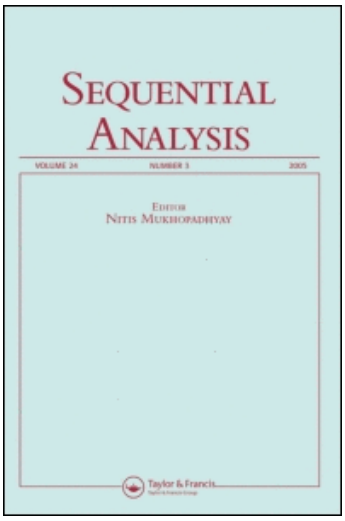

Sequential Analysis
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597296

Stopping Rules in Balanced Allocation Problems: Exact and Asymptotic Distributions

Andrew L. Rukhin [ab]
[a] Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA [b] Department of Math&Stat, UMBC, Baltimore, Maryland, USA

Online Publication Date: 01 July 2008

To cite this Article Rukhin, Andrew L.(2008)'Stopping Rules in Balanced Allocation Problems: Exact and Asymptotic Distributions',Sequential Analysis,27:3,277 — 292
To link to this Article: DOI: 10.1080/07474940802241025
URL: http://dx.doi.org/10.1080/07474940802241025

Sequential Analysis, 27: 277–292, 2008

ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474940802241025

Taylor & Francis
Taylor & Francis Group

Stopping Rules in Balanced Allocation Problems: Exact and Asymptotic Distributions

Andrew L. Rukhin
Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA and Department of Math&Stat, UMBC, Baltimore, Maryland, USA

Abstract: The properties of balanced randomization procedures for several treatments under a general (nonuniform) allocation are reviewed. Under one of such schemes, the truncated multinomial randomization design, the initial assignment probabilities are employed until a treatment fulfills its quota of subjects, after which these probabilities become these of the conditional distribution for the remaining available treatments, and so on. Another design, the random allocation rule, chooses at random one of possible balanced assignments of the given number of subjects per treatment. These two schemes turn out to be quite different: the target fulfillment instants for the truncated multinomial randomization design occur much earlier than for the random allocation rule. Under the truncated multinomial randomization design, the limiting behavior of these instants is determined by normal order statistics with different variances; for the random allocation rule this role is played by the sums of independent but heterogeneous geometric random variables. The formulas for the selection bias and the accidental bias of both procedures are given.

Keywords: Accidental bias; Balanced randomization; Gamma-distribution; Markov randomization design; Negative binomial distribution; Normal order statistics; Occupancy problem; Probability generating function; Schur convexity; Selection bias.

Subject Classifications: 60E05; 60C05; 60G40; 62E20; 62G30; 62P10.

1. INTRODUCTION

A balanced allocation among K treatments for n subjects is achieved by means of a randomization design Π. In our setting, each treatment must meet its

Received August 26, 2007, Revised March 29, 2008, and May 29, 2008,
Accepted June 2, 2008
Recommended by Shelley Zacks
Address correspondence to Andrew L. Rukhin, Statistical Engineering Division, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8980, USA; Fax: 301-975-3144; E-mail: andrew.rukhin@nist.gov

allocation target, a positive integer np_k for the treatment k, $k = 1, \ldots, K$, $\sum_k p_k = 1$. A randomization design is merely a probability distribution over the set of all allowable assignment sequences of length n conditioned by ending up with np_k subjects per treatment k. The assignment of subjects to these treatments is performed according to Π; i.e., at any stage subjects are allocated to a treatment according to their conditional distribution given the previous allocations. If this distribution depends only on the current allocation numbers, Π is called a *Markov* design.

Traditionally the two following randomization schemes are offered for this purpose when $p_k \equiv K^{-1}$. The *random allocation rule* is given by the uniform distribution over all balanced assignment sequences. The *truncated multinomial design* starts with the initial probability distribution of treatments; i.e., the treatment k is originally chosen with probability π_k and proceeds until one of the treatments, say, ξ_1, receives its quota of subjects. At this moment the distribution is changed to the remaining $K - 1$ treatments, whose probabilities become $\pi_k/(1 - \pi_{\xi_1})$, $k \neq \xi_1$. The allocation process continues in this way until there is just one treatment that does not meet its target, and this treatment gets all remaining subjects.

When $K = 2$ and $\pi_1 = p_1 = 1/2$, both of the described (Markov) procedures were introduced by Blackwell and Hodges (1957). These authors demonstrated that the truncated binomial design has the smallest value of the largest selection bias; i.e., is minimax. In the same situation, the random allocation rule has Bayes nature and can effectively eliminate the bias asymptotically (Stigler, 1969). Under a different criterion the truncated binomial design loses its minimaxity and becomes dominated by the random allocation rule (Wei, 1978). Our results for the nonuniform case are similar.

Fair (i.e., randomized) allocation to each of the treatments is a standard practice in comparative studies, especially in clinical trials. Many sequential randomization schemes (Efron's biased coin design, Wei's adaptive urn procedure, etc.) endeavor to equalize treatment assignments being nonbalanced; i.e., they do not guarantee firm quotas of the subjects per treatment. A good survey of these designs and their comparative merits in clinical trials is given by Rosenberger and Lachin (2002). See also Hu and Rosenberger (2006). Possibility of unequal allocation probabilities in randomization of some clinical trials has been suggested by Peto (1978).

The balanced allocation problem is closely related to the classical occupancy problem with sequentially arriving subjects (balls) assigned to one of K treatments (urns). Johnson and Kotz (1977) surveyed the subject. The later work (Anderson et al., 1982; Fang, 1982; Holst, 1986) established the relationship between the gamma-distribution order statistics and the quota fulfillment instants under uniform allocations. A more general model in which the frequency quotas are combined with succession quotas was considered in Sobel and Ebneshahrashoob (1992); the recurrence relations for the sooner and later waiting time for frequency quotas are given by Ling (1992). Rukhin (2004) studied the limiting distribution of the first fulfillment instant and its relationship to classical problems of probability theory. Randomization designs appear in caching algorithms (Flajolet et al., 1992) and in computer storage problems (Rukhin, 2006).

In this article we review the distributions of fulfillment instants of the truncated multinomial design and of the random allocation rule. Conditional and unconditional distributions of these instants for both procedures are discussed in Section 2. In Sections 3 and 4 formulas for the selection bias and the accidental

bias are obtained. Section 5 discusses some issues of statistical inference under a balanced allocation design. This review is concluded with an example in Section 6.

2. DISTRIBUTIONS OF ALLOCATION VECTORS AND THEIR ASYMPTOTICS

The definition of a randomization design Π gives the joint distribution of K-dimensional random vectors $T_1, \ldots, T_n$ representing the allocation of subjects to K treatments. Thus, $T_i = (T_i(1), \ldots, T_i(K))^T$, $T_i(k) = 1$, if at stage i the subject is assigned to treatment $k, k = 1, \ldots, K$, and $T_i(k) = 0$, otherwise. For a Markov randomization design the conditional probability distribution of T_{j+1},

$$P\left(T_{j+1} = \epsilon_r | \sum_{i=1}^{j} T_i = t\right) = p_j(t, t + \epsilon_r),$$

depends only on $\sum_{i=1}^{j} T_i$ with the initial distribution, $P(T_1 = \epsilon_r) = \pi_r$. Here ϵ_r, $r = 1, \ldots, K$, denotes the rth basis vector. Thus, at stage j the allocation is performed according to this conditional distribution. If $\sum_{i=1}^{j} T_i = t$, then $t = (t_1, \ldots, t_K)^T$ is an integer-valued vector such that $\sum_k t_k = j$. In other terms, $\sum_{i=1}^{j} T_i$ forms a Markov chain with time-dependent transition probabilities, $p_j(t, t')$, which can be positive only if $t' - t$ is a basis vector. The randomization design must meet given quotas per treatment after all assignments are made, so that $\sum_{i=1}^{n} T_i(k) = np_k$ with $p_1, \ldots, p_K$ denoting the target probabilities.

Let τ_1 be the stopping rule of the form,

$$\tau_1 = \min\left\{j : \max_k \left[\sum_{i=1}^{j} T_i(k) - np_k\right] \geq 0\right\}. \tag{2.1}$$

In other words, at the instant τ_1 the first target is met, and we denote by ξ_1 the corresponding treatment, $\sum_{i=1}^{\tau_1} T_i(\xi_1) = np_{\xi_1}$. The allocation process continues until the moment τ_2, when some treatment ξ_2, $\xi_2 \neq \xi_1$, receives its quota of np_{ξ_2} subjects,

$$\tau_2 = \min\left\{j : j > \tau_1, \max_{k:k\neq\xi_1} \left[\sum_{i=1}^{j} T_i(k) - np_k\right] \geq 0\right\}.$$

After τ_2, when ξ_2 meets its target, the allocation is performed among the remaining $K - 2$ treatments and so on, until the moment τ_{K-1}, after which only one uncompleted treatment remains. This treatment, ξ_K, receives then the remaining subjects. These two random sequences $0 = \tau_0 < \tau_1 < \cdots < \tau_{K-1} < \tau_K = n$, and $\xi_1, \xi_2, \ldots, \xi_{K-1}, \xi_K$, are principal characteristics of a randomization design. It is convenient to view $\xi = (\xi_1, \xi_2, \ldots, \xi_K)$ as a permutation of the integers $(1, \ldots, K)$.

2.1. Truncated Multinomial Design

The truncated multinomial design Π_{TM} starts with the probability of assignment to the rth treatment being $\pi_r > 0$. After the instant τ_1, the probability of assignment

to each of the remaining $K-1$ treatments changes to $\pi_r/(1-\pi_{\xi_1})$, $r \neq \xi_1$, so that the allocation vectors T_j get this distribution over the available treatments. After τ_2, the allocation probabilities become $\pi_r/(1-\pi_{\xi_1}-\pi_{\xi_2})$, $r \neq \xi_1, \xi_2$, and so on, until the moment τ_{K-1}, when only one uncompleted treatment is left.

If $t_{\xi_i} = np_i$, $i = 1, \ldots, I$, $t_{\xi_i} < np_i$ for all other indices, and $\sum_k t_k = j$, i.e., if at stage j, exactly I treatments $\xi_1, \ldots, \xi_I$ are completed, then

$$P\left(T_{j+1} = \epsilon_r | \sum_{i=1}^{j} T_i = t\right) = \frac{\pi_{\xi_r}}{1-\pi_{\xi_1}-\cdots-\pi_{\xi_{I-1}}}, \quad r \neq \xi_1, \ldots, \xi_I.$$

To derive the joint probability distribution of random variables $\tau_1, \ldots, \tau_k$ and ξ, let

$$nb(x,t,p) = \binom{x-1}{t-1} p^t (1-p)^{x-t}, \quad x = t, \ t+1, \ldots$$

denote the negative binomial probabilities.

Lemma 2.1. *The joint probability distribution of random variables $\tau_1, \ldots, \tau_{K-1}$ and $\xi_1, \ldots, \xi_K$ has the form*

$$\begin{aligned} &P[\tau_1 = t_1, \tau_2 = t_2, \ldots, \tau_{K-1} = t_{K-1}, \xi = (\xi_1, \ldots, \xi_K)] \\ &= \prod_{I=1}^{K} nb\left[t_I - n(p_{\xi_1} + \cdots + p_{\xi_{I-1}}), np_{\xi_I}, \frac{\pi_{\xi_I}}{1-\pi_{\xi_1}-\cdots-\pi_{\xi_{I-1}}}\right] \\ &= \prod_{k=1}^{K-1} \binom{t_k - n(p_{\xi_1} + \cdots + p_{\xi_{k-1}}) - 1}{np_{\xi_k} - 1} \pi_{\xi_k}^{np_{\xi_k}} (1-\pi_{\xi_1}-\cdots-\pi_{\xi_k})^{t_k - t_{k+1}}. \quad (2.2) \end{aligned}$$

Here $np_{\xi_1} \leq t_1 \leq n-K+1$, $n(p_{\xi_1}+p_{\xi_2}) \leq t_2 \leq n-K+2, \ldots, n(p_{\xi_1}+\cdots+p_{\xi_{K-1}}) \leq t_{K-1} \leq n-1$, $0 < t_1 < t_2 < \cdots < t_{K-1} < t_K = n$.

Lemma 2.1 (as well as the following lemmas, Proposition 2.2, and Theorem 2.1) are established in Rukhin (2007c).

To obtain the joint probability generating function for the waiting times τ's, put $G_{(0)} = 0$, and denote by $G_{(1)} < G_{(2)} < \cdots < G_{(K)}$ the order statistics based on K independent random variables $G_1, \ldots, G_K$ from gamma-distributions, $\Gamma(np_r, 1/\pi_r)$, $r = 1, \ldots, K$, with the densities,

$$g_r(u) = \frac{\pi_r^{np_r}}{\Gamma(np_r)} u^{np_r - 1} e^{-\pi_r u}, \quad u > 0.$$

Proposition 2.1. *For Π_{TM} the probability generating function of the conditional distribution of $(\tau_1, \tau_2 - \tau_1, \ldots, \tau_{K-1} - \tau_{K-2})$ for given ξ has the form*

$$\begin{aligned} &E\left(z_1^{\tau_1} z_2^{\tau_2 - \tau_1} \cdots z_{K-1}^{\tau_{K-1} - \tau_{K-2}} \mid \xi\right) \\ &= E\left\{e^{\sum_{r=1}^{K} [G_{(r)} - G_{(r-1)}](\pi_{\xi_r} + \cdots + \pi_{\xi_K})(z_r^{-1} - 1)} \mid G_{(k)} = G_{\xi_k}, k = 1, \ldots, K\right\}, \quad (2.3) \end{aligned}$$

and

$$Ez_1^{\tau_1} z_2^{\tau_2 - \tau_1} \cdots z_{K-1}^{\tau_{K-1} - \tau_{K-2}} = E \exp\left\{\sum_{r=1}^{K} [G_{(r)} - G_{(r-1)}](\pi_{\psi_r} + \cdots + \pi_{\psi_K})(z_r^{-1} - 1)\right\},$$

where the permutation ψ is defined by $G_{(k)} = G_{\psi_k}$, $k = 1, \ldots, K$.

In the situation when $\pi_i \equiv p_i$ and the distribution of the allocation vectors does not change after quota fulfillments, Proposition 2.1 was established by Anderson et al. (1982). See also Holst (1986), who used embedding of independent multinomial trials into a continuous time Poisson process. The conditioning event in Proposition 2.1 means that the antirank of the sample $G_1, \ldots, G_K$ is ξ.

To formulate a limit theorem for $(\tau_1, \tau_2, \ldots, \tau_K)$, denote by $X_1, \ldots, X_K$, $X_k \sim N(0, 1/\pi_k)$, $k = 1, \ldots, K$, independent normal random variables and by $X_{(1)} < X_{(2)} < \cdots < X_{(K)}$ their order statistics. Also let

$$\widetilde{X} = \sum_{k=1}^{K} \pi_k X_k = \sum_{r=1}^{K} \pi_{\psi_r} X_{(r)},$$

$X_{(r)} = X_{\psi_r}$, $r = 1, \ldots, K$, i.e., ψ is the antirank of the sample $X_1, \ldots, X_K$.

Theorem 2.1. *When $\pi_r \equiv p_r$, and $n \to \infty$, the conditional distribution of $n^{-1/2}(\tau_2 - \tau_1, \ldots, \tau_{K-1} - \tau_{K-2}, n - \tau_{K-1})$ for given ξ converges to the conditional distribution of $[X_{(2)} - X_{(1)}](1 - p_{\xi_1}), \ldots, [X_{(K-1)} - X_{(K-2)}](1 - p_{\xi_1} - \cdots - p_{\xi_{K-2}})$, $[X_{(K)} - X_{(K-1)}](1 - p_{\xi_1} - \cdots - p_{\xi_{K-1}})$ given that the antirank of the sample $X_1, \ldots, X_K$ is ξ,*

$$P[\xi = (\xi_1, \ldots, \xi_K)] \to Q_\xi = P[X_{(1)} = X_{\xi_1}, X_{(2)} = X_{\xi_2}, \ldots, X_{(K)} = X_{\xi_K}]. \tag{2.4}$$

The distribution of $n^{-1/2}(\tau_2 - \tau_1, \ldots, \tau_{K-1} - \tau_{K-2}, n - \tau_{K-1})$ converges to that of $\{[X_{(2)} - X_{(1)}](1 - p_{\psi_1}), \ldots, [X_{(K)} - X_{(K-1)}](1 - p_{\psi_1} - \cdots - p_{\psi_{K-1}})\}$, where ψ is the antirank of the normal sample $X_1, \ldots, X_K$.

It follows that for fixed r, $r = 1, \ldots, K-1$, the asymptotic behavior of τ_r is determined by the weighted sum of spacings,

$$\frac{\tau_r - n}{\sqrt{n}} \sim \sum_{j=r+1}^{K} p_{\psi_j}[X_{(r)} - X_{(j)}] = X_{(r)} - \widetilde{X} + \sum_{j=1}^{r-1} p_{\psi_j}[X_{(j)} - X_{(r)}]. \tag{2.5}$$

In particular,

$$\tau_1 - n \sim \sqrt{n}[X_{(1)} - X],$$

and

$$n - \tau_{K-1} \sim \sqrt{n} p_{\psi_K}[X_{(K)} - X_{(K-1)}].$$

According to (2.4), the limiting joint distribution of completed treatments is that of the antirank of a (heterogeneous) normal sample,

$$Q_\xi = \prod p_r^{1/2} \int \cdots \int_{-\infty < y_1 < \cdots < y_{K-1} < y_K < \infty} \prod_r \varphi(\sqrt{p_{\xi_r}} y_r) dy_1 \cdots dy_K$$

with φ denoting the standard normal density. Let $[x]$ denote the integer part of x. The next results show how to express these probabilities in terms of only $K - 1$ order statistics from half-normal distributions.

Lemma 2.2. *Under conditions of Theorem* 2.1 *for any fixed* $w_1 \geq w_2 \geq \cdots \geq w_{K-1} \geq 0$, *as* $n \to \infty$,

$$\sqrt{n}\, P[\tau_1 = n - (w_1\sqrt{n}), \tau_2 = n - (w_2\sqrt{n}), \ldots, \tau_{K-1} = n - (w_{K-1}\sqrt{n}), \xi] \sim \prod_{r=1}^{K-1} \gamma_r \varphi(\gamma_r w_r),$$

where for $r = 1, \ldots, K-1$,

$$\gamma_r = \gamma_{\xi,r} \sqrt{\frac{p_{\xi_r}}{(1 - p_{\xi_1} - \cdots - p_{\xi_{r-1}})(1 - p_{\xi_1} - \cdots - p_{\xi_r})}}. \tag{2.6}$$

Proposition 2.2. *With* γ_r *defined by* (2.6),

$$Q_\xi = \int \cdots \int_{w_{K-1} \geq \cdots \geq w_1 \geq 0} \prod_{r=1}^{K-1} \gamma_r \varphi(\gamma_r w_r) dw_r.$$

Also $Q_\xi = Q_{(\xi_K, \xi_{K-1}, \ldots, \xi_1)}$, *and* $Q_\xi \leq 2^{1-K}$, *which is a sharp bound. The function* Q_ξ *is a convex function of* $\gamma_1^2, \ldots, \gamma_{K-1}^2$. *For* $K = 2$,

$$Q_\xi = 1/2,$$

when $K = 3$,

$$Q_\xi = \frac{1}{2\pi} \arctan\left(\sqrt{\frac{p_{\xi_2}}{p_{\xi_1} p_{\xi_3}}}\right) = \frac{1}{2\pi} \arctan\left(\frac{\gamma_{\xi,2}}{\gamma_{\xi,1}}\right),$$

and for $K = 4$,

$$Q_\xi = \frac{1}{8} - \frac{1}{4\pi} \arctan\left[\frac{(p_{\xi_1} + p_{\xi_2})\sqrt{p_{\xi_4}(1 - p_{\xi_1})} + (p_{\xi_3} + p_{\xi_4})\sqrt{p_{\xi_1}(1 - p_{\xi_4})}}{\sqrt{p_{\xi_2} p_{\xi_3}}}\right].$$

Notice that convexity of Q implies Schur-convexity of its symmetrized version (Marshall and Olkin, 1979, Ch. 3, C. 2), so that

$$\frac{1}{K!} \sum_\xi Q_\xi \geq \int \cdots \int_{w_{K-1} \geq \cdots \geq w_1 \geq 0} \prod_{r=1}^{K-1} \varphi(w_r) dw_r.$$

For $K \leq 6$, if $p_1 < p_2 < \cdots < p_K$, then the largest value of Q_ξ is attained at permutations $(1, 3, \ldots, K, K-1, \ldots, 4, 2)$ and $(2, 4, \ldots, K-1, K, \ldots, 3, 1)$, when K is odd, and at $(1, 3, \ldots, K-1, K, \ldots, 4, 2)$ and $(2, 4, \ldots, K, K-1, \ldots, 3, 1)$, when K is an even number. However, this phenomenon does not hold for larger K. For example when $K = 7$, the permutation $(1, 3, 4, 7, 6, 5, 2)$ can be more likely than $(1, 3, 5, 7, 6, 4, 2)$.

2.2. Random Allocation Rule

The random allocation rule Π_{RA} is described by the allocation vectors T_i, $i = 1, \ldots, n$, so that

$$p\left(\sum_{i=1}^{j} T_i = t\right) = \frac{\binom{j}{t_1 \ldots t_K}\binom{n-j}{np_1 - t_1 \ldots np_K - t_K}}{\binom{n}{np_1 \ldots np_K}} \tag{2.7}$$

for any nonnegative integer-valued vector $t = (t_1, \ldots, t_K)^T$, under condition $\sum_{k=1}^{K} t_k = j$. The conditional probability distribution of T_{j+1} is very simple,

$$P\left(T_{j+1} = \epsilon_r \mid \sum_{i=1}^{j} T_i = t\right) = \frac{np_r - t_r}{n - j}$$

with the initial distribution, $P(T_1 = \epsilon_r) = p_r$, $r = 1, \ldots, K$.

Lemma 2.3. *For Π_{RA} the joint probability distribution of the random variables $\tau_1, \ldots, \tau_{K-1}$ and $\xi = (\xi_1, \ldots, \xi_K)$ has the form*

$$\begin{aligned} P(\tau_1 = t_1, \tau_2 = t_2, \ldots, \tau_{K-1} = t_{K-1}, \xi) &= \frac{\prod_{I=1}^{K} \binom{t_I - n(p_{\xi_1} + \cdots + p_{\xi_{I-1}}) - 1}{np_{\xi_1} - 1}}{\binom{n}{np_1 \ldots np_I}} \\ &= \prod_{k=1}^{K-1} \frac{\left[\begin{smallmatrix} t_k - n(p_{\xi_1} + \cdots + p_{\xi_{k-1}}) - 1 \\ np_{\xi_k} - 1 \end{smallmatrix}\right]}{\left[\begin{smallmatrix} n(1 - p_{\xi_1} - \cdots - p_{\xi_{k-1}}) \\ np_{\xi_k} \end{smallmatrix}\right]} \end{aligned} \tag{2.8}$$

Here $0 < t_1 < t_2 < \cdots < t_{K-1} < t_K = n$, $np_{\xi_1} \le t_1 \le n - K + 1$, $n(p_{\xi_1} + p_{\xi_2}) \le t_2 \le n - K + 2, \ldots, n(p_{\xi_1} + \cdots + p_{\xi_{K-1}}) \le t_{K-1} \le n - 1$.

Lemma 2.4. *With $P_{t_1 \ldots t_{K-1} \xi}$ denoting the probabilities in* (2.8), *when $t_k = n - s_k$, $s_1 > s_2 > \cdots > s_{K-1} > s_K = 0$,*

$$\lim_{n \to \infty} P_{t_1 \ldots t_{K-1} \xi} = p_{\xi_1} \ldots p_{\xi_K} \prod_{k=1}^{K-1} (1 - p_{\xi_1} - \cdots - p_{\xi_k})^{s_k - s_{k+1} - 1}.$$

See Rukhin (2007c) for the proofs of these lemmas and of Theorem 2.2. These results show that for Π_{RA}, the joint limiting behavior of the instants $\tau_1, \ldots, \tau_K$ is quite different from these for Π_{TR}.

Theorem 2.2. *For Π_{RA} the conditional distribution of $(\tau_2 - \tau_1, \ldots, \tau_{K-1} - \tau_{K-2}, n - \tau_{K-1})$ for given ξ converges to that of a random vector $(H_1, H_2, \ldots, H_{K-1})$, where H_k, $k = 1, \ldots, K$, are independent geometric random variables with parameters $p_{\xi_1} + \cdots + p_{\xi_k}$,*

$$P(H_k = m) = (1 - p_{\xi_1} - \cdots - p_{\xi_k})^{m-1}(p_{\xi_1} + \cdots + p_{\xi_k}), \quad m = 1, 2, \ldots,$$

and

$$\lim_{n \to \infty} P[\xi = (\xi_1, \ldots, \xi_K)] = \frac{\prod_1^K p_k}{p_{\xi_1}(p_{\xi_1} + p_{\xi_2}) \cdots (p_{\xi_1} + p_{\xi_2} + \cdots + p_{\xi_{K-1}})}. \tag{2.9}$$

In particular, for a given ξ,

$$n - \tau_r \to H_r + \cdots + H_{K-1},$$

where H_k are independent geometric random variables from Theorem 2.2,

$$E(\tau_1 \mid \xi) \sim n - \sum_{k=1}^{K-1} \frac{1}{p_{\xi_1} + \cdots + p_{\xi_k}},$$

and

$$\mathrm{Var}(\tau_1 \mid \xi) \sim \sum_{k=1}^{K-1} \frac{p_{\xi_{k+1}} + \cdots + p_{\xi_K}}{(p_{\xi_1} + \cdots + p_{\xi_k})^2}.$$

The unconditional limiting distribution of τ_r can be easily derived from Theorem 2.2. The distribution of $n - \tau_1$ converges to that of a discrete random variable U such that for $m = K, K+1, \ldots$

$$P(U < m) = \sum_{j_1 \geq 1, \ldots, j_K \geq 1} \binom{m}{j_1 \cdots j_K} p_1^{j_1} \cdots p_K^{j_K}, \tag{2.10}$$

which is the probability that in m independent trials with K outcomes, whose probabilities are p_k, all outcomes occur (or the probability that all cells are occupied in the problem of distribution of m balls in K cells.) In other terms, U is the waiting time before all K outcomes occur. Feller (1968, Section IX.3) discusses the waiting time in sampling from the uniform distribution that leads to the the sum of independent geometric random variables. In this case,

$$P(U = m) = \frac{K!}{K^{m+1}} \left\{ \begin{matrix} m \\ K-1 \end{matrix} \right\}, \quad m = K-1, K, \ldots,$$

where $\left\{ \begin{matrix} m \\ K \end{matrix} \right\}$ denotes Stirling's number of the second kind; i.e., the number of partitions of an m element set into K nonempty subsets. In the general situation,

$$\begin{aligned} EU &= \sum_{k=0}^{K-1} (-1)^{K-k+1} \sum_{i_1 < i_2 < \cdots < i_k} \frac{1}{1 - p_{i_1} - \cdots - p_{i_k}} - 1 \\ &= \prod_1^K p_k \sum_{\xi} \frac{1}{p_{\xi_1}(p_{\xi_1} + p_{\xi_2}) \cdots (p_{\xi_1} + p_{\xi_2} + \cdots + p_{\xi_{K-1}})} \sum_{k=1}^{K-1} \frac{1}{p_{\xi_1} + \cdots + p_{\xi_k}}. \end{aligned}$$

Similar distributions appear in the classical occupancy problem (the coupons collection problem) and in more general waiting time problems (Blom et al., 1994). Direct verification of conditions of Theorem A.4 in Marshall and Olkin (1979) shows that (2.10) as a function of $p_1, \ldots, p_K$ is Schur-concave, so that U is stochastically larger for nonuniform probabilities than its version for the uniform distribution (i.e., unlike some of truncated multinomial designs, the first fulfillment instant occurs earlier for nonuniform allocations).

The probability distribution on permutations of $\{1, \ldots, K\}$ in (2.9) is that of the antirank of ordered independent exponential random variables with expected values

$1/p_k$, $k = 1, \ldots, K$ (or more generally of random variables with survival functions $F_k = F_0^{p_k}$, $k = 1, \ldots, K$ for a fixed absolutely continuous survival function F_0). Such models are discussed in Marden (1995). The distribution (2.9) is such that if $p_1 < p_2 < \cdots < p_K$, then $(1, 2, \ldots, K)$ is the likeliest permutation, while the least probable permutation is $(K, K-1, \ldots, 1)$, and this is in stark contrast with behavior of the probabilities Q_ξ discussed in Proposition 2.2.

3. SELECTION BIAS

The *selection bias* is defined by Blackwell and Hodges (1957) to be the expected value of the correct guesses number if the experimenter tries to assign subjects, say, with the largest expected response to a particular treatment. More precisely, if at stage i the experimenter makes his guess δ_i, $\delta_i \in \{1, \ldots, K\}$, while the statistician using a randomization design Π assigned the ith patient to the treatment ξ_i, the loss is 1, if $\delta_i = \xi_i$, and 0 otherwise. The total loss, $L(\delta, \Pi)$, is defined for the guessing strategy $\delta = (\delta_1, \ldots, \delta_n)$ of the experimenter and the randomization design Π of the statistician and is equal to the sum of stage-wise losses. Its expected value is then the selection bias of Π and δ.

If at stage j the transition probabilities are $p_j(t, t')$, the common guessing strategy $\delta^0 = (\delta_1^0, \ldots, \delta_n^0)$ of the experimenter is to make the decision according to the largest probability of this conditional distribution; i.e.,

$$\delta_j^0(t) = \arg\max_{t'} p_j(t, t').$$

Proposition 3.1. *In the notation of Theorem* 2.1 *for the truncated multinomial design,*

$$E[L(\delta^0, \Pi_{TM}) \mid \xi] = np_{(K)}\sqrt{n}\sum_{I=1}^{K}\max_{k\geq I} p_{\xi_k} E[X_{(I)} - X_{(I-1)} \mid \xi] + o(\sqrt{n}), \tag{3.11}$$

with $p_{(K)} = \max_k p_k$. *For the random allocation rule,*

$$E[L(\delta^0, \Pi_{RA}) \mid \xi] = np_{(K)} - \sum_{I=1}^{K-1} \frac{p_{(K)} - \max_{k\geq I+1} + p_{\xi_k}}{p_{\xi_1} + \cdots + p_{\xi_I}} + o(1). \tag{3.12}$$

When $p_k \equiv K^{-1}$, according to (3.11),

$$EL(\delta^0, \Pi_{TM}) = \frac{n}{K} + \frac{\sqrt{n}}{K} EX_{(K)} + o(\sqrt{n}),$$

and (3.12) shows that

$$EL(\delta^0, \Pi_{RA}) = \frac{n}{K} + o(1).$$

Under Π_{RA}, the maximum likelihood or convergence strategy, $\hat{\delta}$, is to guess at stage i the treatment k, for which the difference $np_k - \sum_{j=1}^{i} T_j(k)$ is the largest. This strategy maximizes the expected number of the correct guesses,

$$EL(\hat{\delta}, \Pi_{RA}) = \max_{\delta} EL(\delta, \Pi_{RA}).$$

Rukhin (2007b) proves Proposition 3.1 and shows that

$$EL(\hat{\delta}, \Pi_{RA}) = np_{(K)} + \frac{\pi\sqrt{np_{(K)}}}{2} EZ_{(q)} + o(\sqrt{n}), \tag{3.13}$$

where q is the number of ps equal to their maximum, $p_{(K)}$, and $Z_{(1)} < Z_{(2)} < \cdots < Z_{(q)}$ are the order statistics of a standard normal random sample of size q. For the equally likely allocations, $[q = K, p_{(K)} = 1/K]$,

$$EL(\hat{\delta}, \Pi_{RA}) = \frac{n}{K} + \frac{\pi\sqrt{n}}{2\sqrt{K}} EZ_{(K)} + o(\sqrt{n}).$$

Minimaxity of Π_{TR} proven by Blackwell and Hodges (1957) for $K = 2$ and equal probabilities may not hold for nonuniform allocations. Indeed, the comparison of (3.11) and (3.13) shows that if $q = 1$, Π_{TR} has larger selection bias than Π_{RA}, $EL(\delta^0, \Pi_{TM}) > EL(\hat{\delta}, \Pi_{RA})$ for large n. However, if $q = K$, this inequality is reverse.

As the function $\sqrt{n}$ is subadditive, Proposition 3.1 and (3.13) show that for large n, both $EL(\delta^0, \Pi_{TM})$ and $EL(\hat{\delta}, \Pi_{RA})$ as functions of n are subadditive. Therefore, if $n = n_1 + n_2$, and the allocation process is performed in two stages with n_1 subjects at the first stage and n_2 subjects at the second, then for each of our designs the bias obtained in one compound allocation process is smaller than the sum of selection biases corresponding to separate stages.

4. MOMENTS AND ACCIDENTAL BIAS

Here the form of the first two moments of the random allocation vectors T_i assigning subjects to treatments according to Π_{TR} or Π_{RA} is given. The expectations (let alone covariance structures) of these vectors in the nonuniform case are quite different. The results of this section are proven in Rukhin (2007a).

Denote by $\Xi = \mathrm{diag}(p)$, the diagonal matrix formed by the target probability vector p, and let the K-dimensional vector $a_{\xi_1 \ldots \xi_I}$ have coordinates $p_k/(1 - p_{\xi_1} - \cdots - p_{\xi_I})$ for $k \neq \xi_1, \ldots, \xi_I$, $I \geq 1$, with all other coordinates zero (when $I = 0$ this vector coincides with p.) For brevity, we write $P(A, \xi)$ or $P(A, \xi_1, \xi_2, \ldots, \xi_K)$ for $P[A, \xi = (\xi_1, \xi_2, \ldots, \xi_K)]$ and put $C_{i,j} = \mathrm{Cov}(T_i, T_j)$.

Assume, as in Theorem 2.1, that for Π_{TR}, $\pi_k \equiv p_k$. Somewhat surprisingly, for this design the probabilities of different treatments at stage i in general depend on that stage, $t_j = ET_j = \{P[T_j(1) = 1], \ldots, P[T_j(K) = 1]\}^T \neq p$, although, of course, $\sum_1^n t_j = np$.

Proposition 4.1. *Under* Π_{TR} *for any* $i = 1, \ldots, n$,

$$t_i = \sum_{I=0}^{K-1} \sum_{\xi_1, \ldots, \xi_I} P(\tau_I < i \leq \tau_{I+1}, \xi_1, \ldots, \xi_I) a_{\xi_1} \ldots \xi_I, \tag{4.14}$$

$$\mathrm{Var}(T_i) = \mathrm{diag}(t_i) - t_i t_i^T, \tag{4.15}$$

and for $i < j$,

$$C_{i,j} = \sum_{\substack{0 \leq I \leq J < K \\ \xi_1, \ldots, \xi_J}} P(\tau_I < i \leq \tau_{I+1}, \tau_J < j \leq \tau_{J+1}, \xi_1, \ldots, \xi_J) a_{\xi_1 \ldots \xi_I} a_{\xi_1 \ldots \xi_J}^T - t_i t_j^T. \tag{4.16}$$

For Π_{RA},

$$ET_i = p, \quad \mathrm{Var}(T_i) = \Xi - pp^T = \Sigma_0, \tag{4.17}$$

and for $i \neq j$,

$$C_{ij} = -\frac{1}{n-1}\Sigma_0.$$

For $x > 0$, when $n \to \infty$,

$$t_{n-[x\sqrt{n}]} \sim k(x) = P(\widetilde{X} - X_{(1)} < x)p + \sum_{I=1}^{K-1} \sum_{\xi} P\Bigg\{ \sum_{k=I+2}^{K} p_{\xi_k}[X_{(k)} - X_{(I+1)}] \leq x < \sum_{k=I+1}^{K} p_{\xi_k}[X_{(k)} - X_{(I)}], \xi \Bigg\} a_{\xi_1 \ldots \xi_I},$$

$$\mathrm{Var}[T_{n-(x\sqrt{n})}] \sim \mathrm{diag}[k(x)] - k(x)k(x)^T = L(x, x),$$

and for $0 < y < x$,

$$C_{n-(x\sqrt{n}),n-(y\sqrt{n})} \to L(x, y)$$
$$= \sum_{0 \leq I \leq J < K} \sum_{\sigma} P\Bigg\{ \sum_{k=I+2}^{K} p_{\sigma_k}[X_{(k)} - X_{(I+1)}] \leq x < \sum_{k=I+1}^{K} p_{\sigma_k}[X_{(k)} - X_{(I)}],$$
$$\sum_{k=J+2}^{K} p_{\xi_k}[X_{(k)} - X_{(J+1)}] \leq y < \sum_{k=J+1}^{K} p_{\xi_k}[X_{(k)} - X_{(J)}], \xi \Bigg\} a_{\xi_1 \ldots \xi_I} a^T_{\xi_1 \ldots \xi_J} - k(x)k(y).$$

Efron (1971) introduced the concept of *accidental bias*, which is a measure of the expected bias of the treatment effect arising from a linear regression model if important covariates are ignored. In our situation, if e_m denotes the m-dimensional vector with unit coordinates, let z be a nK-dimensional vector of (deterministic) covariates left out of the model such that $z^T e_{nK} = 0$. The degree of susceptibility to accidental bias of a random allocation vector T with the covariance matrix Σ is given by $E[z^T(T - ET)]^2 = z^T \Sigma z$.

Namely, define the accidental bias of pooled nK-allocation vector with the covariance matrix Σ as

$$\rho = \max_{z : z^T e_{nK} = 0} \frac{z^T \Sigma z}{z^T z};$$

i.e., ρ is the largest eigenvalue of $P_{nK} \Sigma P_{nK}$, where $P_m = I - e_m e_m^T/m$ denotes the projection onto the orthogonal subspace of the vectors with equal coordinates. In our balanced case, $\sum_1^n T_i = np$, $\Sigma e_{nK} = 0$, so that $\Sigma = P_{nK} \Sigma P_{nK}$. Thus, ρ coincides with the maximum eigenvalue $\lambda_{\max}(\Sigma)$.

For Π_{RA}, the eigenvalues of Σ are products of eigenvalues of the matrix $\frac{n}{n-1} P_n$ and eigenvalues of Σ_0. The latter are bounded from above by $p_{(K)}$, so that $\lambda_{\max}(\Sigma) \leq np_{(K)}/(n-1)$.

The next result summarizes the asymptotic behavior of these eigenvalues.

Proposition 4.2. *For* Π_{TR}, $\rho = O(n^{1/2})$. *More precisely, if* $\lambda_{\max}(\mathscr{L})$ *is the largest eigenvalue of the bounded integral operator* $\mathscr{L}$ *in the space of square integrable* K*-dimensional vector-functions on* $[0, \infty)$ *determined by the symmetric kernel* $L(x, y)$ *from Proposition* 4.1, *then*

$$\frac{\rho}{\sqrt{n}} = \frac{\lambda_{\max}(\Sigma)}{\sqrt{n}} \to \lambda_{\max}(\mathscr{L}).$$

For Π_{RA}, *the largest eigenvalue* $\lambda_{\max}(\Sigma)$ *is bounded,*

$$\rho = \lambda_{\max}(\Sigma) = \frac{n}{n-1}\lambda_{\max}(\Sigma_0) \le \frac{n}{n-1}p_{(K)}.$$

A conclusion is that the truncated multinomial design can be susceptible to a high degree of accidental bias. In contrast, for the random allocation rule the accidental bias is between $p_{(K)}$ and $p_{(K-1)}$ for large n.

5. STATISTICAL INFERENCE UNDER BALANCED RANDOMIZATION

Since the allocation vectors are not independent, standard inference procedures may not be appropriate. A permutation test is used when the sequence of subject responses is deterministic and only the sequence of treatment assignments is random. If a_{jn} is a score coefficient of subject j out of n subjects, $\sum_1^n a_{jn} = 0$, the basic form of the linear rank statistic employed in a permutation test is $\sum_1^n a_{jn}T_j$. Typical score functions include the simple ranks or van der Waerden scores. The asymptotic distribution of such permutation tests have been derived for various specialized randomization schemes (Rosenberger, 1993; Smythe, 1998; Smythe and Wei, 1983; Wei et al., 1986). Rosenberger and Lachin (2002, Section 7.5) discuss their use for the data from clinical trials when $K = 2$.

The permutation test is given by a vector rank statistic $\sum_{j=1}^n a_{jn}(T_j - t_j)$. For Π_{RA}, its asymptotic normality under a Lindeberg-type condition on the scores,

$$\max_{1\le j\le n} a_{jn}^2 / \sum_{j=1}^n a_{jn}^2 \to 0, \tag{5.18}$$

follows from Hajek et al. (1999, Section 6.1.5, Theorem 1). See details in Rosenberger and Lachin (2002, Section 14.3) when $K = 2$, $p_1 = p_2 = 1/2$.

Under Π_{TR} in the uniform case, $p_k = K^{-1}$, Rukhin (2007b) demonstrated that if for any $r = 1, \ldots, K-1$,

$$\lim_{n\to\infty} \frac{\sum_{i<j} a_{in}a_{jn}P(\tau_r < i)}{\sum_{j=1}^n a_{jn}^2} = 0, \tag{5.19}$$

then

$$\mathrm{Var}\left[\sum_{j=1}^n a_{jn}\left(T_j - \frac{1}{K}e_K\right)\right] \sim \sum_j a_{jn}^2\Sigma_0.$$

Here is an extension of Proposition 2 in Rosenberger and Rukhin (2002) and of Theorem 1 in Zhang and Rosenberger (2005) for $K = 2$ with the statistic L_n written in the form,

$$L_n = \left(\sum_j a_{jn}^2 \right)^{-1/2} \sum_{j=1}^{n} a_{jn} \left(T_j - \frac{1}{K} e_K \right). \tag{5.20}$$

Proposition 5.1. *Assume that when* $n \to \infty$, (5.18) *and* (5.19) *hold, and for a sequence* δ_n *such that* $\delta_n < n$ *and* $n^{-1/2}(n - \delta_n) \to \infty$,

$$\lim_{n \to \infty} \frac{\sum_{j \le \delta_n} a_{jn}^2}{\sum_{j=1}^{n} a_{jn}^2} = 1. \tag{5.21}$$

Then for Π_{TR} *the limiting distribution of* L_n *in* (5.20) *is a* K*-dimensional normal distribution* $N_K(0, \Sigma_0)$.

A similar result can be obtained for a nonparametric confidence interval for the contrast of treatment means (Wei et al., 1989).

6. EXAMPLE

In the psychophysical trials performed by the University Clinic of the Freiburg University (Germany) there were three scenarios (experimental conditions)

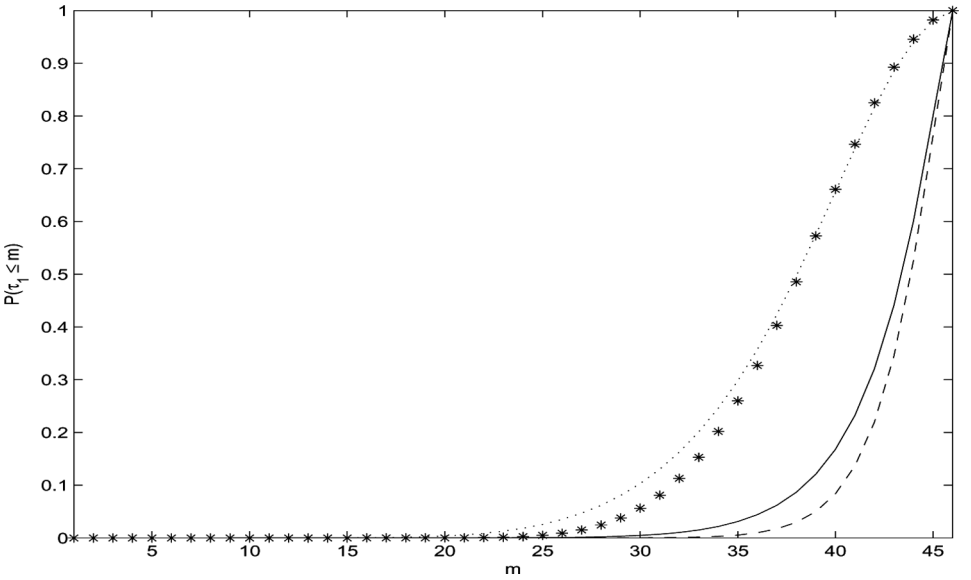

Figure 1. Plot of distribution functions of τ_1 when $K = 3$, $n = 48$, $p_1 = 10/48$, $p_2 = 15/48$, $p_3 = 23/48$ and for equal probabilities $p_i \equiv 1/3$. The solid line stands for Π_{RA}, the dotted line (:) for Π_{TR}, the line marked by $*$ corresponds to Π_{TR} with equally likely allocations, and the dashed (––) line for the uniform Π_{RA}.

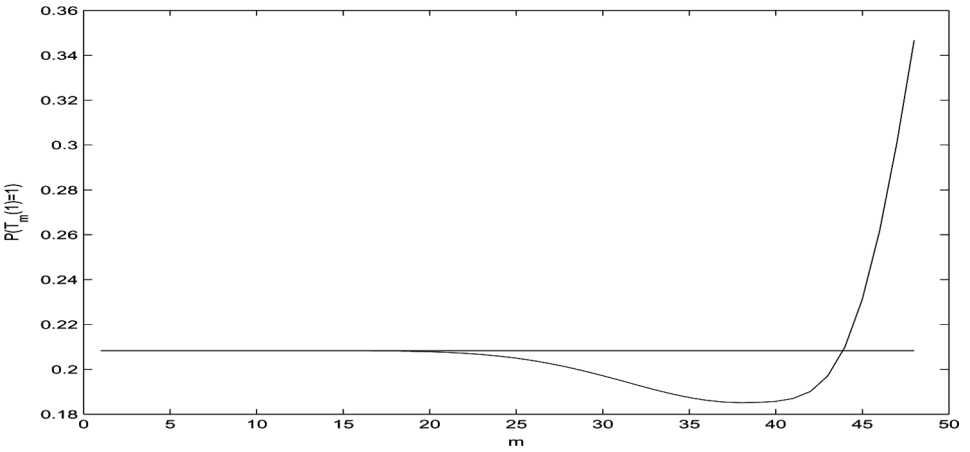

Figure 2. Plot of $P[T_m(1) = 1]$ when $K = 3$, $p_1 = 10/48$, $p_2 = 15/48$, $p_3 = 23/48$ and $n = 48$ for Π_{TR}. The horizontal line corresponds to p_1.

employing different generators of random numbers producing sounds which are made audible to a subject (Braeunig and Faul, 2006).

Two of these generators (Markov and Bernoulli, say, D_1 and D_2) produce really random numbers, but the third is deterministic and can be treated as a placebo (P). The treatment consisted in a random assignment of a patient to one of these three actions under restriction that a condition was not allowed to be followed by the same condition. Thus the total number of such scenarios was $3 \times 2 \times 2 = 12$, out of which assignments to a sequence (P, D_i, P) should be relatively rare.

It makes sense under these circumstances to assign a fairly small probability p_1 to any sequence of the form (P, D_i, P) with exactly two letters P in it, a somewhat larger probability to the sequences (D_i, P, D_j) with one letter P, and even larger probability to the sequences (D_i, D_j, D_i), $i = j$. Thus, it may be reasonable to

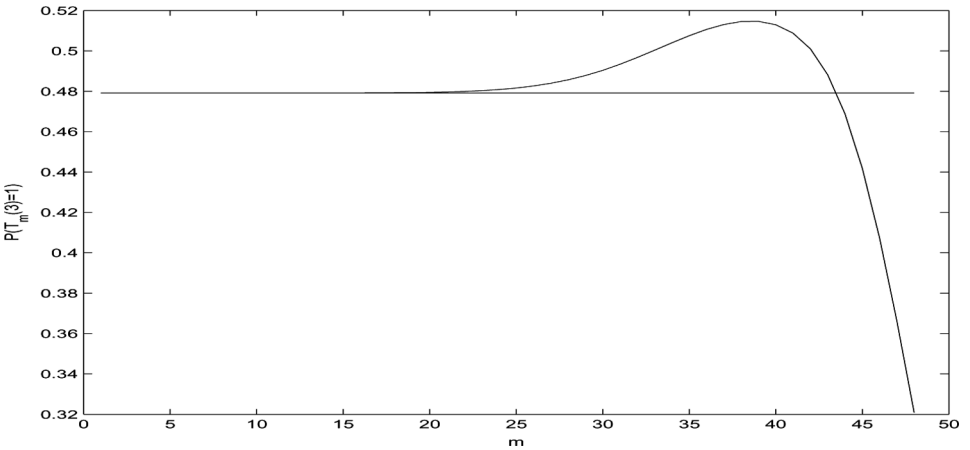

Figure 3. Plot of $P[T_m(3) = 1]$ for Π_{TR} as a function of m. The horizontal line corresponds to $p_3 = 23/48$.

Table 1. Exact and asymptotic distribution of ξ_1 for both designs

$P(\xi_i = k)$	Π_{TR}		Π_{RA}	
k	Exact	Asymptotic	Exact	Asymptotic
1	0.3775	0.3619	0.5053	0.5053
2	0.3309	0.3319	0.3178	0.3178
3	0.2916	0.3062	0.1769	0.1769

interpret a treatment as an assignment sequence with a given number (0, 1, or 2) of letters P in it. In the following numerical example involving Figures 1–3 and Table 1, $K = 3$, $p_1 = 10/48$, $p_2 = 15/48$, $p_3 = 23/48$ and $n = 48$.

Figure 1 shows graphs of distribution functions of τ_1 in this situation for both designs. Obviously this variable for the random allocation rule is stochastically larger than that for the truncated binomial design. For the random allocation rule under the uniform probabilities ($p_i \equiv 1/3$), the instant τ_1 is stochastically larger than for the nonuniform probabilities.

Table 1 contains exact and asymptotic probabilities of different values of ξ_1 (the first treatment to be completed) for both designs. Under each of them, the least frequently used treatment ($P = 2$) is the likeliest to be completed first, but under Π_{RA}, this probability is considerably larger than that for Π_{TR}. Also notice that the asymptotic distribution of ξ_1 is virtually identical to the exact distribution for Π_{RA} ($n = 48$). For the truncated multinomial design there is a substantial difference between asymptotic and exact distribution, and these facts hold more generally.

In some situations it is desirable to test the allocation sequence for randomness. Because of Proposition 4.1, the classical monobit test based on constancy of the assignment probabilities cannot be employed under Π_{TR}. Indeed, Figures 2 and 3 show rather irregular plots of the assignment probabilities, $P[T_m(1) = 1]$ and $P[T_m(3) = 1]$, under Π_{TR}. The probability $P[T_m(1) = 1]$ remains close to p_1 during the first half of the trials, drops to 0.18 at $m = 35$, and then jumps up to the value 0.34, while $P[T_m(3) = 1]$ increases to 0.52 at $m = 40$, when it drops down to 0.32. A patient who wants to be assigned to the treatments according to the given unlike probabilities p_i should come in the beginning of the trials, the one who desires a more uniform distribution in the end!

REFERENCES

Anderson, K., Sobel, M., and Uppuluri, R. R. (1982). Quota Fulfillment Times, *Canadian Journal of Statistics* 10: 73–88.

Blackwell, D. and Hodges, J. L. Jr. (1957). Design for the Control of Selection Bias, *Annals of Mathematical Statistics* 28: 449–460.

Blom, G., Holst, L., and Sandell, D. (1994). *Problems and Snapshots from the World of Probability*, New York: Springer-Verlag.

Braeunig, M. and Faul, T. (2006). T.REG—A Triggered Random Event Generator for Basic Research in Complementary and Alternative Medicine, http://kompmed.uniklinik-freiburg.de

Efron, B. (1971). Forcing a Sequential Experiment to be Balanced, *Biometrika* 58: 403–417.

Fang, K.-T. (1982). A Restricted Occupancy Problem, *Journal of Applied Probability* 19: 707–711.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol I, 3rd ed., New York: Wiley.

Flajolet, P., Gardy, D., and Thimonier, L. (1992). Birthday Paradox, Coupon Collectors, Caching Algorithms and Self-Organized Search, *Discrete Applied Mathematics* 39: 207–229.

Hajek, J., Sidak, Z., and Sen, P. K. (1999). *Theory of Rank Tests*, San Diego: Academic Press.

Holst, L. (1986). On Birthday, Collectors', Occupancy and Other Classical Urn Problems, *International Statistical Review* 54: 15–27.

Hu, F. and Rosenberger, W. F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*, New York: Wiley.

Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Applications. An Approach to Modern Discrete Probability Theory*, New York: Wiley.

Ling, K. D. (1992). A Generalization of the Sooner and Later Waiting Times for Bernoulli Trials: Frequency Quotas, *Statistics & Probability Letters* 14: 401–406.

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*, London: Chapman & Hall.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*, New York: Academic Press.

Peto, R. (1978). Clinical Trials Methodology, *Biomedicine* 28: 24–36.

Rosenberger, W. F. (1993). Asymptotic Inference with Response-Adaptive Treatment Allocation Designs, *Annals of Statistics* 21: 2098–2107.

Rosenberger, W. F. and Lachin, J. M. (2002). *Randomization in Clinical Trials: Theory and Practice*, New York: Wiley.

Rosenberger, W. F. and Rukhin, A. L. (2002). Bias Properties and Non-parametric Inference for Truncated Binomial Randomization, *Journal of Nonparametric Statistics* 15: 455–465.

Rukhin, A. L. (2004). Limiting Distributions in Sequential Occupancy Problems, *Sequential Analysis* 23: 141–158.

Rukhin, A. L. (2006). Gamma-Distribution Order Statistics, Maximal Multinomial Frequency and Randomization Designs, *Journal of Statistical Planning and Inference* 136: 2213–2226.

Rukhin, A. L. (2007a). Randomization with Non-Uniform Allocations: Fulfillment Time Distributions and Bias Properties, *Statistics* 41: 11–30.

Rukhin, A. L. (2007b). Nonparametric Inference for Balanced Randomization Designs, *Journal of Statistical Planning and Inference* 137: 967–984.

Rukhin, A. L. (2007c). Normal Order Statistics and Sums of Geometric Random Variables in Treatment Allocation Problems, *Statistics & Probability Letters* 77: 1312–1321.

Smythe, R. T. (1998). Conditional Inference for Balanced Randomization Designs, *Annals of Statistics* 16: 1155–1161.

Smythe, R. T. and Wei, L. J. (1983). Significance Tests with Balanced Randomization Design, *Biometrika* 70: 496–500.

Sobel, M. and Ebneshahrashoob, M. (1992). Quota Sampling for Multinomial via Dirichlet, *Journal of Statistical Planning and Inference* 33: 157–164.

Stigler, S. M. (1969). The Use of Random Allocation for the Control of Selection Bias, *Biometrika* 59: 553–560.

Wei, L. J. (1978). On the Random Allocation Design for the Control of Selection Bias in Sequential Experiments, *Biometrika* 65: 79–84.

Wei, L. J., Smythe, R. T., and Smith, R. L. (1986). K-treatment Comparisons with Restricted Randomization Rules in Clinical Trials, *Annals of Statistics* 14: 265–274.

Wei, L. J., Smythe, R. T., and Mehta, C. R. (1989). Interval Estimation with Restricted Randomization Rules, *Biometrika* 76: 363–368.

Zhang, Y. and Rosenberger, W. F. (2005). On Linear Rank Tests for Truncated Binomial Randomization, *Statistics & Probability Letters* 72: 83–92.