Random allocations in comparative experiments of known size: balance without blocking

Michael A. Newton [1]
Department of Biostatistics
& Medical Informatics
University of Wisconsin–Madison
600 Highland Avenue
Madison WI 53792-4675
Phone: 608 263 0357
Email: newton@stat.wisc.edu
http: //www.stat.wisc.edu/

Bret Larget
Department of Mathematics and
Computer Science
Duquesne University
440 College Hall
Pittsburgh PA 15282
Phone: 412-396-6469
Email: larget@mathcs.duq.edu
http://www.mathcs.duq.edu/

Rick Chappell
Department Biostatistics
& Medical Informatics
University of Wisconsin-Madison
600 Highland Avenue
Madison, WI 53792-4675
Phone: 608 263 5572
Email: chappell@stat.wisc.edu

Russell F. Jacoby
Department of Medicine
University of Wisconsin-Madison
600 Highland Avenue
Madison, WI 53792-2454
Phone: 608 262 6743
Email: rfjacoby@facstaff.wisc.edu

Issued December, 1997. UW Madison Department of Biostatistics and Medical Informatics Technical Report 124.

[1] To whom correspondence should be addressed.

Random allocations in comparative experiments of known size: balance without blocking

Abstract

We consider the problem of randomizing a known number of subjects into two or more treatment groups when recruitment occurs over an extended time period, and thus the potential for confounding factors related to time is a concern. Our proposed design generates more balanced allocations than complete randomization, but can have higher entropy than randomized-block designs. These conclusions follow a careful analysis of the probability distribution induced on the allocations. Furthermore, analysis and simulation indicate that this new design can exhibit power advantages over randomized blocking and complete randomization. Randomization tests have been applied in a series of cancer chemoprevention trials. The design may be useful in other experiments where experimental units are confined to some linear order, such as arrangement in space.

Keywords: clinical trials, experimental design, one-way layout, permutation test, restricted randomization, relay randomization, urn sampling.

1 RELAY RANDOMIZATION

Il cimento dell'armonia e dell'inventione
- A. Vivaldi, 1725

The inscription given above, which translates to "The contest between harmony and invention," was chosen by Antonio Vivaldi to title his masterful Opus 8, 12 concerti for chamber orchestra and violin. It captures a tension between harmonic requirements of the rules of music on the one hand and originality of a composition on the other. Similar tension characterizes the rather more mundane task of randomizing subjects in clinical trials or other experiments. In order to avoid systematic errors, these groups should be "harmonious," that is balanced in composition with respect to important features such as the order of the subjects' appearance. One extreme design is to allocate subjects deterministically, for example by cycling through the treatment groups in some fixed order. This however would violate the need for probabilistic "invention," or random assignment, which assures that bias due to possibly unknown confounding factors is minimized. Of course randomization also provides a framework for statistical inference (see Fisher, 1971; Cochran and Cox, 1992). The present article offers a new randomization method which effectively protects against time-related confounding factors in a one-way layout.

Operationally, we can implement a "complete" randomization of $n = 2m$ subjects into two groups of size m by sampling without replacement from an urn containing m green balls and m red balls, for example. The color determines the treatment group into which the subject is allocated. Of course the presence of more than two groups creates no problems, as we simply put balls of more than two colors into the urn in the first place. A chemoprevention experiment involving the authors called for $n = 56$ mice to be allocated randomly into $a = 7$ groups of size $m = 8$ (Jacoby *et al.*, 1996). The treatment groups corresponded to dietary dose levels of two compounds thought to modify the growth of intestinal tumours. The mice carried a specific genetic defect which leads to spontaneous intestinal tumour growth and, as they came from a relatively small breeding colony, it would take several months to recruit $n = 56$. The principal investigator was concerned that one or another of the treatment arms would fill up too quickly if complete randomization was used. Indeed, in that case, it would be difficult to interpret the experimental results if some factor associated with time affected the number of tumours, the primary endpoint. For instance, if the control group filled up after one month, and if one of the treatment groups had animals from the second month only, then an observed significant difference could be attributed to altered experimental conditions, perhaps, or to any factor confounded with the treatment factor. The purpose of experimental design, of course, is to anticipate the major factors affecting the response, and to guard against unanticipated factors by some form of randomization. It seems likely that after data are collected, one may always identify *some* factor confounded with the treatment factor (i.e., we speak in correlations not causes!). The question becomes how to best formulate the design so that we retain high power to detect treatment differences while at the same time accounting for potential time-related effects.

One obvious solution is to invoke a simple randomized-block design, in which complete randomization is applied independently to blocks of $a = 7$ mice at a time. The drawback of randomized blocking is that the vector encoding allocations to groups has a severely limited range, and this results in a loss in power if time-related effects are not present (Section 4). The calculation reported in Jacoby *et al.* (1996) introduced a more flexible method which we refer to as relay randomization. Recall again that complete randomization uses an urn with $m = 8$ balls of each of $a = 7$ different colors, and sampling proceeds without replacement. In relay randomization, the contents of this urn are duplicated, so there are $2m$ balls of each color ($2n$ balls total). Subjects present themselves sequentially for allocation, as before. The first subject samples repeatedly, without replacement, until two balls of a common color have been collected. The color of this pair indicates the treatment group into which the subject is allocated. In the process of sampling, the first subject probably will have accumulated a number of other balls, all of different colors, and certainly none matching the color corresponding to the first allocation. These extra balls are passed to the next subject in line, i.e., "relayed," before that second subject begins sampling the urn. Next, the second subject samples the urn, again without replacement, until, in the *combined set* of relayed and newly sampled balls, two balls of one color appear. As before, this color indicates the treatment group of the second subject, who then relays unused balls to the third subject.

Sampling and relaying continue until all subjects are allocated. The urn, which started with two balls for each subject, is emptied because each subject uses two balls to become allocated.

Relay randomization starts by duplicating the contents of an urn set up for complete randomization. There is nothing special about two copies, and indeed a whole family of designs manifests itself if this number is a design parameter, b say. For example, $b = 10$ requires that a subject continues sampling until 10 balls of one color have accumulated. After an allocation, the number of relayed balls is random and tends to become ever larger, with larger b. Note that $b = 1$ defines complete randomization. Figure 1 illustrates one realization of relay randomization in the case where $n = 12$ subjects are allocated into $a = 3$ groups of size $m = 4$, and we use $b = 2$.

Two features of relay randomization are apparent. First, a tendency for balance throughout the experiment is built into the design when $b > 1$. As the allocations proceed, an under-represented group probably gets a head start for the next allocation since the relay is likely to contain balls for that group. The design encourages balance but does not force it, at least until the last subjects are allocated. One expects that the filling up of one treatment group long before another is rather more unlikely than it would be in complete randomization. We show in Section 2 that relay randomization for large b is the same as the simple randomized-block design except for the presence of dependence among blocks. A second feature of this new design is that every balanced allocation of subjects to treatment groups is possible *a priori*. There are no hard restrictions such as those imposed by randomized-block designs. In Section 3 we quantify various aspects of "harmony" and "invention" implicit in relay randomization.

Relay randomization is not the only method that encourages balance in the allocation of subjects to treatment groups. Efron (1971) suggested that subjects be allocated by independent coin tosses in which assignment probability is biased against the group currently holding more subjects. A fair coin determines the allocation if the groups are balanced, and in the event of imbalance only the direction of the imbalance is important. Wei (1977) introduced a more flexible class of designs in which the bias of the coin depends on the magnitude of the imbalance. These designs, extensions thereof, and associated inference have been studied extensively (e.g., Kalish and Begg, 1985). A critical distinction between these designs and relay randomization is that relay randomization considers allocation of a fixed number of subjects and balance is forced at the end of the study. This suggests that complete randomization and randomized-block designs are the natural designs for comparison. There are many variations of blocking (e.g., Cochran and Cox, 1992), but we restrict attention to simple randomized blocking described above.

Having invoked randomization in the experimental design, we have a natural scale on which to assess the significance of observed differences among the treatment groups. That the analysis should reflect the design is a cornerstone of Fisher's theory of significance testing (Fisher, 1971, page 41). As an example, Table 1 records tumour counts resulting from the aforementioned chemoprevention trial. The null hypothesis of no treatment effect can be tested exactly, conditional on the observed counts, using the randomization distribution of the one-way $F-$statistic. Figure 2 shows a histogram of 4999 samples from this distribution, where the randomness comes by repeatedly reallocating mice by relay randomization ($b = 2$), and noting that their tumour counts remain fixed under the null hypothesis. Section 4 presents an analysis of the power of relay randomization, both in the absence and presence of time-related confounding factors. Relay randomization may be nearly as powerful as complete randomization when there is no confounding, but can be more powerful than both complete randomization and simple randomized blocking in the presence of confounding factors. Thus, our initial investigation has revealed cases where relay randomization is optimal in some sense, although the magnitude of the improvements appears to be small.

In Section 5 we elaborate briefly on a with-replacement version of relay randomization. There are fascinating structures here and connections to the earliest discussions of probability.

2 ALLOCATIONS AND THEIR PROBABILITY

A thorough analysis of relay randomization requires that we consider the probability distribution over allocations. We start our discussion with the case of $a = 2$ treatment groups, labeled **g** and **r** for green and red. When $b = 2$, two balls of some color must be accumulated to allocate one subject and the allocation of all subjects $\{1, 2, \ldots, n\}$ is determined by a "primary" sequence of $2n = N$ balls. The sequence **ggrgrrrgrggr** is one of the $\binom{2n}{n} = 924$ possible primary sequences when $n = 6$ subjects are being assigned to two groups of size $m = 3$, for example. Working from left to right and following the allocate-relay rule, we obtain the allocation sequence **grrggr**.

Each primary sequence of $N = nb$ balls may be identified with a monotone non-decreasing path from $(0, 0)$ to $(N/2, N/2)$ in the integer lattice, as shown in the left panel of Figure 3. The sampling of **g** or **r** in the primary sequence corresponds to a horizontal or vertical move, respectively. For example, the shaded path in this panel is the primary sequence **ggrgrrrgrggr** introduced earlier. Furthermore, the induced allocation sequence may be read by following the path formed by this primary sequence. The m vertical and m horizontal heavy lines in the left panel of Figure 3 appear at every bth grid line. Moving from the lower left, there is an allocation each time the path reaches one of the $n = 2m$ heavy lines.

Relay randomization is equivalent to sampling one of the paths uniformly at random from the $(N/2 \times N/2)$ lattice, and then applying the allocate-relay rule to realize an allocation sequence. The probability of an allocation sequence is proportional to the number of paths that would yield this sequence. This number is readily obtained by a simple path-counting algorithm, as demonstrated in the right panel of Figure 3. Starting from the lower left corner, we determine at each lattice point the number of different paths from the start to that point consistent with the allocation by adding the counts of its leftward and downward neighbors. Thus, the allocation **grrggr** has probability $55/924$, somewhat larger than $1/20$, its probability under complete randomization. Table 2 records ten of the twenty possible allocations and their probabilities under relay randomization with $b = 2$. The remaining ten cases may be inferred by exchanging **g** with **r** and noting symmetry. Four of the displayed allocations are blocked, as both **g** and **r** appear in subsequent pairs, and these all have higher probability than the other allocations. The tendency for balance when b is greater than 1 is reflected in these results.

Relay randomization becomes quite different from complete randomization as we increase b, i.e., as we insist that more balls of one color be in hand prior to allocation. Probabilities may be obtained by counting paths as before, but now this is done on a finer grid.

Table 3 shows the effect on allocation probabilities caused by increasing b when $n = 6$. Calculating probabilities by path counting is cumbersome for large b, and a simple asymptotic approximation is both helpful and informative. For notation, we code the random allocation sequence with the vector $\boldsymbol{V}_n^b = (V_{n,1}^b, V_{n,2}^b, \ldots, V_{n,n}^b)$ where each $V_{n,j}^b$ takes values in $\{1, -1\}$ instead of $\{\mathbf{g}, \mathbf{r}\}$. We say that a possible allocation sequence $\boldsymbol{v} = (v_1, \ldots, v_n)$ is blocked if for all odd j, $v_j = -v_{j+1}$. Recall that a blocked allocation has $m = n/2$ blocks of size 2.

Theorem 1 *As $b \to \infty$, $\boldsymbol{V}_n^b$ converges in distribution to $\tilde{\boldsymbol{V}}_n$ where*

$$\mathrm{P}(\tilde{\boldsymbol{V}}_n = \boldsymbol{v}) = \begin{cases} \mathrm{P}(\mathrm{sign}(X_k) = v_{2k-1}, \text{ for all } k = 1, 2, \ldots, m) & \text{if } \boldsymbol{v} \text{ is blocked} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ is a mean zero multivariate normal random vector with covariance matrix $\boldsymbol{\Sigma} = (\sigma_{i,j})$, $\sigma_{i,j} = \sigma_{j,i} = i(m-j)/m^2$ for $1 \le i \le j < m$, and $\sigma_{i,m} = 1[i = m]$ for $1 \le i \le m$.

Thus, for large b, relay randomization is similar to simple randomized blocking, except that allocations in different blocks are not necessarily independent. The allocation vector $\tilde{\boldsymbol{V}}_n$ would correspond to simple randomized blocking if $\sigma_{i,j} = 0$ for all $i \neq j$. Curiously, in the limit, the last block is uncorrelated with the others, although convergence to this uncorrelated state appears to be slow. Blocked allocations which alternate treatment groups have highest probability. For general m there is no closed form expression for (1), although the numerical method in Moran (1986) may enable calculation. Note that the set of blocked allocations is much smaller than the set of possible allocations via complete randomization, i.e., $2^m < \binom{2m}{m}$ for $m > 1$.

A preliminary step to proving Theorem 1 is to show that the probability that allocations occur in blocks approaches one asymptotically in b. We do this by analyzing the allocation time $T_{b,c,j}$ which is the random position in the primary sequence where the $j$bth ball of color c appears, allowing a treatment groups. Noting the allocation rule, $T_{b,c,j}$ is time in $\{1, 2, \ldots, N = nb\}$ corresponding to the jth allocation of color c.

Lemma 1 *For every $j = 1, 2, \ldots, m$, $c = 1, 2, \ldots, a$, and $\epsilon > 0$,*

$$\lim_{b \to \infty} \mathrm{P}\left(\left| \frac{T_{b,c,j}}{N} - \frac{j}{m} \right| > \epsilon \right) = 0.$$

In other words, asymptotically in b, the jth allocation for any group occurs very close to the j/m fraction of the primary sequence. The first allocation of each group must occur near $1/m$, the second allocation for each group must occur near $2/m$, and so on, implying that the allocation sequence is blocked. We prove Lemma 1 in Appendix A.

The second step in our proof of (1) is to recognize a well-known property of the primary sequence when we restrict again to two treatment groups. For notation, we code the primary sequence as $\boldsymbol{U}_N = (U_{N,1}, U_{N,2}, \ldots, U_{N,N})$, where $N = nb$ and $U_{N,i} \in \{1, -1\}$ instead of $\{\mathbf{g}, \mathbf{r}\}$. We noted above that relay randomization amounts to processing a random path from the $(N/2 \times N/2)$ lattice. Equivalently, the primary sequence $\boldsymbol{U}_N$ is a random permutation of

$$(\underbrace{-1, -1, \ldots, -1}_{mb=N/2}, \underbrace{1, 1, \ldots, 1}_{mb=N/2}),$$

and thus, from Billingsley (1968, page 209), as $N \to \infty$,

$$\mathbb{U}_N(t) := \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor Nt \rfloor} U_{N,i} \qquad \text{converges weakly to } \mathbb{U}(t)\text{, a Brownian bridge.} \tag{2}$$

These processes are indexed by time $t \in [0, 1]$, thought of as relative position in the primary sequence, and $\lfloor \cdot \rfloor$ is the floor (integer part) function. We recall that $\mathbb{U}(t)$ is a mean zero Gaussian process with $\text{cov}(\mathbb{U}(s), \mathbb{U}(t)) = s(1 - t)$ for $s \leq t$. It remains to relate the limiting allocation vector $\tilde{\boldsymbol{V}}_n$ to the limiting process $\mathbb{U}(t)$. Details are in Appendix A.

A second interesting consequence of (2) concerns fluctuations in the allocation sequence itself when b is fixed but n is large. Predictions from this approximation are compared with a simulation study of imbalance in Section 3. With allocations $V_{n,j}^b$ defined as above (before Theorem 1), we define the allocation process

$$\mathbb{V}_n^b(t) = \sqrt{\frac{b}{n}} \sum_{j=1}^{\lfloor nt \rfloor} V_{n,j}^b$$

for $t \in [0, 1]$ denoting the relative position in the allocation sequence. Evaluated at the jth allocation, $\mathbb{V}_n^b(j/n)$ is proportional to the imbalance up to that point in the study, that is the difference in the number of allocations to each group. In Appendix A we prove

Theorem 2 *As $n \to \infty$ with b fixed, $\mathbb{V}_n^b(t)$ converges weakly to a Brownian bridge.*

3 HARMONY VERSUS INVENTION

We consider further in this section the extent to which one can predict allocations generated by relay randomization. Complete randomization, which produces uniformly distributed allocations, serves as a natural reference point. Among the many properties of a design, we focus here on the absolute imbalance as a function of time into the study, and on the entropy of the allocation distribution.

Imbalance at the jth allocation is the difference in the numbers of subjects allocated to each of the two treatment groups. Large imbalance is problematic, as discussed in Section 1, because the experiment becomes vulnerable to time-related confounding factors. On the other hand, very small imbalance corresponds to a loss of randomness, which, as we discuss in Section 4, can lead to reduced power in the absence of such factors. A reduction in randomness also lowers the investigator's defenses against other potential confounding factors.

Figure 4 illustrates the expected absolute imbalance as a function of allocation time for several designs. These numbers are calculated by simulation. Not surprisingly, relay randomization is intermediate between complete randomization and simple randomized blocking. Table 4 reports an imbalance statistic for a wider range of settings. The mean expected absolute imbalance is simply the arithmetic mean over allocations of expected absolute imbalance. From Theorem 2, the imbalance at allocation j is approximated by a mean 0 normal random variable with variance $j(n-j)/(nb)$ when n is relatively large. This gives us a second approximation to mean expected absolute imbalance, as reported in Table 4. We observe that imbalance decreases with b for each sample size n, and increases with n for each b. Also the rate at which imbalance increases with n diminishes for large values of b.

Classically, the entropy of a distribution measures its level of randomness, with maximum entropy at the uniform distribution. The probability distribution induced on allocations by a given design may be represented by a long vector $p = (p_1, p_2, \ldots, p_k)$ where $k = \binom{2m}{m}$ is the total number of ways to allocate $2m$ subjects into two equal groups. The entropy of p is defined

$$H = -\sum_{i=1}^{k} p_i \log(p_i) \tag{3}$$

a concave function maximized when $p_i = 1/k$ for all i. Thus, complete randomization has entropy $\log \binom{2m}{m}$. Simple randomized blocking is uniform over a subset of allocations, and has entropy $2m\log(2)$. The entropy for relay randomization is difficult to calculate exactly, but it is readily approximated by simulation, noting that H is the expectation of minus the log probability mass, and that probability mass can be calculated by path counting. In Table 5 we report H for a range of designs and sample sizes. As expected, the entropy is a decreasing function of b for fixed sample size. The first row of Table 5, showing $n = 2m = 6$, illustrates a general phenomenon that the entropy of relay randomization drops below that of simple randomized blocking for sufficiently large b. This is related to the positive dependence among blocks shown in Section 2.

4 POWER CHARACTERISTICS

4.1 Set up

Experimental designs may be compared by calculating the probability of detecting departures from a null hypothesis of interest, i.e., by calculating power. For example, when testing for a difference between two population means, the unpaired design may be compared to a paired design. Pairing (i.e., blocking, matching) is known to be advantageous if, under the null hypothesis, measurements within a pair are more similar than measurements in different pairs because the background noise is effectively reduced. On the other hand, the paired design is not superior to the unpaired design if this within-pair correlation is small, and a well-known trade-off presents itself (e.g., Box, Hunter, and Hunter, 1978, page 103; Lehmann, 1986, page 264). Being intermediate between complete randomization (unpaired design) and randomized blocking (paired design), at least for moderate b, relay randomization may have power advantages over both extreme designs in certain situations.

To study power, we first decompose the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ as $\boldsymbol{Y} = \theta\boldsymbol{V} + \boldsymbol{W}$ where $\boldsymbol{V}$ is the column vector of binary allocations ($V_i \in \{1, -1\}$), and $\boldsymbol{W}$ represents some kind of deviation vector. (We simplify the subscripting notation in this section since we are not emphasizing limiting operations.) The two groups are different because an amount θ is added to or removed from W_i, depending on the allocation V_i. All that we require of the deviations W_i is that they be stochastic with well-defined mean and covariance, and that they be independent of the statistician-supplied allocation vector $\boldsymbol{V}$. Internally they can have arbitrary dependence structure. In any event, 2θ records the difference in expected response for an observation allocated to $V_i = 1$ versus $V_i = -1$. The difference in sample means between the two groups is proportional to $\boldsymbol{V}^t\boldsymbol{Y}$.

A randomization test compares the observed value of a test statistic to the randomization distribution, i.e., its conditional distribution given $\boldsymbol{Y}$. Of course, randomness here enters through hypothetical new allocations, and the data are fixed. In a one-sided test of $H_0 : \theta = 0$ versus $H_a : \theta > 0$, we reject H_0 if the observed test statistic exceeds the $(1-\alpha)$-quantile of this randomization distribution. A natural test statistic in this case is the observed group difference $\boldsymbol{V}^t\boldsymbol{Y}$. By calculating with respect to the distribution of $\boldsymbol{V}$,

$$\mathrm{E}(\boldsymbol{V}^t\boldsymbol{Y}|\boldsymbol{Y}) = 0, \qquad \text{and} \qquad \mathrm{var}(\boldsymbol{V}^t\boldsymbol{Y}|\boldsymbol{Y}) = \boldsymbol{Y}^t\boldsymbol{C}\boldsymbol{Y}$$

where $\boldsymbol{C} = (c_{i,j})$ is the covariance matrix of $\boldsymbol{V}$, a matrix carrying information about the design. Appendix B gives the randomization variance for some standard designs. Note that since V_i is binary, and $E(V_i) = 0$, we have $c_{i,j} = 2\mathrm{P}(V_i = V_j) - 1$ and $\sum_j c_{i,j} = 0$ for all i.

Some simplifying approximations are required to compare power analytically across the different designs. First, we suppose that the randomization distribution is normal, and so the decision rule is determined by the significance level α and the randomization variance $\boldsymbol{Y}^t\boldsymbol{C}\boldsymbol{Y}$ only. Combinatorial central limit theory justifies this step to some extent. Second, we suppose that fluctuations in $\boldsymbol{Y}^t\boldsymbol{C}\boldsymbol{Y}$ are of a smaller order than those of $\boldsymbol{V}^t\boldsymbol{Y}$, and so we can approximate $\boldsymbol{Y}^t\boldsymbol{C}\boldsymbol{Y}$ by its marginal expectation. Here is the first place where we require a sampling model for $\boldsymbol{W}$ in the decomposition $\boldsymbol{Y} = \theta\boldsymbol{V} + \boldsymbol{W}$. Calculating under the alternative hypothesis,

$$\begin{aligned} E(\boldsymbol{Y}^t\boldsymbol{C}\boldsymbol{Y}) &= E[(\theta\boldsymbol{V} + \boldsymbol{W})^t\boldsymbol{C}(\theta\boldsymbol{V} + \boldsymbol{W})] \\ &= \theta^2 E(\boldsymbol{V}^t\boldsymbol{C}\boldsymbol{V}) + E(\boldsymbol{W}^t\boldsymbol{C}\boldsymbol{W}) \end{aligned}$$

using independence of $\boldsymbol{V}$ and $\boldsymbol{W}$. With $\rho_{i,j} = E(W_iW_j)$, the expected variance simplifies to

$$E(\boldsymbol{Y}^t\boldsymbol{C}\boldsymbol{Y}) = \theta^2\sum_{i,j} c_{i,j}^2 + \sum_{i,j} c_{i,j}\rho_{i,j}.$$

Both sums are over all pairs (i, j) on the $n \times n$ lattice, a notation which we suppress below. In this approximate randomization test, the decision rule is to reject H_0 if

$$\boldsymbol{V}^t \boldsymbol{Y} \geq z_\alpha \sqrt{\theta^2 \sum c_{i,j}^2 + \sum c_{i,j} \rho_{i,j}} \tag{4}$$

where z_α is the $(1-\alpha)$-quantile of a standard normal distribution. Although the test is derived under H_0, properties of the test must be established in general, and hence θ appears in the approximation (4).

Power is calculated by noting that $\boldsymbol{V}^t \boldsymbol{Y} = n\theta + \boldsymbol{V}^t \boldsymbol{W}$ and by taking a second normal approximation, this time for the marginal distribution of $\boldsymbol{V}^t \boldsymbol{W}$. The expectation of $\boldsymbol{V}^t \boldsymbol{W}$ is 0, and its variance is $\sum_{i,j} c_{i,j} \rho_{i,j}$. Therefore, the approximate power for detecting an effect θ is the probability that a standard normal variate exceeds

$$z^* = \frac{-n\theta}{\sqrt{\sum c_{i,j} \rho_{i,j}}} + z_\alpha \sqrt{1 + \frac{\theta^2 \sum c_{i,j}^2}{\sum c_{i,j} \rho_{i,j}}}. \tag{5}$$

Evidently, the power depends on the design through $c_{i,j}$ and on potential confounding factors through $\rho_{i,j}$. Several interesting observations can be derived from (5), including the well-known trade-off between the paired and unpaired designs. That such a phenomenon can be revealed using these simple normal approximations is somewhat unexpected, since a discussion of degrees of freedom in a t-distribution is usually required (Box Hunter and Hunter, 1978; Lehmann, 1986). In any event, (5) allows us to compare the power of different designs.

4.2 No Confounders

In the decomposition $\boldsymbol{Y} = \theta \boldsymbol{V} + \boldsymbol{W}$, the deviations $\boldsymbol{W}$ account for all variation in the measurements not accounted for by group differences. We decompose $\boldsymbol{W}$ further by $W_i = \mu + S_i + \epsilon_i$, where ϵ_i are independent, mean 0 and variance σ^2 random errors, and S_i is some kind of unobserved confounding variable, and μ is an overall mean.

The case of no confounding factors can be expressed $S_i = 0$ for all i, and so $\rho_{i,j} = 0$ when $i \neq j$. The percentile in (5) becomes

$$z^* = \frac{-\sqrt{n}\theta}{\sigma} + z_\alpha \sqrt{1 + \frac{\theta^2 \sum c_{i,j}^2}{n\sigma^2}}.$$

We can calculate

$$\sum c_{i,j}^2 = \begin{cases} n + n/(n-1) & \text{for complete randomization} \\ 2n & \text{for simple randomized blocking} \end{cases}$$

which indicates that complete randomization indeed is a more powerful design than randomized blocking. For small to moderate b, the power of relay randomization appears to be in between these extremes. We have been unable to simplify $\sum c_{i,j}^2$, but it is easily approximated by simulating allocations. The relationship between power and b is calculated for one specific case as the top line of Figure 5a.

4.3 Confounders

In the presence of confounding factors, complete randomization may not be as powerful as other designs. One way to show this is with random-walk confounding:

$$S_i = \sum_{l=1}^{i} \delta_l$$

where δ_l are iid mean 0 random variables with variance τ^2. The same model has been used in agricultural field trials to model fluctuations in fertility (e.g., Besag and Kempton, 1986; Besag and Higdon, 1993). It is readily calculated that

$$\rho_{i,j} = \begin{cases} \mu^2 + \sigma^2 + i\tau^2 & \text{if } i = j \\ \mu^2 + \min(i,j)\,\tau^2 & \text{if } i \neq j \end{cases} \tag{6}$$

from which we can calculate percentile z^* needed to obtain approximate power. For complete randomization,

$$z^* = \frac{-\sqrt{n}\theta}{\sqrt{\sigma^2 + \tau^2(n+1)/6}} + z_\alpha \sqrt{1 + \frac{\theta^2 n/(n-1)}{\sigma^2 + \tau^2(n+1)/6}},$$

in contrast to

$$z^* = \frac{-\sqrt{n}\theta}{\sqrt{\sigma^2 + \tau^2/2}} + z_\alpha \sqrt{1 + \frac{2\theta^2}{\sigma^2 + \tau^2/2}}$$

for simple randomized blocking. As expected, these agree when $n = 2$ but simple randomized blocking becomes more powerful for larger n. Indeed, the power of complete randomization does not converge to one if there is random-walk confounding.

Figure 5a compares the power of different designs using the approximation (5) both in the absence and the presence of random-walk confounders. The case under consideration has two groups of $m = 10$ subjects, with noise variance $\sigma^2 = 1$, a slightly larger difference in means, $2\theta = 7/6$, and a one-sided .05 test. We observe that in the absence of such confounders, the power drops as b increases, and relay randomization becomes less powerful than simple randomized blocking at about $b = 7$. In the presence of confounders, complete randomization loses more power than the relay designs, creating a power mode at some $b > 1$. When the size of the confounding is large ($\tau^2 = 1/5$), the best relay design is no better than simple randomized blocking, but for mild confounding, this best relay design is superior to both complete randomization and randomized blocking. The τ^2 value for the three cases of confounding may be translated into a standard deviation of S_{20}, the confounding factor at the study's conclusion. These numbers are 0, 1, and 2 units, respectively.

The same qualitative features characterize a simulation-based approximation of these powers, summarized in Figure 5b. For each design, and each specification of confounders, 10000 data sets were simulated. Normally distributed errors and confounders were used. A randomization test using 499 random allocations was performed for each data set, and the decision rule was recorded. One clear difference between Figure 5a and 5b is the boost in power for the simulation study. This is not surprising since the latter uses a much better approximation to the rejection region. Relay randomization retains high power in the absence of confounders and loses relatively little power in their presence.

The reader may object to the use of random walk confounders in the definition of W_i, and certainly other forms could be considered. One possibility that may occur in a laboratory study is

$$W_i = \mu + \epsilon_i + \kappa 1[T \leq i]$$

where T is some random allocation time, perhaps uniformly distributed on $\{1, 2, \ldots, n\}$, κ indicates a shift in expected response after the change point T, and ϵ_i is the same noise as before. This could model the effects of viral infection in the breeding colony, for example, or the effects of changing the lab technician. Interestingly, $E(W_i W_j)$ becomes

$$\rho_{i,j} = \begin{cases} \mu^2 + \sigma^2 + i\kappa^2/n & \text{if } i = j \\ \mu^2 + \min(i,j)\,\kappa^2/n & \text{if } i \neq j \end{cases}$$

and this is identical in form to (6). In other words, with n fixed, one cannot distinguish the two types of confounders using covariance only.

5 WITH REPLACEMENT RELAYS

There is a natural version of relay randomization based on with-replacement rather than without-replacement urn sampling. Subjects present themselves sequentially for allocation, as above. The first subject rolls an $a-$sided die repeatedly until b of one type appear. Allocation is to the group corresponding to this plurality type. The remaining outcomes are relayed to the next subject, who treats these as initial rolls in an experiment to obtain b outcomes of a plurality type. Allocation, relaying, and sampling continue in this fashion. This design encourages balance as before, but the size of each treatment group is always random.

The reader may recognize structural similarities with some well-known problems. Take the case of $a = 2$ treatment groups. Allocation for the first subject is equivalent to determining the winner in a best of $2b - 1$ game series when each team has probability of $1/2$ of winning and games are independent. The number of games won by the series loser is the size of the relay. If such a series is interrupted prior to completion, one may ask the probability that a particular team ultimately would have won. This is the famous "problem of points" (also called the sharing problem) posed by early Italian probabilists and correctly solved by Pascal and Fermat, independently in 1650 (Jordan, 1972). Sobel and Frankowski (1994) discuss some recent developments. An open question is whether or not with-replacement relay randomization has advantages over the designs suggested by Efron (1971) and Wei (1977).

6 CONCLUDING REMARKS

We have introduced a simple allocation method for one-way layouts. The method achieves balance throughout the experiment without sacrificing randomness, and this gives it advantages in power over existing standard designs in certain situations. It may be particularly useful when an experiment is drawn out over time, and the investigator is concerned about the effects of confounding factors. It compares well to randomized blocking as long as confounding effects are not severe, and somewhat simpler to implement. There remain a number of open questions concerning our proposed randomization method. The advantages over standard designs need to be explored more fully before relay randomization can be generally recommended. Additionally, there is the question of how to modify this design for more complex situations than the one-way layout.

ACKNOWLEDGEMENTS

The authors thank Tom Kurtz for helpful insight on Theorem 2. The first author was supported in part by a contract from the National Cancer Institute. The second author's time in Madison was supported in part by the NSF through the Probability Intern Program. Code to simulate relay randomization designs and to calculate allocation probabilities is available at the first author's web site `http://www.stat.wisc.edu/~newton/newton.html`

Appendix A

Proof of Lemma 1: Let $T = T_{b,c,j}$ for fixed b, c, and j. The exact distribution of T, which does not depend on c due to symmetry, satisfies

$$\mathrm{P}(T = k) = \frac{\binom{k-1}{jb-1}\binom{amb-k}{mb-jb}}{\binom{amb}{mb}}$$

because for the jbth ball of a color to appear on exactly the kth draw of the primary sequence, there must be exactly $jb-1$ balls of that color in the first $k-1$ draws and $mb-jb$ balls in the last $amb-k$ draws, and there are $\binom{amb}{mb}$ ways to place the balls of a given color in the primary sequence.

Using the fact that $k\binom{k-1}{x-1} = x\binom{k}{x}$, the well-known combinatorical identity

$$\sum_k \binom{r+k}{x}\binom{s-k}{y} = \binom{r+s+1}{x+y+1}$$

and straightforward algebraic manipulation, we find from the definition of expectation that

$$\mathrm{E}[T] = \frac{jb(amb+1)}{(mb+1)} = ajb + o(b)$$

A similar calculation yields

$$\mathrm{E}[T(T+1)] = \frac{(jb+1)(jb)(amb+2)(amb+1)}{(mb+2)(mb+1)}$$

It follows that

$$\begin{aligned} \mathrm{var}[T] &= \mathrm{E}[T(T+1)] - \mathrm{E}[T] - (\mathrm{E}[T])^2 \\ &= \frac{(jb)(amb+1)((a-1)mb)}{(mb+1)(mb+2)} = o(b^2) \end{aligned}$$

Now $\mathrm{E}[T/amb] = j/m + o(b)/b$ and $\mathrm{var}[T/amb] = o(b^2)/b^2$ and the lemma follows from Chebyshev's inequality and letting $b \to \infty$. $\square$

Proof of Theorem 1: By Lemma 1, the jth block of allocations occurs at the rescaled time of j/m. The event that the jth block is $(-1,1)$ or $(1,-1)$ corresponds with whether $\mathbb{U}(j/m)$ is negative or positive, respectively, for $j < m$. The order of the last block cannot be determined from the limiting process, but is equally likely to be either possibility by symmetry. The last block of allocations will occur at a scaled time just prior to 1. The correlation of the random variables at times s and $1-\epsilon$ in a Brownian bridge equals $\sqrt{s\epsilon/((1-s)(1-\epsilon))}$ which vanishes as ϵ approaches 0 for each s and we see that the last block is independent of all other blocks. (This follows intuitively because the path of a Brownian bridge crosses 0 infinitely often in a neighborhood of 1 and the probability that our urn will contain equal numbers of each color ball at some time after the penultimate block is allocated regardless of past allocations approaches one.) Hence, the limiting probability of any allocation is determined by the signs of the random vector $\boldsymbol{X} = (\mathbb{U}(1/m), \mathbb{U}(2/m), \ldots, \mathbb{U}((m-1)/m), Z)$ where Z is a complete normal random variable independent of everything else. $\square$

Proof of Theorem 2: For $1 \leq k \leq N$ a position in the primary sequence,

$$G_k = (k + \sqrt{N}\mathbb{U}_N(k/N))/2$$

indicates the number of green balls (i.e., $1's$) sampled up to that point, and so $k - G_k$ records the number of red balls sampled. Allocations in the first k sampled balls must be $\lfloor G_k/b \rfloor$ to green, and $\lfloor (k - G_k)/b \rfloor$ to red, and so we can define a scaled allocation imbalance process as a function of relative position in the primary sequence:

$$\begin{aligned}
\mathbb{A}_N(t) &= \sqrt{\frac{b}{n}}\left(\left\lfloor \frac{G_{\lfloor Nt \rfloor}}{b} \right\rfloor - \left\lfloor \frac{\lfloor Nt \rfloor - G_{\lfloor Nt \rfloor}}{b} \right\rfloor\right) \\
&= \frac{b}{\sqrt{N}}\left(\left\lfloor \frac{(\lfloor Nt \rfloor + \sqrt{N}\mathbb{U}_N(\lfloor Nt \rfloor))/2}{b} \right\rfloor - \left\lfloor \frac{(\lfloor Nt \rfloor - \sqrt{N}\mathbb{U}_N(\lfloor Nt \rfloor))/2}{b} \right\rfloor\right) \\
&= \frac{b}{\sqrt{N}}\left(\frac{\sqrt{N}\mathbb{U}_N(\lfloor Nt \rfloor)}{b} + O(1)\right) \\
&= \mathbb{U}_N(t) + \frac{O(b)}{\sqrt{N}}
\end{aligned}$$

If we let $n \to \infty$ for b fixed, $N = nb \to \infty$ and the left term converges in distribution to the Brownian bridge $\mathbb{U}(t)$, while the right term vanishes. The third equality follows because for any two real numbers x and y, $\lfloor x \rfloor - \lfloor y \rfloor = x - y + O(1)$.

Now, for any sequence of draws, $\mathbb{A}_N(t)$ and $\mathbb{V}_n^b(t)$ are both step functions on $[0,1]$ with the same n step values occurring in the same order. The jumps in $\mathbb{V}_n^b(t)$ occur at times $1/n, 2/n, \ldots, 1$, while the jumps in $\mathbb{A}_N(t)$ occur at random times $t_1, t_2, \ldots, t_n = 1$ where $i/n \leq t_i < (i+1)/n$ for every possible sequence of draws. Because $|(i+1)/n - i/n| \to 0$ as $n \to \infty$, it is evident that $\mathbb{V}_n^b(t)$ also converges in distribution to a Brownian bridge. $\square$

Appendix B

For a simple randomized-block design, $V_j = -V_{j+1}$ for odd i, and allocations in different blocks are independent. The randomization variance of $\boldsymbol{V}^t\boldsymbol{Y}$ is proportional to

$$\sum_{\text{odd } j} d_j^2$$

where $d_j = Y_j - Y_{j+1}$. Interestingly, this is almost the same as the sample variance of differences one would use in a paired t-test. That test standardizes by $\sum_{\text{odd } j}(d_j - \bar{d})^2$, but to center at $\theta = 0$ seems more natural from the perspective of the null hypothesis. If the design is complete randomization, then $c_{i,i} = 1$ and $c_{i,j} = -1/(n-1)$ for $i \neq j$. The randomization variance in this case becomes

$$\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

which is proportional to the sample variance of the full data set. By contrast, a two-sample unpaired t-test uses the pooled complete deviation, i.e., the mean of the sample variances within each group. Again, from the perspective of the null hypothesis, the overall sample variance is more natural.

References

Besag, J. and Higdon, D. (1993). Bayesian inference for agricultural field experiments, *Bull. Inst. Internat. Statist.* **55**: 121–136.

Besag, J. and Kempton, R. (1986). Statistical analysis of field experiments, *Biometrics* **78**: 301–304.

Box, G., Hunter, W. and Hunter, J. (1978). *Statistics for Experimenters*, Wiley, New York.

Cochran, W. G. and Cox, G. (1992). *Experimental Designs*, 2nd edn, Wiley, New York.

Efron, B. (1971). Forcing a sequential experiment to be balanced, *Biometrika* **58**: 403–417.

Fisher, R. (1971). *The Design of Experiments*, ninth edn, Macmillan, New York.

Jacoby, R., Marshall, D., Newton, M., Novakovic, K., Tutsch, K., Cole, C., Lubet, R., Kelloff, G., Verma, A., Moser, A. and Dove, W. (1996). Chemoprevention of spontaneous intestinal adenomas in the Apc^{Min} mouse model by the non-steroidal anti-inflammatory drug piroxicam, *Cancer Research* **56**: 710–714.

Jordan, K. (1972). Chapters on the classical calculus of probability, *Disquisitiones Mathematicae Hungaricae*, Akadémiai Kiadó, Budapest, Hungary, pp. 440–449. 4.

Kalish, L. and Begg, C. (1985). Treatment allocation methods in clinical trials: A review, *Statistics in Medicine* **4**: 129–144.

Lehmann, E. (1986). *Testing Statistical Hypotheses*, 2nd edn, Wiley, New York.

Moran, P. (1986). Orthant probabilities and gaussian markov processes, *Journal of Applied Probability* **23**: 413–417.

Sobel, M. and Frankowski, K. (1994). The 500th anniversary of the sharing problem (the oldest problem in the theory of probability), *The American Mathematical Monthly* **101**: 833–847.

Vivaldi, A. (1725). *Gazette d'Amsterdam.* December 14.

Wei, L.-J. (1977). A class of designs for sequential clinical trials, *Journal of the American Statistical Association* **72**: 382–386.

Table 1: Tumor Counts across Treatment Groups: Subscript indicates the order in which the mice entered the study. NA means not available.

Group	Tumor Count							
1	19_2	13_{14}	8_{20}	30_{29}	NA_{33}	13_{39}	16_{46}	22_{53}
2	12_3	20_9	7_{15}	1_{17}	27_{27}	5_{28}	15_{35}	25_{52}
3	7_6	5_{10}	2_{13}	16_{22}	19_{26}	3_{38}	15_{47}	7_{49}
4	2_5	0_{18}	3_{24}	41_{36}	17_{40}	9_{45}	19_{51}	5_{56}
5	9_7	6_{12}	3_{19}	8_{25}	3_{30}	9_{34}	3_{48}	0_{55}
6	4_4	4_8	6_{21}	2_{23}	4_{37}	4_{42}	1_{44}	10_{54}
7	5_1	0_{11}	2_{16}	0_{31}	1_{32}	7_{41}	0_{43}	NA_{50}

Table 2: Probabilities of Allocations for $n = 6$ Subjects and Two Groups. In boldface are ten allocations. Each count is the number of primary sequences (out of 924) leading to the corresponding allocation. Allocations obtained by swapping **r** and **g** have equal probability to the ones shown. Blocked allocations are those in which **r** and **g** appear in each of the three pairs of assignments.

Blocked	
grgrgr	99
grgrrg	75
grrgrg	71
grrggr	55

Unblocked	
ggrrgr	47
ggrrrg	31
grggrr	31
ggrgrr	27
grrrgg	19
gggrrr	7

Table 3: The Effect of Increasing b: Tabulated are probabilities assigned to each allocation for various designs. RB is randomized block, and b indicates which relay randomization.

allocation	$b=1$	$b=2$	$b=3$	$b=5$	$b=10$	$b=100$	$b=\infty$	RB
grgrgr	1/20	.11	.14	.18	.19	.18	1/6	1/8
grgrrg	1/20	.08	.10	.12	.14	.16	1/6	1/8
grrgrg	1/20	.08	.09	.09	.09	.09	1/12	1/8
grrggr	1/20	.06	.06	.07	.07	.08	1/12	1/8
ggrrgr	1/20	.05	.04	.02	.00	.00	0	0
ggrrrg	1/20	.03	.02	.01	.00	.00	0	0
grggrr	1/20	.03	.02	.00	.00	.00	0	0
ggrgrr	1/20	.03	.01	.00	.00	.00	0	0
grrrgg	1/20	.02	.01	.00	.00	.00	0	0
gggrrr	1/20	.01	.00	.00	.00	.00	0	0

Table 4: Mean Expected Imbalance: Mean expected absolute imbalances in treatment assignments for designs with $a = 2$ groups and $m = 6, 10, 20$ and 50 subjects per group. The designs are relay randomization with $b = 1, 2, 5,$ and 10 and simple randomized blocking (RB). Numbers in parentheses are normal approximations. All other figures are either exact or computed from simulations large enough such that the Monte-Carlo standard error is less than .005.

Design	$n = 6$	$n = 10$	$n = 20$	$n = 50$
$b = 1$	.80	1.02	1.41	2.21
	(.71)	(.96)	(1.38)	(2.21)
$b = 2$	.65	.78	1.04	1.60
	(.50)	(.68)	(.98)	(1.56)
$b = 5$	.53	.59	.73	1.05
	(.32)	(.43)	(.62)	(.99)
$b = 10$	.50	.52	.59	.80
	(.23)	(.30)	(.44)	(.70)
RB	.50	.50	.50	.50

Table 5: Entropy as a function of sample size and design: RB refers to simple randomized block. Monte Carlo simulation was used to calculate entropy for relay designs with $b > 1$. The figures are accurate to the number of significant digits reported.

n	$b = 1$	$b = 2$	$b = 5$	$b = 10$	RB
6	3.0	2.8	2.3	2.0	2.1
10	5.5	5.2	4.2	3.5	3.5
20	12.1	11.4	9.3	7.6	6.9
50	32.5	30.8	25.1	20.2	17.3

Figure 1: Relay Randomization with queue size $b = 2$, $a = 3$ groups and $m = 4$ subjects per group: Shown is one particular realization to allocate $n = 12$ subjects. The top row shows the contents of an urn with balls of three types, and the second row shows one permutation of these contents representing the sequence of sampled balls from left to right. Vertical lines indicate times of allocation, with shaded boxes giving the group into which the allocation is made. Clear boxes contain the relayed balls.

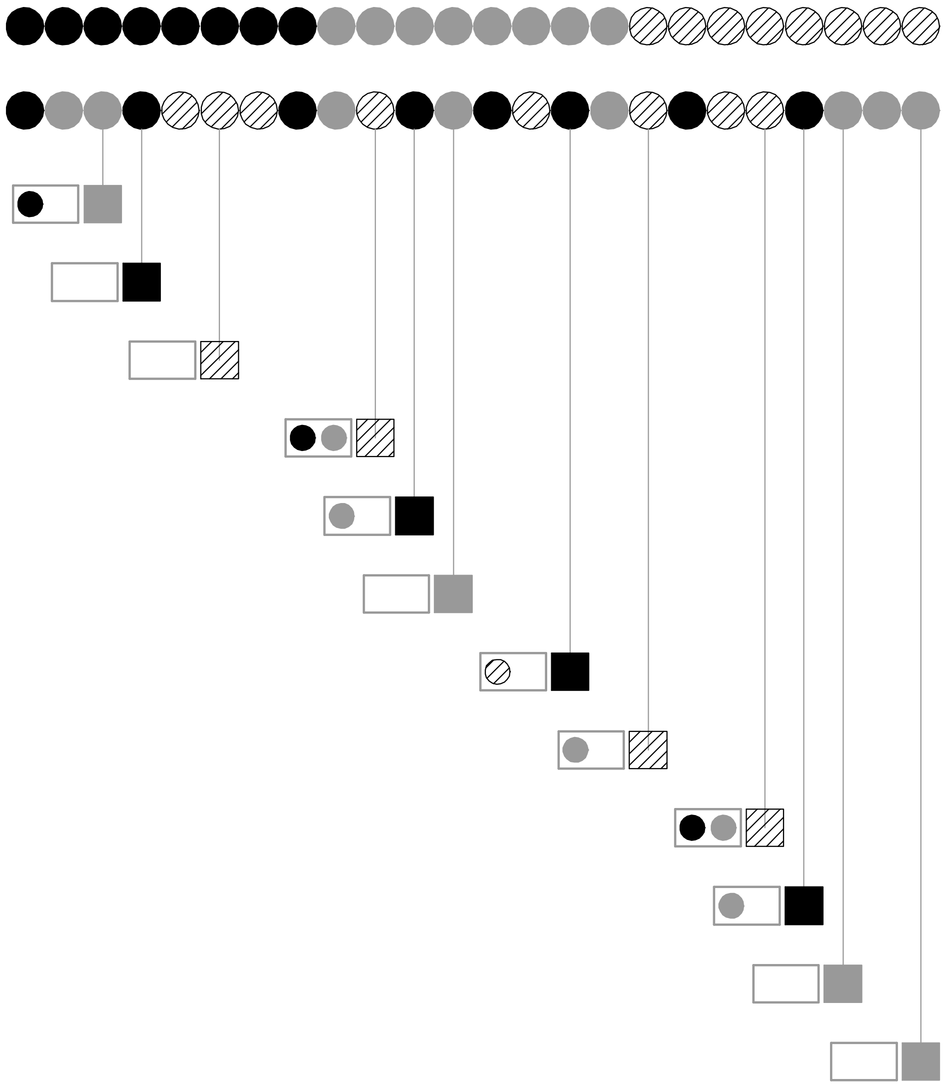

Figure 2: Randomization Distribution of F-statistic: Solid line is $F_{6,49}$ density, the normal theory approximation. The observed statistic is 3.97, yielding an exact Monte Carlo p-value of .0018 using 4999 simulated statistics.

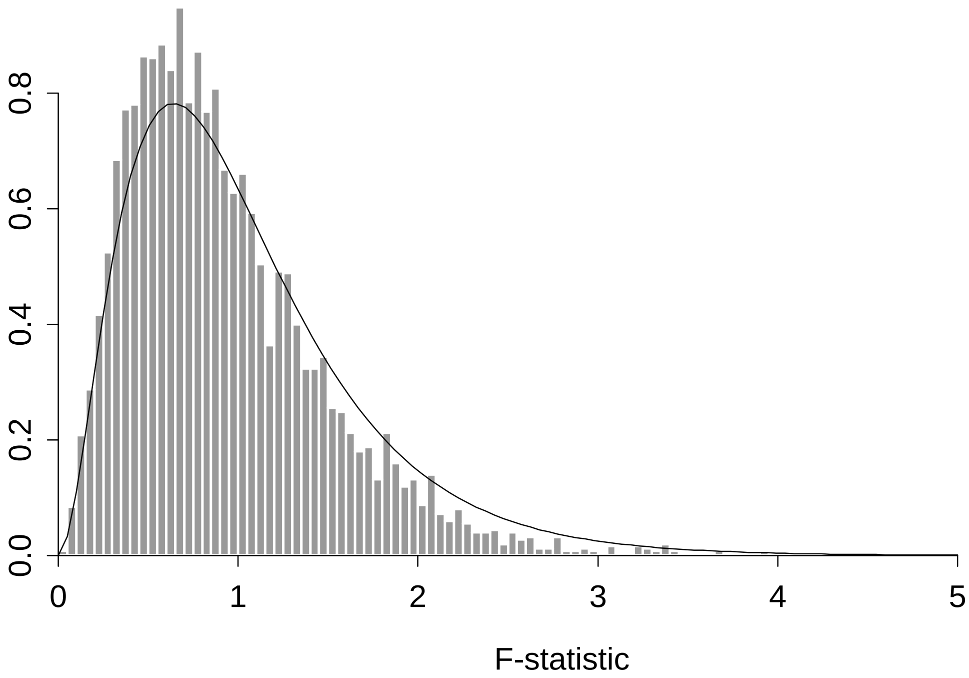

Figure 3: Lattice Representation: Each primary sequence is a monotone nondecreasing path from **start** to **finish**, such as **ggrgrrrgrggr** indicated in grey in the left panel. An allocation occurs when the path reaches one of the heavy grid lines. Any path through the shaded region in the right panel yields the same allocation **grrggr**. Each number indicates how many legal paths pass through the grid point just below and to the left of the number.

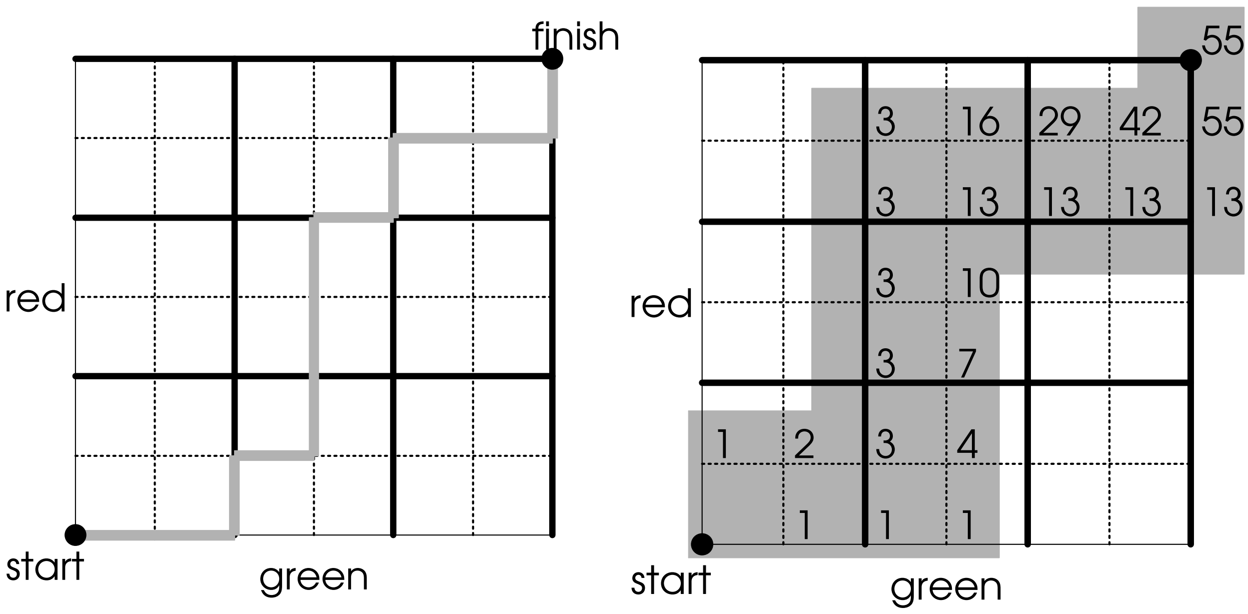

Figure 4: Expected absolute imbalance versus assignment number for designs with $a = 2$ groups and $m = 10$ subjects per group. Imbalance at even assignments is plotted.

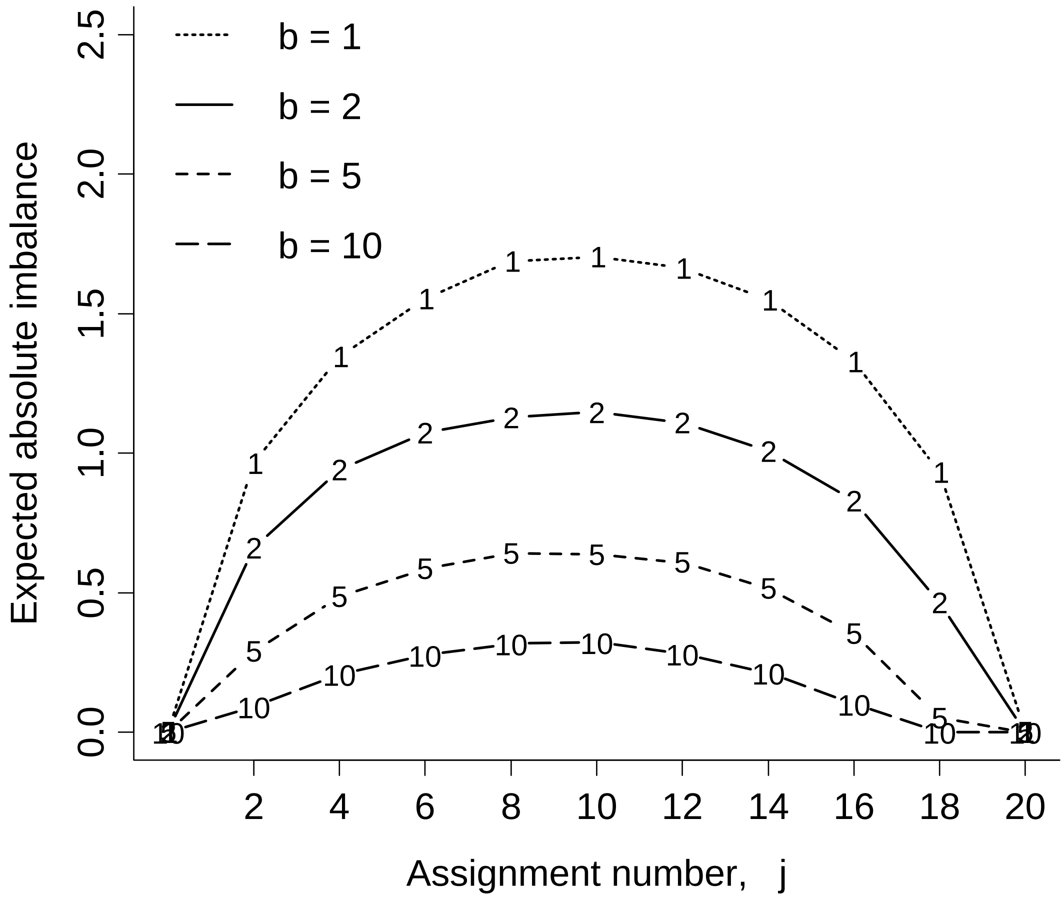

Figure 5: Power Study: Analytical (a) and simulation (b) approximations to the power of detecting a mean difference $2\theta = 7/6$ with a sample of $n = 20$ for different designs. Horizontal axis is the design parameter b. In all cases the background noise is $\sigma^2 = 1$. The three cases are no confounding, $\tau^2 = 0$ (1 =solid curve), mild confounding $\tau^2 = 1/20$ (2 =dotted), and substantial confounding $\tau^2 = 1/5$ (3 =dashed curve). Horizontal lines show the corresponding power for the randomized-block design. The vertical bar in the lower right of (b) has length equal to .009, approximately 2 Monte Carlo standard errors.