The Annals of Statistics
2002, Vol. 30, No. 1, 122–139

ASYMPTOTIC PROPERTIES OF ADAPTIVE DESIGNS FOR CLINICAL TRIALS WITH DELAYED RESPONSE

BY Z. D. BAI[1], FEIFANG HU[1] AND WILLIAM F. ROSENBERGER[2]

National University of Singapore, University of Virginia and University of Maryland

For adaptive clinical trials using a generalized Friedman's urn design, we derive the limiting distribution of the urn composition under staggered entry and delayed response. The stochastic delay mechanism is assumed to depend on both the treatment assigned and the patient's response. A very general setup is employed with K treatments and L responses. When $L = K = 2$, one example of a generalized Friedman's urn design is the randomized play-the-winner rule. An application of this rule occurred in a clinical trial of depression, which had staggered entry and delayed response. We show that maximum likelihood estimators from such a trial have the usual asymptotic properties.

1. Preliminaries.

1.1. *Introduction.* Adaptive designs for clinical trials use sequentially accruing outcome data to dynamically update the probability of assignment to one of two or more treatments. The idea is to skew these probabilities to favor the treatment that has been the most effective thus far in the trial, thus making the randomization strategy more attractive to physicians and their patients than standard equal allocation. A typical probability model for adaptive clinical trials is the generalized Friedman's urn model [cf. Athreya and Karlin (1968)]. Initially, a vector $\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{1K})$ of balls of type $1, \ldots, K$ is placed in an "urn" (computer generated). Patients sequentially enter the trial. When a patient is ready to be randomized, a ball is drawn at random and replaced. If it was type i, the ith treatment is assigned. We then wait for a random variable ξ (whose probability distribution depends on i) to be observed. An additional d_{ij} balls are added to the urn of type $j = 1, \ldots, K$, where $d_{ij}(\xi)$ is some function on the sample space of ξ. The algorithm is repeated through n stages.

Let $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nK})$ be the urn composition when the nth patient arrives to be randomized. Then the probability that the patient will be randomized to

Received February 2000; revised September 2001.

[1]Supported by a grant from the National University of Singapore.

[2]Supported by Grant R29-DK51017-05 from the National Institute of Diabetes and Digestive and Kidney Diseases.

AMS 2000 subject classification. 62G10.

Key words and phrases. Generalized Friedman's urn, martingales, randomization, randomized play-the-winner rule, staggered entry, treatment allocation, urn models.

treatment j is given by $Y_{nj}/|\mathbf{Y}_n|$, where $|\mathbf{Y}_n| = \sum_{i=1}^{K} Y_{ni}$. Let $\mathbf{D}(\xi) = ((d_{ij}(\xi), i, j = 1, \ldots, K))$. First-order asymptotics for the generalized Friedman's urn are determined by the generating matrix of the urn, given by $\mathbf{H} = E\{\mathbf{D}(\xi)\}$. Provided $\mathbf{H}$ is positive regular and $\Pr\{d_{ij} = 0 \ \forall j\} = 0$ for all i,

$$\frac{Y_{nj}}{|\mathbf{Y}_n|} \to v_j \quad \text{almost surely,} \tag{1}$$

$j = 1, \ldots, K$, where $\mathbf{v} = (v_1, \ldots, v_K)$ is the normalized left eigenvector corresponding to the maximal eigenvalue of $\mathbf{H}$ [cf. Athreya and Karlin (1968)].

As a simple example, ξ might be the primary outcome of a clinical trial, such as death or cure. Assuming that $\mathbf{Y}_1$ is deterministic, let $d_{ij} = (K - 1)\delta_{ij}$ if cure on treatment i, and $d_{ij} = (1 - \delta_{ij})$ if death on treatment i, where δ_{ij} is the Kronecker delta. Assuming that ξ is immediately observable after the patient is randomized, we have $|\mathbf{Y}_n| = |\mathbf{Y}_1| + (K - 1)(n - 1)$. When $K = 2$, this is the randomized play-the-winner rule of Wei and Durham (1978), which has been used occasionally in clinical trials [see, e.g., Bartlett, Roloff, Cornell, Andrews, Dillon and Zwischenberger (1985) and Tamura, Faries, Andersen and Heiligenstein (1994)]. Wei, Smythe, Lin and Park (1990) gave a simple probability model for the randomized play-the-winner rule, letting p_1 be the probability of success on treatment 1 and p_2 be the probability of success on treatment 2. Under this model, by (1),

$$\frac{Y_{n1}}{Y_{11} + Y_{12} + n - 1} \to \frac{1 - p_2}{2 - p_1 - p_2} \quad \text{almost surely}$$

[Rosenberger (1996), page 140] and hence the rule allocates according to the relative risk of failure on treatment 2 versus treatment 1. Wei (1979) first proposed using the generalized Friedman's urn to develop a broad class of allocation rules for clinical trials. The generalized Friedman's urn also has been used in other medical applications [see, e.g., Rosenberger (1996) and Rosenberger and Grill (1997)].

Typically, clinical trials do not result in immediate outcomes, and urn models are simply not appropriate for today's oft-performed long-term survival trials, where outcomes may not be ascertainable for many years. However, there are many trials where many or most outcomes are available during the recruitment period, even though individual patient outcomes may not be immediately available prior to the randomization of the next patient. Consequently, the urn can be updated when outcomes become available, and this does not involve any additional logistical complexities. Wei (1988) suggested such updating for the randomized play-the-winner rule and introduced an indicator function, δ_{jk}, $j < k$, that takes the value 1 if the response of patient j occurs before patient k is randomized and 0 otherwise. He did not explore its properties. Later, Bandyopadhyay and Biswas (1996) explored properties of a simple probability model, which assumes that $P(\delta_{jk} = 1)$ is a constant depending only on the lag $k - j$, for a modified version of the randomized play-the-winner rule. Delays in response can slow the adaptation process considerably, and simulation studies show that the expected allocation

proportions generated by the design are more conservative than if outcomes are immediately ascertainable [see Rosenberger (1999)]. However, the adaptive nature of the design still accomplishes its purpose: more patients, on average, are assigned to the better treatment. In practice, time to response in clinical trials can depend on the treatment assigned and the response observed. Heretofore, what has not been known is how such delayed response with staggered entry affects the limiting distribution of the urn composition given by (1).

In this paper, we verify that stochastic staggered entry and delay mechanisms do not affect the limiting distribution of the urn for a wide class of designs defined by the generalized Friedman's urn. We then show that the maximum likelihood estimators have the usual asymptotic properties. This extends the work of Rosenberger, Flournoy and Durham (1997), who investigated properties of maximum likelihood estimators from a generalized Friedman's urn design with immediate response. In our proofs, we assume that patients arrive sequentially and their arrival process has independent increments, and that time to response has a distribution that depends on both the treatment assigned and the patient's response. We investigate a very general adaptive randomization scheme with K treatments and L outcomes, based on the generalized Friedman's urn model.

1.2. *Motivating example.* Tamura, Faries, Andersen and Heiligenstein (1994) described an adaptive placebo-controlled clinical trial of fluoxetine in depression. Patients were stratified by shortened or normal rapid eye movement latency (REML) and then were randomized according to the randomized play-the-winner rule. Separate urns were used in the REML strata. The outcome on which the adaptive randomization was based is a 50% reduction in the Hamilton Depression Scale ($HAMD_{17}$) score in two consecutive visits after at least three weeks of therapy. This outcome was obviously not ascertainable immediately, and hence the urn was updated based on the available data. Exploratory data analysis [Rosenberger and Hu (1999)] indicated that entry time was approximately uniformly distributed over a 270-day time interval. Time to response was similar among the two REML strata and two treatment groups, but differed according to patient outcomes. Time to response was approximately normal with mean 43 days and variance 122 days; in nonresponders, time to determination of nonresponse was approximately uniform on the interval (20 days, 75 days).

When there is immediate response, Wei, Smythe, Lin and Park (1990) showed that the usual maximum likelihood estimator of the treatment effect from a clinical trial employing the randomized play-the-winner rule has the usual asymptotic properties, namely, consistency and asymptotic normality. However, this result is not applicable to clinical trials with delayed response, such as the fluoxetine trial; hence, we have the motivation for this paper.

REMARK 1. The fluoxetine trial had a number of subtleties that necessitated nonstandard analyses; in particular, the outcome on which the adaptation was

based was a surrogate outcome for the true primary outcome. We note that Tamura, Faries, Andersen and Heiligenstein (1994) simulated the joint distribution of these outcomes and the time delay in order to make inferences about the treatment effect, and they also performed a Bayesian analysis. They found that fluoxetine was modestly effective in the shortened REML stratum and not effective in the normal REML stratum. The results of our paper suggest a simple alternative analysis, but only for the surrogate outcome, based on the asymptotic distribution of the treatment effect. However, the accuracy of such a test may be appropriately questioned since there were approximately 40 patients in each stratum. With large numbers of patients, enumerating or simulating the exact distribution of the test statistic may be computationally intensive; in which case the asymptotic distribution given in this paper may be attractive.

1.3. *Organization of the paper.* In Section 2, we derive the limiting distribution of the urn composition under delayed response for a multi-armed trial. In Theorem 1, we prove that the urn composition, $\mathbf{Y}_n$, suitably normalized, tends to the same limit as in (1). In Theorem 2, we show that the urn composition tends to a multivariate normal distribution in law, and we give a form of the variance–covariance matrix. The main assumption of the theorems is that the delay cannot be very large relative to the patient entry stream. The important observation is made that the limiting distribution does not depend on the delay mechanism, but, in practice, the delay mechanism must be taken into account in estimating the variance–covariance matrix. In Section 3, we derive the full likelihood and show that the usual asymptotic inference results can be applied to data arising from a generalized Friedman's urn design when there is staggered entry and delayed response. We conclude our paper with some observations in Section 4. Proofs are relegated to the Appendix.

2. Asymptotic properties of the urn composition. We assume a multinomial response model with responses $\xi_n = l$ if patient n had response l, $l = 1, \ldots, L$. Let J_n be the treatment indicator for the nth patient, that is, $J_n = j$ if patient n was randomized to treatment $j = 1, \ldots, K$, and let $\mathbf{X}_n = (X_{n1}, \ldots, X_{nK})$ satisfy $X_{nJ_n} = 1$ and all other elements 0. We assume that the entry time of the nth patient is t_n, where $\{t_n - t_{n-1}\}$ are i.i.d. for all n. The response time of the nth patient is denoted by $\tau_n(j, l)$, which has distribution g_{jl}, $j = 1, \ldots, K$, $l = 1, \ldots, L$, for all n, so that the distribution of the response times can depend on both the treatment assigned and the response observed. For the nth patient randomized to treatment j, we define an indicator function $M_{jl}(n, m)$ as follows: $M_{jl}(n, m) = 0$ if $\xi_n \neq l$ and $M_{jl}(n, m) = 1_{[\text{response time} \in (t_{n+m}, t_{n+m+1})]}$ if $\xi_n = l$. We assume for $n = 1, 2, \ldots$ that, given j, $\{M_{j\xi_n}(n, m)\}$ are i.i.d. By definition, for every pair of n and j, there is only one pair (l, m) such that $M_{jl}(n, m) = 1$ and $M_{jl'}(n, m') = 0$ for all $(l, m) \neq (l', m')$. We can define $\mu_{jlm} = E\{M_{jl}(n, m)\}$ as the probability that a patient on treatment j with response l will respond after m

more patients are enrolled and before $m+1$ more patients are enrolled. Thus we have

$$\sum_{l,m} \mu_{jlm} = 1 \quad \text{for } j = 1, \ldots, K.$$

For patient n, after observing $\xi_n = l,\ J_n = i$, we add $d_{ij}(\xi_n = l)$ balls of type j to the urn, where the total number of balls added at each stage is constant; that is, $\sum_{j=1}^{K} d_{ij}(\xi_n) = \beta$, where $\beta > 0$. Without loss of generality, we can assume $\beta = 1$; otherwise, we can consider the sequence $\{\mathbf{Y}_n/\beta\}$ instead. Note that the number of balls added to the urn does not have to be an integer, as in the models of Andersen, Faries and Tamura (1994). Let $\mathbf{D}(l) = ((d_{ij}(\xi_n = l),\ i, j = 1, \ldots, K))$.

REMARK 2. It is possible to generalize our results to the case in which the total number of balls added at each stage is random, provided that the expected number of balls added is a positive constant.

For given n and m, if $M_{jl}(n, m) = 1$, then we add balls at the $(n+m)$th stage according to the rule $\mathbf{X}_n\mathbf{D}(l)$. $\mathbf{X}_n$ contains the randomness in J_n, and $\mathbf{D}(l)$ contains the randomness in ξ_n, conditioned on J_n. We can now write a recursive formula for the urn composition,

$$\mathbf{Y}_n = \mathbf{Y}_{n-1} + \mathbf{W}_n,$$

where $\mathbf{W}_n$ is the number of balls of each type added at the nth stage, given by

$$\mathbf{W}_n = \sum_{m=0}^{n-2} M_{J_{n-m-1},\xi_{n-m-1}}(n-m-1, m)\mathbf{X}_{n-m-1}\mathbf{D}(\xi_{n-m-1}).$$

Denote by $\mathcal{F}_n$ the sigma algebra generated by $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ and let $E_n\{\cdot\} = E\{\cdot|\mathcal{F}_n\}$. We have

$$\begin{aligned} &E_{n-1}\{M_{J_{n-m-1},\xi_{n-m-1}}(n-m-1, m)\mathbf{X}_{n-m-1}\mathbf{D}(\xi_{n-m-1})\} \\ &\quad = \sum_{l=1}^{L} \frac{\mathbf{Y}_{n-m-1}}{|\mathbf{Y}_{n-m-1}|}\boldsymbol{\mu}_{lm}\mathbf{D}(l), \end{aligned}$$

where $\boldsymbol{\mu}_{lm}$ is a $K \times K$ diagonal matrix with the jth diagonal element μ_{jlm}. Then

$$E_{n-1}\{\mathbf{W}_n\} = \sum_{m=0}^{n-2} \frac{\mathbf{Y}_{n-m-1}}{|\mathbf{Y}_{n-m-1}|}\left(\sum_{l=1}^{L} \boldsymbol{\mu}_{lm}\mathbf{D}(l)\right).$$

It turns out that it is easier to work with the recursive formula

$$\mathbf{Y}_n = \mathbf{Y}_{n-1} + E_{n-1}\{\mathbf{W}_n\} + (\mathbf{W}_n - E_{n-1}\{\mathbf{W}_n\}).$$

Setting $\mathbf{Q}_n = \mathbf{W}_n - E_{n-1}\{\mathbf{W}_n\}$, we obtain the recursive formula

$$\mathbf{Y}_n = \mathbf{Y}_{n-1} + \sum_{m=0}^{n-2} \frac{\mathbf{Y}_{n-m-1}}{|\mathbf{Y}_{n-m-1}|}\left(\sum_{l=1}^{L} \boldsymbol{\mu}_{lm}\mathbf{D}(l)\right) + \mathbf{Q}_n. \tag{2}$$

We will use (2) as the pivotal recursion formula to prove asymptotic properties of $\mathbf{Y}_n$. But first we will require the following assumptions:

ASSUMPTION 1. For some $c \in (0, 1]$,

$$\sum_{i=m}^{\infty} \mu_{jli} = o(m^{-c}) \qquad \forall j, l. \tag{3}$$

REMARK 3. Assumption 1 implies that the probability that at least m additional patients will arrive prior to a patient's response is of order $o(m^{-c})$. Hence, in practical examples, the delay cannot be very large relative to the entry stream. In practice, it is convenient to verify this assumption by examining the time-to-response variable $\tau_n(j, l)$ and the entry times t_n. If (i) $E[\tau_n(j, l)]^{c_1} < \infty$ for each j, l and $c_1 > c$ and (ii) $E(t_i - t_{i-1}) > 0$ and $E(t_i - t_{i-1})^2 < \infty$, then Assumption 1 is satisfied. This is because

$$\mu_{jlm} = P\{\tau_n(j, l) \in (t_{n+m}, t_{n+m+1})\} = P\{\tau(j, l) \in (S_m, S_{m+1})\},$$

where $S_m = \sum_{i=1}^{m}(t_i - t_{i-1})$ $(t_0 = 0)$. Then

$$\sum_{i=m}^{\infty} \mu_{jli} = P\{\tau(j, l) \in (S_m, \infty)\}.$$

Since $S_m/m \to E(t_i - t_{i-1}) = E(t_1)$ almost surely as $m \to \infty$,

$$P\{\tau(j, l) \in (S_m, \infty)\} \le P\{\tau(j, l) \in (mE(t_1)/2, \infty)\} + P(S_m \le mE(t_1)/2).$$

By the Markov inequality, we have

$$\begin{aligned} P\{\tau(j, l) \in (mE(t_1)/2, \infty)\} &= P\{\tau(j, l) > mE(t_1)/2\} \\ &\le E[\tau(j, l)]^{c_1}/(mE(t_1)/2)^{c_1} \\ &= O(m^{-c_1}) = o(m^{-c}) \end{aligned}$$

and

$$P(S_m \le mE(t_1)/2) = P(S_m - ES_m \le -mE(t_1)/2) \le O(m^{-1}).$$

Consequently, Assumption 1 is not very stringent.

ASSUMPTION 2. Using the notation in Section 1, let $\mathbf{H} = E(\mathbf{D})$ and let $\mathbf{v}$ be the normalized left eigenvector of $\mathbf{H}$ corresponding to its maximal eigenvalue. Assume that $\mathbf{H}$ has the following Jordan decomposition:

$$\mathbf{T}^{-1}\mathbf{H}\mathbf{T} = \operatorname{diag}[1, \mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_s],$$

where $\mathbf{\Psi}_t$ is a $\nu_t \times \nu_t$ matrix (defining ν_t to be the block size of the Jordan form), given by

$$\mathbf{\Psi}_t = \begin{bmatrix} \lambda_t & 1 & 0 & \cdots & 0 \\ 0 & \lambda_t & 1 & \cdots & 0 \\ 0 & 0 & \lambda_t & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_t \end{bmatrix}.$$

We may select the matrix $\mathbf{T}$ so that its first column is $\mathbf{1}'$ and the first row of $\mathbf{T}^{-1}$ is $\mathbf{v}$. Let $\lambda = \max\{\mathrm{Re}(\lambda_1), \ldots, \mathrm{Re}(\lambda_s)\}$ and $\nu = \max_j\{\nu_j \text{ such that } \mathrm{Re}(\lambda_j) = \lambda\}$, where Re( ) is the real part of the eigenvalue.

THEOREM 1. *Under Assumptions* 1 *and* 2, *if* $c > 0$ *and* $\lambda < 1$, *then* $\mathbf{Y}_n/|\mathbf{Y}_n| \to \mathbf{v}$ *almost surely*.

PROOF. See the Appendix.

We can extend Theorem 1 to apply not only to the urn composition, but also to the sample fractions assigned to each treatment. Let $\mathbf{N}_n = (N_{n1}, \ldots, N_{nK})$, where N_{nj} is the number of patients assigned to treatment j, $j = 1, \ldots, K$, after n stages.

COROLLARY 1. *Under the assumptions of Theorem* 1, $\mathbf{N}_n/n \to \mathbf{v}$ *almost surely*.

PROOF. See the Appendix.

We now give the central limit result.

THEOREM 2. *Under Assumptions* 1 *and* 2, *for* $c > 1/2$ *and* $\lambda < 1/2$, *we have* $n^{1/2}(\mathbf{Y}_n/|\mathbf{Y}_n| - \mathbf{v})$ *converges in law to* $N(0, \mathbf{\Sigma})$, *where the form of* $\mathbf{\Sigma}$ *is given in* (22).

PROOF. See the Appendix.

REMARK 4. If $\lambda = 1/2$, the asymptotic normality holds, but with a different norming, given by $n \log^{2\nu-1} n$. In this case, we can derive $\mathbf{\Sigma}$ using techniques similar to those in the proof of Theorem 2.

REMARK 5. Because $\mathbf{\Sigma}$ depends on $\sum_{m=0}^{\infty}\mu_{jlm} = \Pr\{\xi_n = l | J_n = j\}$ through (21), we see that $\mathbf{\Sigma}$ does not depend on the delay mechanism. But this is a limiting result. In practice, we need to estimate $\mathbf{\Sigma}$ using (19) and (22), and the estimate will involve the delayed-response mechanism, $M_{jl}(n,m)$. We can estimate $\mathbf{\Sigma}$ in practice using the following procedure:

(i) Estimate $\mathbf{H}$ by

$$\hat{\mathbf{H}} = \frac{\sum_{i=2}^{n}\sum_{m=0}^{i-2} M_{J_{i-m-1},\xi_{i-m-1}}(i-m-1,m)\operatorname{diag}(\mathbf{X}_{i-m-1})\mathbf{D}(\xi_{i-m-1})}{\sum_{i=2}^{n}\sum_{m=0}^{i-2} M_{J_{i-m-1},\xi_{i-m-1}}(i-m-1,m)},$$

where $M_{J_{i-m-1},\xi_{i-m-1}}(i-m-1,m)$, $\mathbf{X}_{i-m-1}$ and $\mathbf{D}(\xi_{i-m-1})$ are observed during the trial.

(ii) Estimate $\mathbf{B}_{ni}$ by

$$\hat{\mathbf{B}}_{ni} = \prod_{j=i+1}^{n}(\mathbf{I} + j^{-1}\hat{\mathbf{H}}).$$

(iii) Estimate $\mathbf{\Sigma}$ by

$$\hat{\mathbf{\Sigma}} = \bigl(\mathbf{I} - (\mathbf{Y}_n'/|\mathbf{Y}_n|)\mathbf{l}\bigr)\left[n^{-1}\sum_{i=1}^{n}\hat{\mathbf{B}}_{ni}'(\mathbf{W}_i - \bar{\mathbf{W}})'(\mathbf{W}_i - \bar{\mathbf{W}})\hat{\mathbf{B}}_{ni}\right]\bigl(\mathbf{I} - \mathbf{l}'\mathbf{Y}_n/|\mathbf{Y}_n|\bigr),$$

where $\bar{\mathbf{W}} = n^{-1}\sum_{i=1}^{n}\mathbf{W}_i$. The $\mathbf{W}_i$ are the number of balls added to the urn at stage i, which are observed during the trial.

REMARK 6. For the sample fractions assigned to each treatment, we know that

$$n^{-1/2}(\mathbf{N}_n/n - \mathbf{v}) = n^{-1/2}\sum_{i=1}^{n}\bigl[\mathbf{X}_i - E(X_i|\mathcal{F}_i)\bigr] \tag{4}$$

$$+ n^{1/2}\sum_{i=1}^{n}\bigl(\mathbf{Y}_i/|\mathbf{Y}_i| - \mathbf{v}\bigr). \tag{5}$$

The asymptotic normality of the first term on the right-hand side of (4) follows from a multivariate version of the martingale central limit theorem. However, we still have not derived the asymptotic distribution of (5) and the correlation between the two terms, and we leave this as an additional research topic. Smythe (1996) proved the asymptotic joint normality of the sample fractions for the generalized Friedman's urn with immediate updating of the urn.

3. Likelihood results. Let $\mathbf{Y}^n = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ be the history of the urn composition, where $\mathbf{Y}_i$ is defined in Section 1. Let $\mathbf{J}^n = (J_1, \ldots, J_n)$ be the history of treatment assignments, $\boldsymbol{\xi}^n = (\xi_1, \ldots, \xi_n)$ be the history of patient responses,

$\boldsymbol{\tau}^n = (\tau_1, \ldots, \tau_n)$ be the history of response times and $\mathbf{t}^n = (t_1, \ldots, t_n)$ be the history of entry times. Then the full likelihood is given by

$$\begin{aligned}
\mathcal{L}_n &= \mathcal{L}(\boldsymbol{\tau}^n, \boldsymbol{\xi}^n, \mathbf{J}^n, \mathbf{Y}^n, \mathbf{t}^n) \\
&= \mathcal{L}(\tau_n | \boldsymbol{\tau}^{n-1}, \boldsymbol{\xi}^n, \mathbf{J}^n, \mathbf{Y}^n, \mathbf{t}^n) \mathcal{L}(\xi_n | \boldsymbol{\tau}^{n-1}, \boldsymbol{\xi}^{n-1}, \mathbf{J}^n, \mathbf{Y}^n, \mathbf{t}^n) \\
&\quad \times \mathcal{L}(J_n | \boldsymbol{\tau}^{n-1}, \boldsymbol{\xi}^{n-1}, \mathbf{J}^{n-1}, \mathbf{Y}^n, \mathbf{t}^n) \mathcal{L}(\mathbf{Y}_n | \boldsymbol{\tau}^{n-1}, \boldsymbol{\xi}^{n-1}, \mathbf{J}^{n-1}, \mathbf{Y}^{n-1}, \mathbf{t}^n) \\
&\quad \times \mathcal{L}(t_n | \boldsymbol{\tau}^{n-1}, \boldsymbol{\xi}^{n-1}, \mathbf{J}^{n-1}, \mathbf{Y}^{n-1}, \mathbf{t}^{n-1}) \mathcal{L}_{n-1} \\
&= \mathcal{L}(\tau_n | J_n, \xi_n) \mathcal{L}(\xi_n | J_n) \mathcal{L}(J_n | \mathbf{Y}_n) \mathcal{L}(t_n) \mathcal{L}_{n-1} \\
&= \prod_{i=1}^{n} \mathcal{L}(\tau_i | J_i, \xi_i) \mathcal{L}(\xi_i | J_i) \mathcal{L}(J_i | \mathbf{Y}_i) \mathcal{L}(t_i) \\
&\propto \prod_{i=1}^{n} \mathcal{L}(\xi_i | J_i).
\end{aligned}$$

Note that the allocation proportions are random and, together with treatment responses, form a sufficient statistic, unlike in the i.i.d. case with fixed allocation.

For the problem we have formulated, we have a product multinomial likelihood with $p_{jl} = \Pr\{\xi_n = l | J_n = j\}$ for all n, and $j = 1, \ldots, K$, $l = 1, \ldots, L-1$ and $p_{jL} = 1 - p_{j1} - \cdots - p_{j,L-1}$. Standard martingale techniques can be used to prove the consistency and asymptotic normality of the maximum likelihood estimators $\hat{p}_{jl}$ from this likelihood. Rosenberger, Flournoy and Durham (1997) gave a convenient set of sufficient conditions. In our case, only their condition (A3) is nontrival. Using their notation, let $L_i = \log(\mathcal{L}_i / \mathcal{L}_{i-1})$, where $\mathcal{L}_0 = 1$. Then condition (A3) requires

$$-n^{-1} \sum_{i=1}^{n} E_{i-1} \left\{ \frac{\partial^2 L_i}{\partial p_{jl} \partial p_{km}} \right\} \to \gamma_{jklm} \quad \text{almost surely,} \tag{6}$$

where γ_{jklm} is a constant, $j, k = 1, \ldots, K$, $l, m = 1, \ldots, L-1$. Using the multinomial likelihood, it is easy to show that the left-hand side of (6) is 0 when $j \neq k$ and, for $j = k$, is given by

$$\left(\frac{1}{p_{jl}} + \frac{1}{p_{jL}} \right) n^{-1} \sum_{i=1}^{n} \frac{Y_{ij}}{|\mathbf{Y}_i|}, \qquad l = m \tag{7}$$

and

$$\frac{1}{p_{jL}} n^{-1} \sum_{i=1}^{n} \frac{Y_{ij}}{|\mathbf{Y}_i|}, \qquad l \neq m. \tag{8}$$

From Theorem 1, (7) converges almost surely to $v_j / p_{jl} + v_j / p_{jL}$ and (8) converges almost surely to v_j / p_{jL}. Hence,

$$\boldsymbol{\Gamma} = \big((\gamma_{jklm}) \big) = \frac{v_j}{p_{jl}} \mathbf{I} + \frac{v_j}{p_{jL}} \mathbf{J},$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{J} = \mathbf{1}\mathbf{1}'$. Then, by the theorem of Rosenberger, Flournoy and Durham (1997), page 71, we obtain the following result:

THEOREM 3. *For fixed $j = 1, \ldots, K$, the vector with components*

$$\left(\left(n^{1/2}\{\hat{p}_{jl} - p_{jl}\}\right)\right)_{l=1}^{L-1}$$

is asymptotically multivariate normal with mean vector $\mathbf{0}$ and variance–covariance matrix

$$\mathbf{\Gamma}^{-1} = \frac{p_{jl}}{v_j}\mathbf{I} - \frac{p_{jl}^2}{v_j(p_{jL} + (L-1)p_{jl})}\mathbf{J}.$$

Moreover, the K vectors are asymptotically independent.

Consequently, the usual asymptotic χ^2 tests can be used to investigate the treatment effect. For $K = L = 2$, we can use standard Z tests of the simple difference of proportions or the odds ratio.

4. Conclusions. Results on the asymptotic properties of the generalized Friedman's urn when there is a stochastic delay in updating the urn are interesting in their own right, from a probabilistic perspective. But the main contribution of this paper is in showing that randomized clinical trials using the generalized Friedman's urn for randomization can now use standard maximum likelihood estimation following the trial, under the standard clinical trial conditions of staggered entry and delayed response. We have also demonstrated, in Remark 3, that the assumptions on the entry stream and delay mechanism are typically not stringent.

We have not examined properties of estimators in this paper. For example, the joint distribution of sufficient statistics could be used to develop inferential tests, as an alternative to maximum likelihood. It would be interesting to develop several types of estimators and compare their efficiencies under different delay mechanisms, but we will leave that topic for future research.

Finally, asymptotic theory is becoming less important in this age of rapid algorithms for computing exact distributions. Hardwick and Stout (1998) performed seminal work in this area for adaptive designs and generally found samples as large as $n = 75$ to be amenable to exact computations, using parallel processing and networking algorithms. How one would implement such algorithms with a stochastic delay mechanism may be an interesting topic for further research. The third author has had some success with simulating the distribution of test statistics for adaptive designs with delayed response, using priority queues [see, e.g., Rosenberger and Seshaiyer (1997)]. However, the computational simplicity of an asymptotic normal test based on the maximum likelihood estimator, we presume, will always make it an attractive tool, and, in this paper, we have provided the necessary theory to justify its use.

APPENDIX

Because of the delayed response, the total number of balls in the urn at each stage will be a random variable, depending on which patients have already responded. To prove Theorem 1, we will need to take care of this complication, which we do in the following lemma:

LEMMA 1. (i) *For the total urn composition*, $\mathbf{Y}_n$, $n^{-1}|\mathbf{Y}_n| \to 1$ *in probability*. (ii) *If Assumption* 1 *is true*,

$$n^{-1}|\mathbf{Y}_n| = 1 + o_p(n^{-c'}) \quad \textit{for any } c' < c$$

and

$$n^{-1}|\mathbf{Y}_n| = 1 + o(n^{-c'}) \quad \textit{almost surely for any } c' < c/2,$$

where the constant c is defined in Assumption 1.

PROOF. Recall that we have assumed, without loss of generality, that the number of balls added to the urn at each stage is 1. Also, assume, without loss of generality, that $|\mathbf{Y}_1| = 1$. Then the number of balls at stage n will be n minus the balls not added due to a patient's nonresponse by stage n. We can write this mathematically as

$$|\mathbf{Y}_n| = n - \sum_{m=1}^{n-1} \sum_{i=n-m}^{\infty} M_{J_m,\xi_m}(m,i), \tag{9}$$

by noting that

$$\sum_{i=0}^{\infty} M_{J_m,\xi_m}(m,i) = 1.$$

Now, since

$$\sum_{i=m}^{\infty} \mu_{jli} \to 0 \ [(= o(m^{-c}) \text{ under Assumption 1}] \qquad \text{as } m \to \infty,$$

we have

$$\begin{aligned} &E\left(\sum_{m=1}^{n-1} \sum_{i=n-m}^{\infty} M_{J_m,\xi_m}(m,i) \right) \\ &\quad = \sum_{m=1}^{n-1} \sum_{i=n-m}^{\infty} E\{M_{J_m,\xi_m}(m,i)\} = \sum_{m=1}^{n-1} \sum_{i=n-m}^{\infty} \sum_{j=1}^{K} \sum_{l=1}^{L} \mu_{jli} \\ &\quad = \begin{cases} o(n), & \text{without Assumption 1,} \\ o(n^{1-c}), & \text{under Assumption 1 and } 0 < c < 1, \\ o(\log n), & \text{under Assumption 1 and } c = 1. \end{cases} \end{aligned}$$

This proves conclusion (i) and the first part of conclusion (ii) of Lemma 1 by the Markov inequality.

Now, choose ρ such that $\rho(c-c') > 1$ and $\rho c' < 1$ ($c' < c/2$). Define $n_k = [k^\rho]$, where $[\cdot]$ is the greatest integer function. Then, for any $\varepsilon > 0$,

$$\begin{aligned} P\big(n_k^{-1+c'}\big||\mathbf{Y}_{n_k}| - n_k\big| \geq \varepsilon\big) &\leq \varepsilon^{-1} n_k^{-1+c'} E\big||\mathbf{Y}_{n_k}| - n_k\big| \\ &\leq \varepsilon^{-1} n_k^{-1+c'} n_k^{1-c)} \log n_k \\ &\leq C k^{-\rho(c-c')} \log k \end{aligned}$$

for some constant C. Note that the right-hand side of the preceding inequality is summable. Thus, by the Borel–Cantelli lemma,

$$n_k^{-1+c'}(|\mathbf{Y}_{n_k}| - n_k) \to 0 \tag{10}$$

almost surely. To complete the proof of the lemma, we need to show that

$$\max_{n_{k-1} < n \leq n_k} \big|n^{-1+c'}(|\mathbf{Y}_n| - n)\big| \to 0$$

almost surely. It is easy to see that

$$|\mathbf{Y}_{n_{k-1}}| - n_{k-1} - (n_k - n_{k-1}) \leq |\mathbf{Y}_n| - n \leq |\mathbf{Y}_{n_k}| - n_k + (n_k - n_{k-1}).$$

From (10), we have

$$n^{-1+c'}(|\mathbf{Y}_{n_{k-1}}| - n_{k-1}) \to 0 \quad \text{and} \quad n^{-1+c'}(|\mathbf{Y}_{n_k}| - n_k) \to 0$$

almost surely. By the selection of ρ,

$$n^{-1+c'}(n_k - n_{k-1}) \leq n_{k-1}^{-1+c'} C n_k k^{-1} \leq C k^{\rho c' - 1} \to 0.$$

Therefore, we have proved the second part of (ii). $\square$

PROOF OF THEOREM 1. From (2), we have

$$\mathbf{Y}_n = \mathbf{Y}_{n-1} + \sum_{m=1}^{n-1} \frac{\mathbf{Y}_m}{m} \mathbf{G}(n-m-1) + \mathbf{Q}_n + \mathbf{R}_n, \tag{11}$$

where

$$\mathbf{G}(m) = \sum_{l=1}^{L} \boldsymbol{\mu}_{lm} \mathbf{D}(l)$$

and

$$\mathbf{R}_n = \sum_{m=1}^{n-1} \frac{\mathbf{Y}_m(m - |\mathbf{Y}_m|)}{m|\mathbf{Y}_m|} \mathbf{G}(n-m-1).$$

Recalling the definition of $\mathbf{Q}_n$ from (2), we can derive the following using (11) [letting $\mathbf{G}(-1)$ be the identity matrix $\mathbf{I}$ for convenience]:

$$\begin{aligned}\mathbf{Y}_n &= \mathbf{Y}_1 + \sum_{i=2}^{n}\sum_{m=1}^{i-1}\frac{\mathbf{Y}_m}{m}\mathbf{G}(i-m-1) + \sum_{i=2}^{n}(\mathbf{Q}_i+\mathbf{R}_i)\\ &= \sum_{m=1}^{n-1}\sum_{i=m}^{n-1}\frac{\mathbf{Y}_m}{m}\mathbf{G}(i-m) + \sum_{i=1}^{n}(\mathbf{Q}_i+\mathbf{R}_i) + \mathbf{Y}_1 \qquad (12)\\ &= \sum_{m=1}^{n-1}\frac{\mathbf{Y}_m}{m}\sum_{i=0}^{n-m-1}\mathbf{G}(i) + \sum_{i=1}^{n}(\mathbf{Q}_i+\mathbf{R}_i) + \mathbf{Y}_1.\end{aligned}$$

Here $\mathbf{Q}_1 = \mathbf{R}_1 = 0$. By the definition of $\mathbf{G}(m)$, we have

$$\sum_{m=0}^{\infty}\mathbf{G}(m) = \sum_{m=0}^{\infty}\sum_{l=1}^{L}\boldsymbol{\mu}_{lm}\mathbf{D}(l) = \mathbf{H},$$

recalling that $\mathbf{H} = E(\mathbf{D})$. Then, by (12), we have

$$\mathbf{Y}_n = \sum_{m=1}^{n-1}\frac{\mathbf{Y}_m\mathbf{H}}{m} + \sum_{i=1}^{n}\mathbf{Q}_i + \mathbf{Y}_1 + \mathbf{R}_n^{(1)}, \qquad (13)$$

where

$$\mathbf{R}_n^{(1)} = \sum_{i=1}^{n}\mathbf{R}_i - \sum_{m=1}^{n-1}\frac{\mathbf{Y}_m}{m}\sum_{i=n-m}^{\infty}\mathbf{G}(i).$$

Under condition (3) and by the result of Lemma 1, we further have

$$\mathbf{R}_n^{(1)} = \sum_{m=1}^{n-1}\frac{\mathbf{Y}_m(m-|\mathbf{Y}_m|)}{m|\mathbf{Y}_m|}\sum_{i=0}^{n-m-1}\mathbf{G}(i) - \sum_{m=1}^{n-1}\frac{\mathbf{Y}_m}{m}\sum_{i=n-m}^{\infty}\mathbf{G}(i) = o_p(n^{1-c'}),$$

and, by Lemma 1, we can strengthen this to almost sure convergence. From (13), we have

$$\mathbf{Y}_n = \mathbf{Y}_{n-1}\big(\mathbf{I} + (n-1)^{-1}\mathbf{H}\big) + \mathbf{Q}_n + \mathbf{R}_n^{(2)},$$

where

$$\mathbf{R}_n^{(2)} = \mathbf{R}_n^{(1)} - \mathbf{R}_{n-1}^{(1)} = o(n^{1-c'}).$$

Furthermore,

$$\mathbf{Y}_n = \sum_{i=2}^{n}\mathbf{Q}_i\mathbf{B}_{ni} + \mathbf{Y}_1\mathbf{B}_{n1} + \sum_{i=2}^{n}\mathbf{R}_i^{(2)}\mathbf{B}_{ni}, \qquad (14)$$

where

$$\mathbf{B}_{ni} = \prod_{j=i+1}^{n}\big(\mathbf{I} + (j-1)^{-1}\mathbf{H}\big)$$

(with the convention that $\mathbf{B}_{nn}$ is the identity matrix $\mathbf{I}$).

Let $\mathbf{z}_n = n^{-1}\mathbf{Y}_n\mathbf{T}$, where $\mathbf{T}$ is defined in Assumption 2. We wish to show that $\mathbf{z}_n$ converges to $(1, 0, 0, \ldots, 0)$ almost surely, which then implies that

$$n^{-1}\mathbf{Y}_n \to (1, 0, 0, \ldots, 0)\mathbf{T}^{-1} = \mathbf{v} \quad \text{almost surely.} \tag{15}$$

We have already shown that the first element of $\mathbf{z}_n$ converges almost surely to 1. Hence, we can focus on $\mathbf{z}_n\mathbf{E}$, where $\mathbf{E}' = [\mathbf{0} : \mathbf{I}]_{K-1\times K}$. Then, from (14), we obtain

$$\mathbf{z}_n\mathbf{E} = n^{-1}\sum_{i=2}^{n}\mathbf{Q}_i\mathbf{B}_{ni}\mathbf{T}\mathbf{E} + n^{-1}\mathbf{Y}_1\mathbf{B}_{n1}\mathbf{T}\mathbf{E} + n^{-1}\sum_{i=2}^{n}\mathbf{R}_i^{(2)}\mathbf{B}_{ni}\mathbf{T}\mathbf{E}. \tag{16}$$

Let $\tilde{\mathbf{B}}_{ni} = \mathbf{T}^{-1}\mathbf{B}_{ni}\mathbf{T}$. The second term of (16) can be written as $n^{-1}\mathbf{z}_1\tilde{\mathbf{B}}_{n1}\mathbf{E}$, which converges almost surely to 0, as $n^{-1}\mathbf{z}_1\tilde{\mathbf{B}}_{n1}$ converges almost surely to $(z_{11}, 0, \ldots, 0)$, where z_{11} is the first element of $\mathbf{z}_1$. This is because $n^{-1}\tilde{\mathbf{B}}_{n1}$ converges to $(1, 0, \ldots, 0)'(1, 0, \ldots, 0)$ under the condition $\lambda < 1$.

The third term of (16) requires a careful analysis. We write

$$\begin{aligned} n^{-1}\sum_{i=2}^{n}\mathbf{R}_i^{(2)}\mathbf{B}_{ni}\mathbf{T}\mathbf{E} &= n^{-1}\mathbf{R}_n^{(1)}\mathbf{B}_{nn}\mathbf{T}\mathbf{E} + n^{-1}\sum_{i=2}^{n-1}\mathbf{R}_i^{(1)}(\mathbf{H}/i)\mathbf{B}_{n,i+1}\mathbf{T}\mathbf{E} \\ &\quad - n^{-1}\mathbf{R}_1^{(1)}\mathbf{B}_{n2}\mathbf{T}\mathbf{E}. \end{aligned} \tag{17}$$

The first term of (17) is $o(n^{-c'})$. We can write the second term of (17) as

$$n^{-1}\sum_{i=2}^{n-1}\mathbf{R}_i^{(1)}(\mathbf{H}/i)\mathbf{T}\tilde{\mathbf{B}}_{n,i+1}\mathbf{E}.$$

For the analysis of $\tilde{\mathbf{B}}_{n,i+1}$, recalling the definitions of λ and $\mathbf{\Psi}_t$ in Assumption 2, we see that, for $\varepsilon > 0$,

$$n^{-(\lambda+\varepsilon)}\prod_{j=i+1}^{n}(\mathbf{I} + j^{-1}\mathbf{\Psi}_t) = o(i^{-\lambda}).$$

Consequently, $\tilde{\mathbf{B}}_{ni}$ is of order $o(n^{\lambda+\varepsilon}/i^{\lambda})$, and the second term is of order $o(n^{-c'+\varepsilon})$ if $\lambda + c' < 1$ and $o(n^{\lambda+\varepsilon-1})$ if $\lambda + c' > 1$. Finally, the third term of (17) can be written as

$$-n^{-1}\mathbf{R}_1^{(1)}\mathbf{T}\tilde{\mathbf{B}}_{n2}\mathbf{E},$$

and this term is $o(n^{\lambda+\varepsilon-1})$.

To complete the analysis of (16), we inspect the first term. The variance of the jth ($j > 1$) term is given by

$$n^{-2}\sum_{i=1}^{n}\mathrm{Var}(\mathbf{Q}_i\mathbf{B}_{ni}\mathbf{T}\mathbf{e}_j') = n^{-2}\sum_{i=1}^{n}\mathbf{e}_j'\mathbf{T}^*\mathbf{B}_{ni}'E(\mathbf{Q}_i'\mathbf{Q}_i)\mathbf{B}_{ni}\mathbf{T}\mathbf{e}_j', \tag{18}$$

where $\mathbf{e}_j'$ is the jth column of $\mathbf{E}$ and $\mathbf{T}^*$ is the conjugate transpose of $\mathbf{T}$. The conditional expectation of $\mathbf{Q}_n'\mathbf{Q}_n$ is given by

$$
\begin{aligned}
E_{n-1}(\mathbf{Q}_n'\mathbf{Q}_n) &= E_{n-1}(\mathbf{W}_n'\mathbf{W}_n) - \big(E_{n-1}(\mathbf{W}_n)\big)'\big(E_{n-1}(\mathbf{W}_n)\big) \\
&= \sum_{m=0}^{n-2}\sum_{l=1}^{L}\mathbf{D}(l)'\mathrm{Diag}\bigg(\frac{Y_{n-m-1,1}}{|\mathbf{Y}_{n-m-1}|}\mu_{1lm},\ldots,\frac{Y_{n-m-1,K}}{|\mathbf{Y}_{n-m-1}|}\mu_{Klm}\bigg)\mathbf{D}(l) \\
&\quad - \big(E_{n-1}(\mathbf{W}_n)\big)'\big(E_{n-1}(\mathbf{W}_n)\big),
\end{aligned}
\tag{19}
$$

and hence $E(\mathbf{Q}_i'\mathbf{Q}_i)$ is bounded. [Equation (19) will be important in determining the variance–covariance structure of the limiting distribution, derived in the proof of Theorem 2.] Since the elements of $\tilde{\mathbf{B}}_{ni}\mathbf{e}_j'$ are controlled by $n^{\lambda+\varepsilon}/i^{\lambda}$, from the developments above, we can see that (18) is $o(n^{\lambda+2\varepsilon-1})$. Because ε can be made small, using the Chebyshev inequality, we conclude that

$$
\Pr\bigg(\bigg|n^{-1}\sum_{i=1}^{n}\mathbf{Q}_i\mathbf{B}_{ni}\mathbf{T}\mathbf{e}_j'\bigg| \geq \delta\bigg) \leq Cn^{-b}
$$

for some constant C and $b > 0$.

Now we define $n_k = [k^{\rho}]$, where ρ satisfies $b\rho > 1$. For the subsequence n_k, the first term of (16) converges almost surely to 0 by the Borel–Cantelli lemma. Hence, for $c' > 0$, and by choosing ε small, the terms of $\mathbf{z}_{n_k}\mathbf{E}$ converge almost surely to 0, and (15) holds on the subsequence n_k, implying that $\mathbf{Y}_{n_k}/n_k$ converges almost surely to $\mathbf{v}$. Applying the subsequence method (use the monotonicity of $\mathbf{Y}_n$), $\mathbf{Y}_n/n$ converges almost surely to $\mathbf{v}$. Then, under Assumption 1 and by Lemma 1, $\mathbf{Y}_n/|\mathbf{Y}_n|$ converges almost surely to $\mathbf{v}$. □

PROOF OF COROLLARY 1. We can write

$$
\begin{aligned}
\mathbf{N}_n = \mathbf{N}_{n-1} + \mathbf{X}_n &= \sum_{i=1}^{n}\mathbf{X}_i \\
&= \sum_{i=1}^{n}\big[\mathbf{X}_i - E(\mathbf{X}_i|\mathcal{F}_i)\big] + \sum_{i=1}^{n}E(\mathbf{X}_i|\mathcal{F}_i) \\
&= \sum_{i=1}^{n}\big[\mathbf{X}_i - E(\mathbf{X}_i|\mathcal{F}_i)\big] + \sum_{i=1}^{n}\mathbf{Y}_i/|\mathbf{Y}_i|.
\end{aligned}
$$

Then

$$
\frac{\mathbf{N}_n}{n} = \frac{1}{n}\sum_{i=1}^{n}\big[\mathbf{X}_i - E(\mathbf{X}_i|\mathcal{F}_i)\big] + \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i/|\mathbf{Y}_i|.
$$

From the martingale strong law [e.g., Theorem 2.18 of Hall and Heyde (1980), page 36], the first term converges to 0 almost surely. The second term

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i/|\mathbf{Y}_i| \to \mathbf{v}$$

almost surely, which follows directly from Theorem 1. □

PROOF OF THEOREM 2. From Lemma 1, we have that

$$n^{1/2}(|\mathbf{Y}_n| - n) \to 0 \quad \text{in probability.}$$

Define $\mathbf{z}_n^{\#} = (n^{-1}\mathbf{Y}_n - \mathbf{v})\mathbf{T}$. We wish to examine the limit of $n^{1/2}\mathbf{z}_n^{\#}$. First, by Lemma 1, the first term converges almost surely to 0. We recall that $\mathbf{E}' = [\mathbf{0} : \mathbf{I}]_{K-1\times K}$ and, since $\mathbf{v}$ is the first row of $\mathbf{T}^{-1}$, we have $\mathbf{vT} = (1, 0, \ldots, 0)$, and $\mathbf{vTE} = 0$. Consequently, from (16),

$$\begin{aligned} n^{1/2}\mathbf{z}_n^{\#}\mathbf{E} &= n^{-1/2}\sum_{i=2}^{n}\mathbf{Q}_i\mathbf{B}_{ni}\mathbf{TE} + n^{-1/2}\mathbf{Y}_1\mathbf{B}_{n1}\mathbf{TE} \\ &\quad + n^{-1/2}\sum_{i=2}^{n}\mathbf{R}_i^{(2)}\mathbf{B}_{ni}\mathbf{TE}. \end{aligned} \tag{20}$$

The third term of (20), $n^{-1/2}\sum_{i=2}^{n}\mathbf{R}_i^{(2)}\mathbf{B}_{ni}\mathbf{TE}$, can be decomposed into three components, as in (17). From the proof of Theorem 1, it follows that the first term is of order $o_p(n^{-c_1+1/2})$, so it tends to 0 for $c > c_1 > 1/2$. The second term is $o_p(n^{-c_1+\varepsilon+1/2})$, and as ε is arbitrarily small, this term tends to 0. The third term is $o_p(n^{\lambda+\varepsilon-1/2})$, and as $\lambda < 1/2$, this also tends to 0.

From the proof of Theorem 1, $\tilde{\mathbf{B}}_{n1}\mathbf{E}$ is of order $o(n^{\lambda+\varepsilon})$, so that the second term of (20), $n^{-1/2}\mathbf{Y}_1\mathbf{B}_{n1}\mathbf{TE} = o(n^{\lambda+\varepsilon-1/2})$, also tends to 0.

Finally, we can use the martingale central limit theorem [e.g., Corollary 3.1 of Hall and Heyde (1980), page 58] to show that the first term of (20)

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{Q}_i\mathbf{B}_{ni}\mathbf{TE} \to N(0, \mathbf{\Sigma}_1) \quad \text{in law.}$$

The form of $\mathbf{\Sigma}_1$ can be obtained by careful derivation, but it is quite messy. Using the same techniques as Bai and Hu (1999), we can derive an exact expression for $\mathbf{\Sigma}_1$. It is given by $\mathbf{\Sigma}_1 = ((\mathbf{\Sigma}_{gh},\ g, h, = 1, \ldots, s))$, where $\mathbf{\Sigma}_{gh}$ is a submatrix with (a, b) element

$$\sum_{a'=0}^{a-1}\sum_{b'=0}^{b-1}\frac{(a'+b')!}{a'!b'!(1-\lambda_g-\bar{\lambda}_h)^{a'+b'+1}}\left[\mathbf{T}_g^{*}E_{\infty}(\mathbf{Q}'\mathbf{Q})\mathbf{T}_h\right]_{a-a',b-b'}$$

($\mathbf{T}^*$ is the conjugate transpose of $\mathbf{T}$ and $\bar{\lambda}$ is the complex conjugate of λ), where $E_\infty(\mathbf{Q}'\mathbf{Q}) = \lim_{n\to\infty} E_{n-1}(\mathbf{Q}_n'\mathbf{Q}_n)$. From (19),

$$E_\infty(\mathbf{Q}'\mathbf{Q}) = \sum_{l=1}^{L} \mathbf{D}(l)' \mathrm{Diag}\left(v_1 \sum_{m=0}^{\infty} \mu_{1lm}, \ldots, v_K \sum_{m=0}^{\infty} \mu_{Klm} \right) \mathbf{D}(l) - \mathbf{v}'\mathbf{v}. \tag{21}$$

Finally, we obtain

$$\mathbf{\Sigma} = (\mathbf{T}^{-1})^* \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{\Sigma}_1 \end{bmatrix} \mathbf{T}^{-1}. \tag{22}$$

□

Acknowledgments. Professor Rosenberger is also affiliated with the Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine. His research was done while visiting the Department of Statistics and Applied Probability, National University of Singapore. He thanks the Department for its hospitality and support. Professor Hu is also affiliated with Department of Statistics and Applied Probability, National University of Singapore. Special thanks go to the referees and Associate Editor for their constructive comments, which led to a much improved version of the paper.

REFERENCES

ANDERSEN, J., FARIES, D. and TAMURA, R. (1994). A randomized play-the-winner design for multi-arm clinical trials. *Comm. Statist. Theory Methods* **23** 309–323.

ATHREYA, K. B. and KARLIN, S. (1968). Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *Ann. Math. Statist.* **39** 1801–1817.

BAI, Z. D. and HU, F. (1999). Asymptotic theorems for urn models with nonhomogeneous generating matrices. *Stochastic Process. Appl.* **80** 87–101.

BANDYOPADHYAY, U. and BISWAS, A. (1996). Delayed response in randomized play-the-winner rule: a decision theoretic outlook. *Calcutta Statist. Assoc. Bull.* **46** 69–88.

BARTLETT, R. H., ROLOFF, D. W., CORNELL, R. G., ANDREWS, A. F., DILLON, P. W. and ZWISCHENBERGER, J. B. (1985). Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* **76** 479–487.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, San Diego.

HARDWICK, J. and STOUT, Q. (1998). Flexible algorithms for creating and analyzing adaptive sampling procedures. In *New Developments and Applications in Experimental Design* (N. Flournoy, W. F. Rosenberger, and W. K. Wong, eds.) 91–105. IMS, Hayward, CA.

ROSENBERGER, W. F. (1996). New directions in adaptive designs. *Statist. Sci.* **11** 137–149.

ROSENBERGER, W. F. (1999). Randomized play-the-winner clinical trials: review and recommendations. *Controlled Clinical Trials* **20** 328–342.

ROSENBERGER, W. F., FLOURNOY, N. and DURHAM, S. D. (1997). Asymptotic normality of maximum likelihood estimators from multiparameter response-driven designs. *J. Statist. Plann. Inference* **60** 69–76.

ROSENBERGER, W. F. and GRILL, S. G. (1997). A sequential design for psychophysical experiments: an application to estimating timing of sensory events. *Statistics in Medicine* **16** 2245–2260.

ROSENBERGER, W. F. and HU, F. (1999). Bootstrap methods for adaptive designs. *Statistics in Medicine* **18** 1757–1767.

ROSENBERGER, W. F. and SESHAIYER, P. (1997). Adaptive survival trials. *Journal of Biopharmaceutical Statistics* **7** 617–624.

SMYTHE, R. T. (1996). Central limit theorems for urn models. *Stochastic Process. Appl.* **65** 115–137.

TAMURA, R. N., FARIES, D. E., ANDERSEN, J. S. and HEILIGENSTEIN, J. H. (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *J. Amer. Statist. Assoc.* **89** 768–776.

WEI, L. J. (1979). The generalized Polya's urn design for sequential medical trials. *Ann. Statist.* **7** 291–296.

WEI, L. J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika* **75** 603–606.

WEI, L. J. and DURHAM, S. (1978). The randomized play-the-winner rule in medical trials. *J. Amer. Statist. Assoc.* **73** 840–843.

WEI, L. J., SMYTHE, R. T., LIN, D. Y. and PARK, T. S. (1990). Statistical inference with data-dependent treatment allocation rules. *J. Amer. Statist. Assoc.* **85** 156–162.

Z. D. BAI
DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 119260

F. HU
DEPARTMENT OF STATISTICS
HALSEY HALL
UNIVERSITY OF VIRGINIA
CHARLOTTESVILLE, VIRGINIA 22904-4135
E-MAIL: fh6e@pitman.stat.virginia.edu

W. F. ROSENBERGER
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF MARYLAND, BALTIMORE COUNTY
1000 HILLTOP CIRCLE
BALTIMORE, MARYLAND 21250