l + Y_l - (\kappa_{lQ} - \kappa_{lP}), \end{aligned} \tag{16}$$

for appropriate constants κ_{lP}, κ_{lQ}.

By similar calculations as those justifying Proposition 3.1, $\mathrm{Var}_P(X_l) = \beta^2 \sum_b k^{-1}_{(l)b} s^2(\mathbf{x}_{(l)b}) = \beta^2 \sum_b (\bar{m}_{(l)b}/k_{(l)b}) s^2(\mathbf{x}_{(l)b})/\bar{m}_{(l)b}$. By A3 and (9), this variance approaches $\beta^2 K s^2_{0x}$. By the discussion following (3), $\mathrm{Var}_P(Y_l) = \sum_b n_{(l)tb}(1 - n_{(l)tb}/n_{(l)b}) s^2(\mathbf{x}_{(l)b}E(\mathbf{x}_{(l)b}))$. By A3, this is $O(E_l^2)$ as $l \uparrow \infty$, where $E_l := \sup_{b,i} |E(x_{(l)bi})|$. By A3 and A4, of course, $E_l \downarrow 0$ as $l \uparrow \infty$; thus $\mathrm{Var}_P(Y_l) \downarrow 0$ as $l \uparrow \infty$. Since (as we have seen) $\mathrm{Var}_P(X_l)$ is $O(1)$, it follows also that $\mathrm{Cov}_P(X_l, Y_l) = O(E_l)$, and overall $\mathrm{Var}_P(X_l + Y_l) \to \beta^2 K s^2_{0x}$ as $l \uparrow \infty$.

Clearly both X_l and Y_l have expectation 0, under P. Since the random term $X_l + Y_l$ is, as in Section 3.2, a sum of totals of simple random samples, its limiting law (under P) must be $N(0, \beta^2 K s^2_{0x})$.

It remains to be shown that $\kappa_{lQ} - \kappa_{lP} \to \frac{1}{2}\beta^2 K s^2_{0x}$. Since $\mathbf{E}_P(dQ/dP)(\mathbf{Z}) = 1$, $\exp\{\kappa_{lQ} - \kappa_{lP}\} = \mathbf{E}_P e^{X_l + Y_l}$. From what was just shown it follows immediately that $e^{X_l+Y_l} \stackrel{P}{\Rightarrow} e^{N(0,\beta^2 K s^2_{0x})}$, the expectation of which equals the moment generating function of the standard Normal distribution evaluated at $\beta K^2 s^2_{0x}$, or $\exp\{\frac{1}{2}\beta^2 s^2_{0x}\}$. So the conclusion follows if we can establish that $\mathbf{E}_P e^{X_l+Y_l}$ converges to $\mathbf{E}e^{N(0,\beta^2 K s^2_{0x})}$. This would follows from uniform integrability of the random variables $e^{X_l+Y_l}$, which would follow in turn from $\sup_n \mathbf{E}_P e^{(1+\epsilon)(X_l+Y_l)} < \infty$, any $\epsilon > 0$.

The rest of the argument verifies this by establishing the technical condition that $\limsup_l \mathbf{E}_P \exp\{\sqrt{2}(X_l + Y_l)\} < \infty$. We make use of a theorem of Hoeffding (1963), to the effect that the expectation of a convex continuous function of a sum of a simple random sample is bounded above by the expectation of the same function of a similarly sized with-replacement sample from the same population, and of the fact from calculus that if for a triangular array $\{c_{ij}\}$ of nonnegative numbers, $\max_j c_{ij} \downarrow 0$ while $\sum_j c_{ij} \to \lambda$, then $\prod_j (1+c_{ij}) \to e^\lambda$. Write $m_{(l)b}(t)$ for the moment generating function of $\mathbf{Z}^t_{(l)b}(\psi_x(\mathbf{x}_{(l)b}) - \overline{\psi_x(x)}_{(l)b})$, so that $\mathbf{E}_P e^{t(X_l+Y_l)} = \prod_b m_{(l)b}(t)$. Under P, $\mathbf{Z}^t_{(l)b}(\psi_x(\mathbf{x}_{(l)b}) - \overline{\psi_x(x)}_{(l)b})$ is the sum of a simple random sample of size $n_{(l)tb}$, and by Hoeffding's theorem $m_{(l)b}(t) \leq (\tilde{m}_{(l)b}(t))^{n_{(l)tb}}$, where $\tilde{m}_{(l)b}(t)$ is the moment generating function of a single draw, $D_{(l)b}$, from $\{\psi_x(\mathbf{x}_{(l)bi}) - \overline{\psi_x(x)}_{(l)bi} : i \leq n_{(l)b}\}$. By Taylor approximation, for each l and b, $\tilde{m}_{(l)b}(\sqrt{2}) = 1 + \mathbf{E}_p D^2_{(l)b} \exp\{t^*_{(l)b} D_{(l)b}\}$, some $t^*_{(l)b} \in [0, \sqrt{2}]$. We now need to show that $\max_b \mathbf{E}_p D^2_{(l)b} \exp\{t^*_{(l)b} D_{(l)b}\} \downarrow 0$ and $\sum_b n_{(l)tb} \mathbf{E}_p D^2_{(l)b} \exp\{t^*_{(l)b} D_{(l)b}\}$ is $O(1)$. By A3, as l increases $D^2_{(l)b} \exp\{t^*_{(l)b} D_{(l)b}\}$ is deterministically bounded by constants tending to 0, entailing $\max_b \mathbf{E}_p D^2_{(l)b} \exp\{t^*_{(l)b} D_{(l)b}\} \downarrow 0$. $\exp\{t^*_{(l)b} D_{(l)b}\}$ also declines to 0 deterministically, so that the sum of $n_{(l)tb} \mathbf{E}_p D^2_{(l)b} \exp\{t^*_{(l)b} D_{(l)b}\}$ is $O(1)$ if $\sum_b n_{(l)tb} \mathbf{E}_p D^2_{(l)b} = \sum_b n_{(l)tb} \sigma^2(\psi_x(\mathbf{x}_{(l)b}))$ is. Now $\sum_b n_{(l)tb} \sigma^2(\psi_x(\mathbf{x}_{(l)b})) = \sum_b n_{(l)tb} \beta^2 \sigma^2(\mathbf{x}_{(l)b}) + \sum_b n_{(l)tb} \sigma^2(\psi_x(\mathbf{x}_{(l)b}) - \beta \mathbf{x}_{(l)b}) + 2 \sum_b n_{(l)tb} \beta \sigma(\mathbf{x}_{(l)b}, \psi_x(\mathbf{x}_{(l)b}) - \beta \mathbf{x}_{(l)b})$. Invoking A3, the first of these three sums may be seen to be $O(1)$, and the latter two $O(E_l^2)$ and $O(E_l)$, respectively, as $l \uparrow \infty$. It follows that $\prod_b (\tilde{m}_{(l)b}(\sqrt{2}))^{n_{(l)tb}}$, and hence $\prod_b m_{(l)b}(\sqrt{2})$, are $O(1)$, confirming that $\{e^{X_l + Y_l} : l = 1, \ldots\}$ is uniformly integrable. □

A.2 Proof of Lemma 5.2.

Write $T_l := \mathrm{Var}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}))^{-1/2} d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)})$. By arguments of Section 3.2, $T_l \stackrel{P}{\Rightarrow} N(0,1)$. Combining this with Lemma 5.1, one has that

$$(T_l, \log \frac{dQ}{dP}(\mathbf{Z}_{(l)})) \stackrel{P}{\Rightarrow} N\left[(0, -\sigma^2/2), \begin{pmatrix} 1 & r \\ r & \sigma^2 \end{pmatrix}\right],$$

for some as yet to be determined r. This establishes the premise of Le Cam's Third Lemma (Le Cam 1960; Hájek and Šidák 1967), the conclusion of which is that the limit law *under* Q of the random variable T_l is $N(r, 1)$. We now calculate r.

Using the notation of (16), $\mathrm{Cov}(T_l, \log \frac{dQ}{dP}(\mathbf{Z}_{(l)})) = \mathrm{Cov}(T_l, X_l) + \mathrm{Cov}(T_l, Y_l)$. Now $|\mathrm{Cov}(T_l, Y_l)| \leq (\mathrm{Var}(T_l)\mathrm{Var}(Y_l))^{1/2} = \mathrm{Var}(Y_l)^{1/2}$, which was shown in the proof of Lemma 5.1 to decline to

0 as l increases. Considering only non-excluded treatment assignments $\mathbf{Z}_{(l)}$,

$$\begin{aligned}
\mathrm{Cov}_P(T_n, X_l) &= V^{-1/2}\mathrm{Cov}_P\left(\sum_{b=1}^{B} w_{(l)b}(k_{(l)b}/\bar{m}_{(l)b})\mathbf{Z}^t_{(l)b}\mathbf{x}_{(l)b}, \sum_b \beta\mathbf{Z}^t_{(l)b}\mathbf{x}_{(l)b}\right) \\
&= \beta V^{-1/2}\sum_b w_b(k_{(l)b}/\bar{m}_{(l)b})\mathrm{Var}_P\left(\mathbf{Z}^t_{(l)b}\mathbf{x}_{(l)b}\right) \\
&= \beta V^{-1/2}\sum_b w_b s^2(\mathbf{x}_{(l)b})/\bar{m}_{(l)b},
\end{aligned}$$

writing V for $\mathrm{Var}_P(d(\mathbf{Z}_{(l)}, \mathbf{x}_{(l)}))$, invoking independence of $\mathbf{Z}_b$ and $\mathbf{Z}_{b'}$, $b \neq b'$, and evaluating $\mathrm{Var}_P(\mathbf{Z}^t_b\mathbf{x}_b)$ in the same manner as led to Proposition 3.1. According to (10), then, $\mathrm{Cov}_P(T_l, X_l) \to \beta\frac{s^2_{wx}}{v_{wx}}$. It follows that $r = \beta\frac{s^2_{wx}}{v_{wx}}$.□