OPERATIONS RESEARCH
Vol. 63, No. 4, July–August 2015, pp. 868–876
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2015.1361

METHODS

The Power of Optimization Over Randomization in Designing Experiments Involving Small Samples

Dimitris Bertsimas
Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, dbertsim@mit.edu

Mac Johnson
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, mac.johnson@sloan.mit.edu

Nathan Kallus
Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, kallus@mit.edu

Random assignment, typically seen as the standard in controlled trials, aims to make experimental groups statistically equivalent before treatment. However, with a small sample, which is a practical reality in many disciplines, randomized groups are often too dissimilar to be useful. We propose an approach based on discrete linear optimization to create groups whose discrepancy in their means and variances is several orders of magnitude smaller than with randomization. We provide theoretical and computational evidence that groups created by optimization have exponentially lower discrepancy than those created by randomization and that this allows for more powerful statistical inference.

Subject classifications: experimental design; clinical trials; partitioning problems.
Area of review: Optimization.
History: Received November 2013; revisions received September 2014, December 2014; accepted January 2015.
Published online in *Articles in Advance* April 1, 2015.

1. Introduction

Experimentation on groups of subjects, similar in all ways but for the application of an experimental treatment, is a cornerstone of modern scientific inquiry. In any controlled experiment, the quality, interpretability, and validity of the measurements and inferences drawn depends on the degree to which the groups are similar at the outset.

For close to a century, randomization of subjects into different groups has been relied on to generate statistically equivalent groups. Where group size is large relative to variability, randomization robustly generates groups that are well matched with respect to any statistic. However, when group sizes are small, the expected discrepancy in any covariate under randomization can be surprisingly large, hindering inference. This problem is further aggravated as the number of groups one needs to populate increases.

This is the situation faced in numerous disciplines in which the rarity or expense of subjects makes assembly of large groups impractical. For example, in the field of oncology research, experimental chemotherapy agents are typically tested first in mouse models of cancer, in which tumor-bearing mice are segregated into groups and dosed with experimental compounds. Because these mouse models are laborious and expensive, group size is kept small (typically 8–10), while the number of groups is relatively large, to accommodate comparison of multiple compounds and doses with standard-of-care compounds and untreated control groups. In this case, it is clear that initial tumor weight is highly correlated with post-treatment tumor weight, in which we measure the effect of treatment. A typical experiment might consist of 40–60 mice segregated into four to six groups of ten, though experiments using fewer mice per group and many more groups are performed as well. Given that the implanted tumors grow quite heterogeneously, a coefficient of variation of 50% or more in pre-treatment tumor size is not unusual.

In such circumstances, common in nearly all research using animal models of disease as well as many other endeavors, simple randomization fails to reliably generate statistically equivalent groups, and therefore fails to generate reliable inference. It is clearly more desirable that experiments be conducted with groups that are similar, particularly in mean and variance of relevant baseline covariates. Here we treat the composition of small statistically equivalent groups as a mathematical optimization problem in which the goal is to minimize the maximum difference in both mean and variance between any two groups. We report one treatment of this problem as well as a study of the size of the discrepancy when group enrollment is optimized compared to other common designs including complete randomization.

Block and orthogonal designs (see Fisher 1935) have been a common way to reduce variability when baseline covariates are categorical, but do not apply to mixed (discrete and continuous) covariates, which is the main focus

of our work. For such cases, apart from randomization, two prominent methods are pairwise matching for controlled trials (see Rosenbaum and Rubin 1985, Greevy et al. 2004) and re-randomization as proposed in Morgan and Rubin 2012.[1] The finite selection model (FSM) proposed by Morris (1979) can also be used for this purpose. In comparisons explored in §4, we find that the balance produced by our proposed optimization-based approach greatly improves on both randomization and these methods.

Pairwise matching is most common in observational studies, where assignment to treatment cannot be controlled (see Rubin 1979 and Rosenbaum and Rubin 1983 for a thorough discussion of the application of pairwise matching and other methods to observational studies). A large impediment to existing practices is that they are based on subject pairs. When sample sizes are small and random there will hardly be any well-matched pairs. Such matching does little to eliminate bias in the statistics that measure the overall average effect size. Instead we consider matching the experimental groups to minimize the en-masse discrepancies in means and variances among groups as formulated in (1).

When discrepancy is minimized, statistics such as the mean difference in subject responses are far more precise, and concentrated tightly around their nominal value, while still being unbiased estimates. Indeed, under optimization, these statistics will no longer follow their usual distributions, which are wider, and traditional tests that rely on knowledge of this distribution, like the Student T test, no longer apply. Beyond estimation, we propose a hypothesis test based on the bootstrap to draw inferences on the differences between treatments. Experimental evidence shows that these inferences are much more powerful than is usually possible.

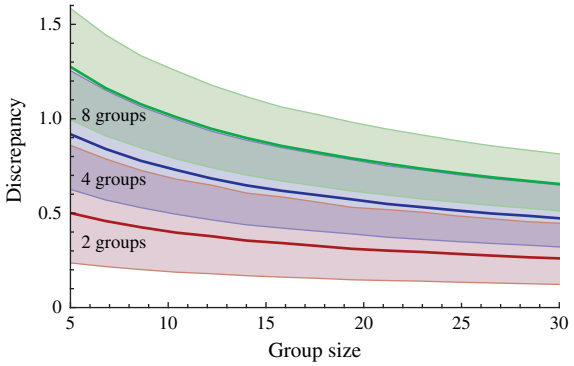
In this paper, we provide theoretical and computational evidence that groups created by optimization have exponentially lower discrepancy in pre-treatment covariates than those created by randomization or by existing matching methods.

2. Limitations of Randomization

Three factors can impair successful matching of the independent variable means of groups assembled using randomization. These are: (a) the group size, (b) the variance of the data, and (c) the number of groups being populated. The specific influence of these three factors is shown graphically in Figure 1. The plot shows the average maximal pairwise discrepancy in means between groups under the conditions indicated for the normal distribution. Average discrepancy is proportional to standard deviation and is therefore reported in units of standard deviations.

It can be seen from the plot that discrepancy increases with the number of groups involved and decreases with increasing group size. When all three factors come into play, i.e., small group size, high standard deviation, and numerous groups, the degree of discrepancy can be substantial. For example, a researcher using randomization to create four groups of ten mice each will be left with an average discrepancy of 0.66 standard deviations between some two of the groups. Because statistical significance is often declared at a mean difference of 1.96 standard deviations ($p \leqslant 0.05$), this introduces enough noise to conceal an effect in comparisons between the mismatched groups or to severely skew the apparent magnitude and statistical significance of a larger effect. Examination of Figure 1 makes it clear that when multiple groups are involved, even apparently large group size can still result in a substantial discrepancy in means between some groups. Doubling the group sizes to twenty each still leaves the researcher with a discrepancy of 0.47 standard deviations.

Figure 1. Average maximal pairwise discrepancy in means among randomly assigned groups of normal variates.



Notes. The vertical axis is in units of standard deviation. The band denotes the average over- and under-shoot: $\mathbb{E}[X \mid X \geqslant \mathbb{E}X]$ and $\mathbb{E}[X \mid X \leqslant \mathbb{E}X]$, where X is maximal pairwise discrepancy.

One solution to this problem is to simply increase group size until discrepancies decrease to acceptable levels. However, the size of the groups needed to do so can be surprisingly large. To reduce the expected discrepancy to below 0.1 standard deviations would require more than 400 subjects per group in the above experiment. For 0.01 standard deviations, more than 40,000 subjects per group would be necessary. With diminishing returns in the reduction of discrepancy with additional subjects, larger increases in the number of subjects enrolled are needed to conduct experiments studying subtler effects.

When considering the effects of this on post-treatment measurements such as mean differences or T statistic, clearly a more precise measurement could be made when groups are well matched at the onset. As we discuss below, well matched groups yield a measurement much closer to the nominal (average or mode) measurement of pure randomization. Indeed, that this distribution of measurements is different (i.e., tighter) means that a naïve application of the Student T test would result in an underestimate of confidence and power, but that the distribution is tighter should allow for much more powerful inference.

3. Optimization Approach

Our proposal is to assign subjects so as to minimize the discrepancies in centered first and second moments, where this assignment is gleaned via integer optimization. After assignment, we randomize which group is given which treatment. This ensures unbiased estimation as discussed in §5.

Given pre-treatment values of subjects w_i, $i = 1, \ldots, n = mk$, we are interested in creating m groups each containing k subjects in such a way that the discrepancy in means and ρ times the discrepancy in second moments is minimized between any two groups. We first preprocess the full sample by normalizing it so that it has zero sample mean and unit sample variance. We set

$$w_i' = (w_i - \hat{\mu})/\hat{\sigma},$$
$$\text{where } \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w_i \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i - \hat{\mu})^2.$$

After construction of k groups, we randomize which treatment is given to which group. Algorithmically, we number the treatments and the groups in any way, shuffle the numbers $1, \ldots, m$ and treat the group in position j with treatment number j. This does not affect the objective value.

The parameter ρ controls the trade-off between the discrepancy of first moments and second moments and is chosen by the researcher. We introduce the decision variable $x_{ip} = 0$ or 1 to denote the assignment of subject i to group p. Using continuous auxilliary variable d and letting

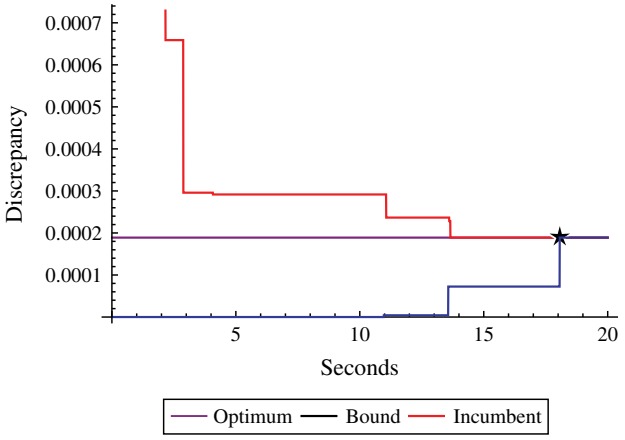
$$\mu_p(x) = \frac{1}{k}\sum_{i=1}^{n} w_i' x_{ip} \quad \text{and} \quad \sigma_p^2(x) = \frac{1}{k}\sum_{i=1}^{n}(w_i')^2 x_{ip},$$

we formulate the problem as follows:

$$\begin{aligned} Z_m^{\text{opt}}(\rho) &= \min_x \ \max_{p\neq q}(|\mu_p(x)-\mu_q(x)|+\rho|\sigma_p^2(x)-\sigma_q^2(x)|) \\ &= \min_{x,d} \ d \qquad (1) \\ &\text{s.t.} \ \ \forall p<q=1,\ldots,m: \\ &\qquad d \geqslant \mu_p(x)-\mu_q(x)+\rho\sigma_p^2(x)-\rho\sigma_q^2(x) \\ &\qquad d \geqslant \mu_p(x)-\mu_q(x)+\rho\sigma_q^2(x)-\rho\sigma_p^2(x) \\ &\qquad d \geqslant \mu_q(x)-\mu_p(x)+\rho\sigma_p^2(x)-\rho\sigma_q^2(x) \\ &\qquad d \geqslant \mu_q(x)-\mu_p(x)+\rho\sigma_q^2(x)-\rho\sigma_p^2(x), \\ &\qquad x_{ip}\in\{0,1\}, \\ &\qquad \sum_{i=1}^{n} x_{ip} = k \quad \forall p=1,\ldots,m, \\ &\qquad \sum_{p=1}^{m} x_{ip} = 1 \quad \forall i=1,\ldots,n, \\ &\qquad x_{ip} = 0 \quad \forall i<p. \end{aligned}$$

As formulated, problem (1) is a mixed integer linear optimization problem with $m(1+2n-m)/2$ binary variables and 1 continuous variable. The last constraint reduces the redundancy in the branch-and-bound tree due to permutation symmetry. Further symmetry reduction is possible by methods described in Kaibel et al. (2011). Symmetry is reintroduced by randomizing which group receives which treatment.

Figure 2. The progress of solving an instance of problem (1) with $n = 40$, $m = 4$.



We implement this optimization model in Gurobi v5.6. For values $n = 40$ and $m = 4$ problem (1) can be solved to full optimality in under twenty seconds on a personal computer with 8 processor cores. Gurobi also has built-in symmetry detection to avoid redundant computations in the branch-and-bound tree. We plot the progress of the branch-and-bound procedure for one example in Figure 2. For larger instances, Gurobi generally finds a near optimal solution with objective value within a few minutes. Finding the optimum can take longer, and proving its optimality even longer.

The formulation of optimization problem (1) extends to multiple covariates. Suppose we are interested in matching the first and second moments in a vector of r covariates where w_{is} denotes the sth covariate of subject i. Again, we normalize the sample to have zero sample mean and identity sample covariance by setting $\mathbf{w}_i' = \Gamma(\mathbf{w}_i - \hat{\mu})$, where Γ is the matrix square root of the (pseudo-)inverse of the sample covariance $\hat{\Sigma} = \sum_{i=1}^{n}(\mathbf{w}_i - \hat{\mu})(\mathbf{w}_i - \hat{\mu})^T/n$. Given the trade-off parameter ρ, we rewrite the optimization problem for this case using $m(1+2n-m)/2$ binary variables and $1+m(m-1)r(r+3)/4$ continuous variables as follows:

$$\begin{aligned} &\min \ d \\ &\text{s.t.} \ \ x\in\{0,1\}^{n\times m}, \ x_{ip} = 0 \quad \forall i<p, \ d \geqslant 0, \\ &\qquad \sum_{i=1}^{n} x_{ip} = k \quad \forall p=1,\ldots,m, \\ &\qquad \sum_{p=1}^{m} x_{ip} = 1 \quad \forall i=1,\ldots,n, \\ &\qquad x_{ip} = 0 \quad \forall i<p, \end{aligned}$$

$$M \in \mathbb{R}^{m(m-1)/2\times r},\ V \in \mathbb{R}^{m(m-1)/2\times r(r+1)/2},$$

$$\forall\, p = 1,\dots,m,\ q = p+1,\dots,m:$$

$$d \geqslant \sum_{s=1}^{r} M_{pqs} + \rho \sum_{s=1}^{r} V_{pqss} + 2\rho \sum_{s=1}^{r} \sum_{s'=s+1}^{r} V_{pqss'},$$

$$\forall\, s = 1,\dots,r:$$

$$M_{pqs} \geqslant \frac{1}{k} \sum_{i=1}^{n} w'_{is}(x_{ip} - x_{iq}),$$

$$M_{pqs} \geqslant \frac{1}{k} \sum_{i=1}^{n} w'_{is}(x_{iq} - x_{ip}),$$

$$\forall\, s = 1,\dots,r,\ s' = s,\dots,r:$$

$$V_{pqss'} \geqslant \frac{1}{k} \sum_{i=1}^{n} w'_{is} w'_{is'}(x_{ip} - x_{iq}),$$

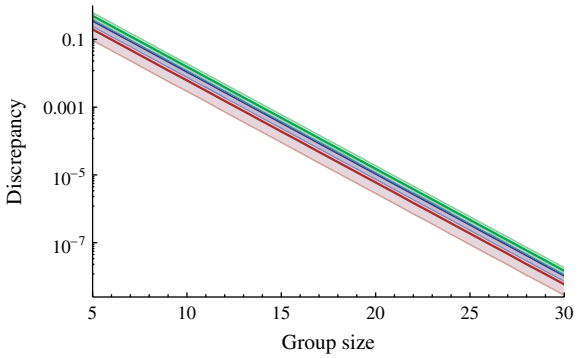
$$V_{pqss'} \geqslant \frac{1}{k} \sum_{i=1}^{n} w'_{is} w'_{is'}(x_{iq} - x_{ip}).$$

The potential extension to even higher moments is straightforward. More generally, such optimization procedures, along with complete randomization and pairwise matching, can all be interpreted under the unifying lens of minimizing worst-case variance; see Kallus (2014).

4. Optimization vs. Randomization in Reducing Discrepancies

Using the above optimization model implemented in Gurobi v5.6, we conducted a series of simulations comparing the results of group assembly using randomization and optimization. Our key finding is that optimization is *starkly* superior to randomization in matching group means under all circumstances tested.

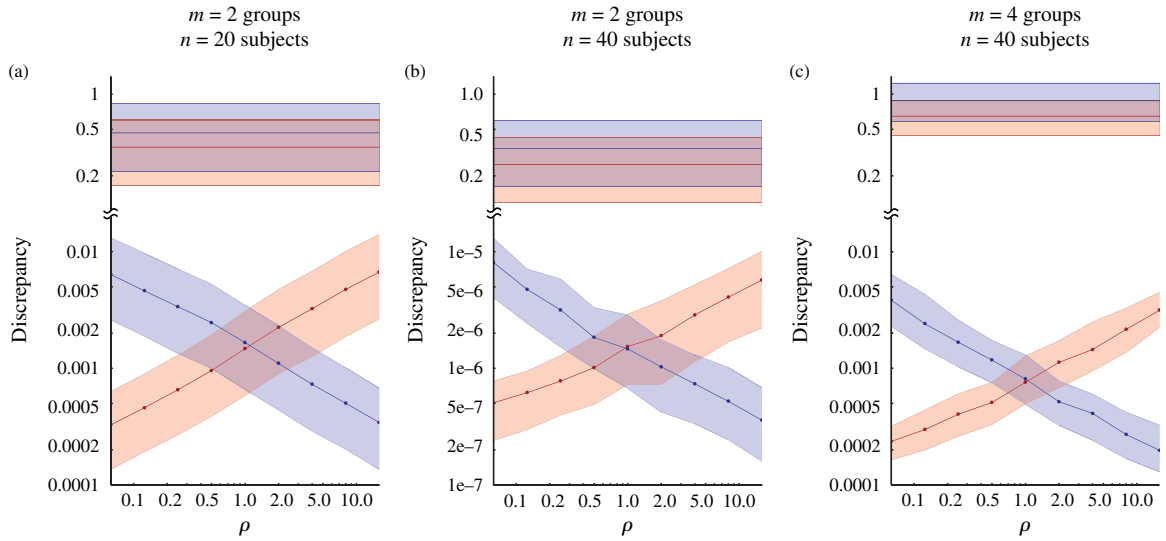
Figure 3. Discrepancy in means among optimally assigned groups of normal variates with $\rho = 0$.



Notes. The colors are as in Figure 1. Note the vertical log scale compared to the absolute scale of Figure 1.

Figure 3 provides the analogue of Figure 1 for optimization and Figure 4 compares side-by-side the mismatch achieved in the first two moments by optimization and randomization. In particular we show for various numbers of groups and group sizes that the achievable range of feasible matchings as ρ varies. For all values of ρ, the pretreatment discrepancy is significantly reduced compared to that seen under randomization, essentially eliminating population variance as a significant source of noise for all but the most extreme circumstances. Noting that discrepancy in either moment is minuscule under optimization using any of the values of ρ shown, we arbitrarily choose $\rho = 0.5$ for all further numerical examples unless otherwise noted. To revisit the example used to illustrate the limitations of randomization, the researcher assembling four groups of 10 mice each under optimization with $\rho = 0.5$ would end

Figure 4. The range of achievable discrepancies under optimization and under randomization.



Notes. The upper halves of the plots correspond to randomization and the lower ones to optimization. Red denotes discrepancy in mean and blue variance. The bands depict average under- and over-shoot. Notice the log scales and the break in the vertical axis.

up with 0.0005 standard deviations of discrepancy in first moment (or a twentieth of that for $\rho = 0$, not shown in figure), compared with 0.66 standard deviations under randomization.

There is some theoretical backing to the experimental evidence that optimization eliminates all discrepancies to such an extreme degree. When $\rho = 0$ and $m = 2$ the problem, scaled by $1/n$, reduces to the well studied balanced number partitioning problem (see Karmarkar and Karp 1982). Let Z_2^{rand} denote the discrepancy in means under randomization. When pre-treatment covariates are random with variance σ^2, we have by Jensen's inequality that

$$\mathbb{E}[Z_2^{\text{rand}}] \leqslant \sqrt{\mathbb{E}[(Z_2^{\text{rand}})^2]} = \sqrt{\frac{2}{k}}\sigma$$

and if they are normally distributed then

$$E[Z_2^{\text{rand}}] = \frac{2}{\sqrt{\pi k}}\sigma.$$

In comparison, an analysis of balanced number partitioning with random weights (see Karmarkar et al. 1986) yields that there is a $C > 0$ such that

$$\text{median}(Z_2^{\text{opt}}(0)) \leqslant \frac{C}{2^{2k}}.$$

Heuristic arguments from spin-glass theory (see Mertens 2001) provide the prediction

$$E[Z_2^{\text{opt}}(0)] = \frac{2\pi\sigma}{2^k},$$

which agrees with our experimental results for large k. Comparing the asymptotic orders of Z_2^{rand} and $Z_2^{\text{opt}}(0)$, we see an *exponential reduction* in discrepancies by optimization versus randomization.

Matching done on a subject-pairwise basis such as caliper matching as done in propensity score matching (see Rubin 1979) does not close this gap either even when the sample-based optimal caliper width is chosen. For simplicity, consider uniformly-distributed pre-treatment covariates so that any subsequent difference of two nearest neighbors are on average $(n+1)^{-1}$. If assignment within each pair is randomized independently, a simple calculation then shows that the average discrepancy is of order $k^{-3/2}$, whereas if assignment is alternating among the sorted covariates, then the average discrepancy is of order k^{-1}. The case is worse for normally distributed covariates as reported below.

Following the average predictions for the normal distribution, if we want to limit discrepancy to some fraction of the standard deviation, $\epsilon\sigma$, we see a dramatic difference in the necessary number of subjects per group, k:

$$k^{\text{Opt}} = \left\lceil \log_2 \frac{2\pi}{\epsilon} \right\rceil, \quad k^{\text{Rand}} = \left\lceil \frac{4}{\pi\epsilon^2} \right\rceil.$$

Table 1. The number of subjects per group needed to guarantee an expected discrepancy no more than $\epsilon\sigma$ for $m = 2$ and $\rho = 0$.

ϵ	k^{Opt}	k^{Rand}	k^{Pair}	k^{RR}
0.1	3	128	9	4
0.01	5	12,833	65	83
0.001	7	1,273,240	514	8,130
0.0001	8	127,323,955	4,354	820,143

In Table 1 we report specific values of k^{Opt} and k^{Rand}, as well as k^{PW} corresponding to optimal pairwise matching and k^{RR} corresponding to the Mahalanobis-distance re-randomization method of Morgan and Rubin (2012) with a fixed acceptance probability of 5%.[2]. This is a clear example of the power of optimization for experiments hindered by small samples. While pairwise matching and re-randomization improve on randomization, they are significantly outperformed by optimization especially when small discrepancy is desired.

A concern may be that by optimizing only the first two moments, and not others, those higher moments may become mismatched. We find, however, that this is not the case even when compared to all the other methods considered above. In Table 2 we tabulate the mismatch in the first five moments and in the generalized moment of log for the various methods when assigning $2k$ subjects with baseline covariates drawn from a standard normal population. In Table 3 we tabulate the mistmatch of multivariate moments for the various methods when assigning $2k$ subjects with multivariate baseline covariates drawn from a three-dimensional standard normal population. For pairwise matching we use the Mahalanobis pairwise distance, for re-randomization we use an acceptance probability of 5%, for FSM we use the method implied by Equation (2.11) of Morris (1979) with $c_i = 1$, $T = I$, and for our method we use $\rho = 0.5$. Note that optimal assignment yields superior balance in the moments considered and that all methods result in similar balance for those moments not directly considered in the optimization problem.

5. Optimization, Randomization, and Bias

Randomization has traditionally been used to address two kinds of bias in experimental design. The first is investigator bias, or the possibility that an investigator may subconsciously or consciously construct experimental groups in a manner that biases toward achieving a particular result. As a fixed, mechanical process, optimization guards against this possibility at least as well as randomization. Indeed it does better because any manual manipulation of the optimized results would make the result less well matched than the reproducible optimum, which is verifiable, whereas no one grouping can ever be verified as the true result of pure randomization.

The second kind of bias is the incidental disproportionate assignment of variables, measured or hidden, that directly

Table 2. The discrepancy in various moments under different assignment mechanisms.

		Moment					
k	Method	1	2	3	4	5	log
5	Opt	0.0513	0.286	1.43	2.67	9.75	0.498
	Rand	0.510	0.689	1.79	3.81	10.3	0.544
	Pair	0.184	0.498	1.27	3.29	8.93	0.345
	Re-rand	0.047	0.711	1.09	3.88	8.47	0.572
	FSM	0.508	0.553	1.76	3.33	10.2	0.440
10	Opt	0.00174	0.0145	0.906	1.47	6.87	0.338
	Rand	0.352	0.504	1.30	2.88	7.79	0.399
	Pair	0.0839	0.259	0.759	2.09	6.06	0.176
	Re-rand	0.0298	0.497	0.764	2.93	6.20	0.389
	FSM	0.374	0.334	1.33	2.26	7.90	0.264
20	Opt	1.23e−6	2.34e−6	0.600	1.04	5.23	0.221
	Rand	0.258	0.345	0.947	2.13	6.13	0.276
	Pair	0.0379	0.140	0.445	1.40	4.24	0.286
	Re-rand	0.0207	0.356	0.565	2.16	4.99	0.284
	FSM	0.249	0.190	0.896	1.50	5.89	0.146

Note. Column i corresponds to the average mismatch in the ith moments between the two groups and the last column corresponds to the mismatch in the generalized moments in $\log|w|$.

affects the treatment. Randomization, given large enough samples, will tend to equalize the apportionment of any one factor. However, just as with the measured covariates w_i, randomization cannot be counted on to eliminate discrepancies in hidden factors when samples are relatively small. Optimization considers the measured covariates w_i when allocating a subject to a particular group. For all factors that are independent with this variable, the allocation remains just as random. Variables that are correlated with the measured covariates in ways such as joint normality will be just as well balanced as the measured covariates and variables with a higher order dependence, such as having a polynomial conditional expectation in w, would be as balanced as seen in Tables 2 and 3.

In general, the observed difference in treatment effects after optimizing the assignment as described here will always be an *unbiased estimator* of the true population average difference, as in a randomized experiment. This is a consequence of randomizing the identity of treatments (while optimizing the partition of subjects) so that the assignment of a single subject is marginally independent of its potential responses to different treatments.[3] Unbiasedness in estimation means that, were the experiment to be repeated many times and the results recorded, the average result would coincide with the true value. In particular, there is no omitted variable bias. That is, neglecting to take into consideration a relevant covariate does not introduce bias in estimation.

Table 3. The discrepancy in various multivariate moments under different assignment mechanisms.

		Moment					
k	Method	w_1	w_1^2	w_1w_2	w_1^3	$w_1^2w_2$	$w_1w_2w_3$
10	Opt	0.0701	0.145	0.183	0.93	0.508	0.337
	Rand	0.360	0.492	0.344	1.29	0.58	0.333
	Pair	0.179	0.383	0.271	0.964	0.478	0.299
	Re-rand	0.141	0.493	0.357	0.883	0.484	0.34
	FSM	0.368	0.606	0.503	1.30	0.574	0.340
15	Opt	0.0230	0.0450	0.117	0.718	0.411	0.292
	Rand	0.292	0.400	0.286	1.05	0.489	0.289
	Pair	0.125	0.290	0.201	0.748	0.38	0.247
	Re-rand	0.113	0.409	0.289	0.714	0.414	0.293
	FSM	0.289	0.597	0.491	1.05	0.488	0.281
25	Opt	0.00302	0.00497	0.0780	0.547	0.315	0.227
	Rand	0.226	0.325	0.222	0.842	0.384	0.227
	Pair	0.0849	0.196	0.143	0.547	0.276	0.172
	Re-rand	0.0863	0.326	0.230	0.566	0.314	0.220
	FSM	0.219	0.592	0.494	0.823	0.388	0.224

Note. Column w_1w_2, for example, corresponds to the average mismatch in the moments of w_1w_2 between the two groups, which by symmetry is the same as that of w_1w_3 or w_2w_3 on average.

6. Optimization vs. Randomization in Making a Conclusion

As we have shown in the previous sections, optimization eliminates nearly all noise due to pre-treatment covariates. One would then expect that it can also offer superior precision in estimating the differences between treatments and superior power in making statistical inferences on these differences.

In randomized trials, randomization tests (see Eddington and Onghena 2007) can be used to draw inferences based directly on the randomness of assignment without normality assumptions, which often fail for small samples. However, for optimization the assignment is not random enough and this test is not applicable. For the purpose of testing differences of treatments in an optimized trial, we propose the following test based on the bootstrap (see Efron and Tibshirani 1993).

Comparing two treatments, we would like to test the null hypothesis that every subject $i = 1, \dots, n$ would have had the same response to treatment whether either of the two treatments were assigned (this is known as the sharp null hypothesis; see Rubin 1980). Let v_i denote the response measured for subject i after it was administered the treatment to which it was assigned. Given subjects with covariates $w_1, \dots, w_n$, the test we propose is as follows:

1. Find an optimal assignment of these to two groups (permuting randomly):

$$\{i_1, \dots, i_{n/2}\} \quad \text{and} \quad \{i_{n/2+1}, \dots, i_n\}.$$

2. Administer treatments and measure responses v_i, which are henceforth fixed.
3. Compute

$$\delta = \frac{1}{k}(v_{i_1} + \cdots + v_{i_{n/2}}) - \frac{1}{k}(v_{i_{n/2+1}} + \cdots + v_{i_n}).$$

4. For $b = 1, \dots, B$:
   (a) Draw a random sample with replacement $w_{b,1}, \dots, w_{b,n}$ from $w_1, \dots, w_n$.
   (b) Find an optimal assignment of these to two groups (permuting randomly):

$$\{i_{b,1}, \dots, i_{b,n/2}\} \quad \text{and} \quad \{i_{b,n/2+1}, \dots, i_{b,n}\}.$$

   (c) Compute

$$\delta_b = \frac{1}{k}(v_{i_{b,1}} + \cdots + v_{i_{b,,n/2}}) - \frac{1}{k}(v_{i_{b,n/2+1}} + \cdots + v_{i_{b,n}}).$$

5. Compute the p-value

$$p = \frac{1}{1+B}\left(1 + \sum_{b=1}^{B} \mathbb{I}[|\delta_b| \geqslant |\delta|]\right).$$

Then, to test our null hypothesis at a significance of α, we only reject it if $p \leqslant \alpha$. The quantity δ above constitutes our estimate of the difference between the two treatments.

To examine the effect of optimization on making a conclusion about the treatments, we again consider the example of a murine tumor study. We consider two groups, each of k mice, with tumor weights initially normally distributed with mean 200 mg and standard deviation 300 mg (truncated to be nonnegative). Two treatments are considered: a placebo and a proposed treatment. Their effect on the tumor, allowed to grow for a period of a day, is of interest to the study.

The effects of treatment and placebo are unknown and are to be inferred from the experiment. We consider a hidden reality where the growth of the tumors is dictated by the Gomp-ex model of tumor growth (see Wheldon 1988). That is, growth is governed by the differential equation:

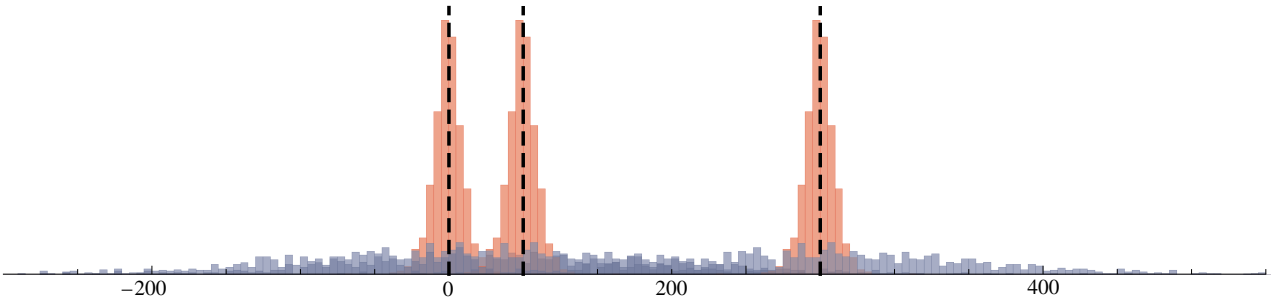
$$\frac{dw}{dt} = w(t)\big(a + \max\{0, b\log(w_c/w(t))\}\big),$$

where a and b are rate parameters and w_c is the critical weight that marks the change between exponential and logistic growth. We arbitrarily choose $a = 1(1/\text{day})$, $b = 5(1/\text{day})$, $w_c = 400$ mg, and $t = 1$ day. We pretend that tumors under either treatment grow according to this equation, but subtract δ_0 from the final weights for the proposed treatment. We consider δ_0 being 0 mg (no effect), 50 mg (small effect), and 250 mg (large effect).

For various values of k and for several draws of initial weights, we consider assignments produced by randomization, our optimization approach ($\rho = 0.5$), pairwise matching, and re-randomization. We consider both the post-treatment estimate of the effect and the inference drawn on it at a significance of $\alpha = 0.05$, using our bootstrap test for our method and the standard randomization test for the others.[4]. In Figure 5 we plot the resulting estimates for $k = 20$ and in Figure 6 we plot the rates at which the null hypothesis is rejected. When there is no effect, this rate should be no more than the significance $\alpha = 0.05$.[5] When there is an effect, we would want the rate to be as close to 1 as possible. In a sense, the complement of this rate is the fraction of experiments squandered in pursuit of an effective drug. The cost-saving benefits of optimization in this case are clear.

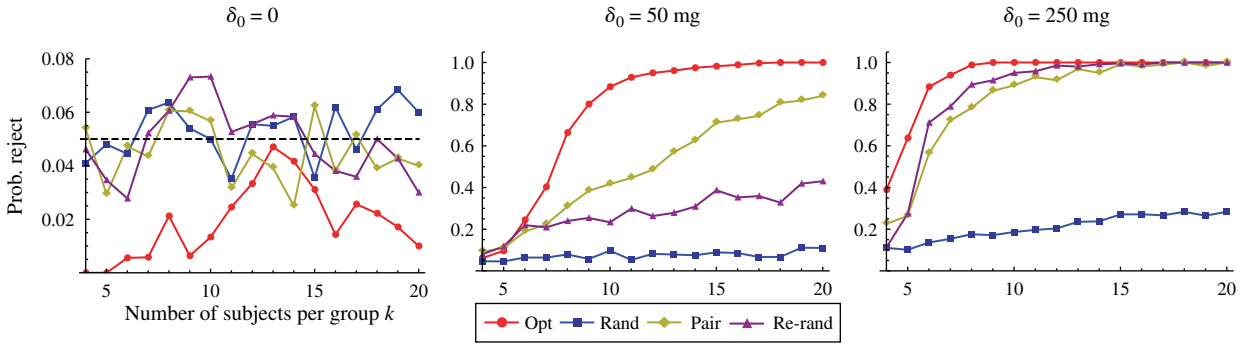
The exact improvements in precision and power depend on the nature of treatment effect. However, comparisons to existing methods are possible. Morgan and Rubin (2012) study reduction in variance due to re-randomization only under the additive treatment model, a very restrictive assumption. In this setting, when setting $\rho = 0$, the same analysis as provided in their Theorem 3.2 provides that the reduction in variance provided by en-masse optimization is exponentially better because the reduction in mean mismatch is exponentially better. Nonetheless, treatment effects usually do depend, albeit perhaps to a lesser extent, on higher orders of the covariates and on their interactions.

Figure 5. The distribution of estimates of effect size under optimization (red) and randomization (blue) for $k = 20$ and effect sizes 0 mg, 50 mg, and 250 mg (dashed lines).



Note. The overlap of estimates under randomization of the nonzero effects and of the zero effect elucidate the low statistical power of randomization in detecting the nonzero effects.

Figure 6. The probability of rejecting the null hypothesis of no effect for various effect sizes.



In Tables 2 and 3 we saw that optimization balances higher and interaction moments no worse than other methods (better for second moments).

7. Practical Significance

Here we present evidence that optimization produces groups that are far more similar in mean and variance than those created by randomization, especially in situations in which group size is small, data variability is large, and numerous groups are needed for a single experiment. For each additional subject per group, optimization roughly halves the discrepancy in the covariate, whereas both randomization and subject-pair matchings offer quickly diminishing reductions. Making groups similar before treatment allows for statistical power beyond what can normally be hoped for with small samples.

We propose that optimization protects against experimental biases at least as well as randomization and that the advantage of optimized groups over randomized groups is substantial. We believe that optimization of experimental group composition, implementable on commonplace software such as Microsoft Excel and on commercial mathematical optimization software, is a practical and desirable alternative to randomization; it can improve experimental power in numerous fields, such as cancer research, neurobiology, immunology, investment analysis, market research, behavioral research, proof-of-concept clinical trials, and others.

Acknowledgments

The authors thank the associate editor and the reviewers for prudent suggestions and thoughtful comments that helped improve the paper. This material is based on work supported by the National Science Foundation Graduate Research Fellowship [Grant 1122374].

Endnotes

1. The work of Morgan and Rubin (2012) can be seen as formalizing and reinterpreting the common informal practice of cherry-picking from several randomizations as a principled heuristic method for matching.
2. Simulation is used to glean k^{opt} for these values of ϵ, for which the asymptotic predictions yield overestimates. Simulation also shows that for FSM, $k^{\text{FSM}} \approx k^{\text{Rand}}$.
3. The correctness of modeling using potential outcomes is contingent on the stable unit treatment value assumption. See Rubin (1986).
4. For non-completely-randomized designs, the randomization test draws random re-assignments according to the method used at the onset. See Eddington and Onghena (2007, Chapter 10).
5. The fact that for our bootstrap test this rate is below 0.05 may be an indication that the test is conservative, i.e., more significant than designed. Nonetheless, despite such conservatism, the test is still more powerful than the other tests.

References

Eddington SE, Onghena P (2007) *Randomization Tests*, Fourth ed. (CRC Press, Boca Raton, FL).

Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap* (CRC Press, Boca Raton, FL).

Fisher A (1935) *The Design of Experiments* (Oliver and Boyd, Edinburgh, UK).

Greevy R, Lu B, Silber JH, Rosenbaum PR (2004) Optimal multivariate matching before randomization. *Biostatistics* 5(2):263–275.

Kaibel V, Peinhardt M, Pfetsch M (2011) Orbitopal fixing. *Discrete Optim.* 8(4):595–610.

Kallus N (2014) Optimal A Priori Balance in the Design of Controlled Experiments. Accessed October 14, 2014, http://arxiv.org/abs/1312.0531.

Karmarkar N, Karp R (1982) Differencing Method of Set Partitioning. Technical Report CSD-83-113, University of California, Berkeley, CA.

Karmarkar N, Karp R, Lueker G, Odlyzko A (1986) Probabilistic analysis of optimum partitioning. *J. Appl. Probab.* 23(3):626–645.

Mertens S (2001) A physicist's approach to number partitioning. *Theoret. Comput. Sci.* 265(1):79–108.

Morris C (1979) A finite selection model for experimental design of the health insurance study. *J. Econom.* 11(1):43–61.

Morgan K, Rubin D (2012) Rerandomization to improve covariate balance in experiments. *Ann. Statist.* 40(2):1263–1282.

Rubin D (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* 74(366):318–328.

Rubin D (1980) Comment: Randomization analysis of experimental data: The Fisher randomization test. *J. Amer. Statist. Assoc.* 75(371): 591–593.

Rubin D (1986) Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.* 81(396):961–962.

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

Rosenbaum PR, Rubin D (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statistician* 39(1):33–38.

Wheldon TE (1988) *Mathematical Models in Cancer Research* (Adam Hilger, Bristol, UK).

Dimitris Bertsimas is the Boeing Professor of Operations Research, Codirector of the Operations Research Center at Massachusetts Institute of Technology, and a member of the National Academy of Engineering. He works on optimization, statistics, stochastics, and their applications. The present paper is part of the authors research in the interface of optimization and statistics.

Mac Johnson leads a cancer drug discovery project at Vertex Pharmaceuticals, Inc., where he is also Head of the Imaging Sciences Group. The present paper grew from an Independent Studies project conducted while he was an EMBA student at MIT's Sloan School of Management.

Nathan Kallus is a Ph.D. candidate in operations research at the Massachusetts Institute of Technology. His interests lie at the intersection of statistics, machine learning, and data science with optimization, operations, and management. The present paper is part of his research into the ramifications of modern optimization for statistical practices.