

Sequential Monitoring of Randomization Tests

William F. Rosenberger

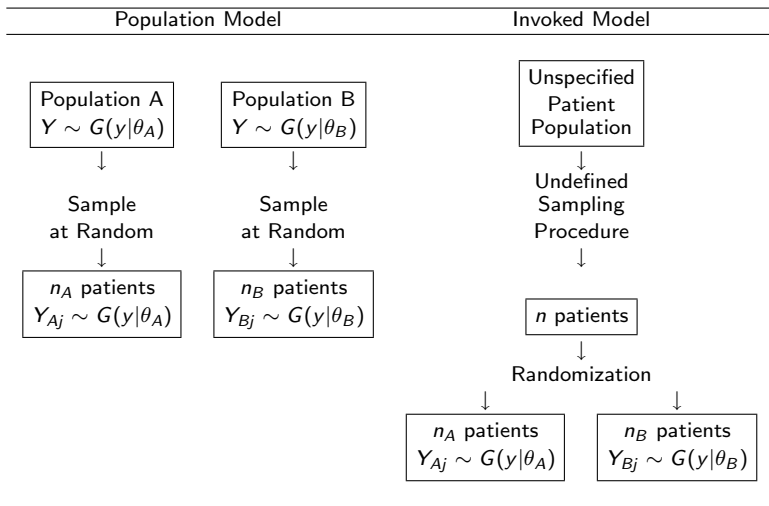
Department of Statistics
George Mason University
Fairfax, VA

Joint work with Yanqiong Zhang and R. T. Smythe

Outline

- The randomization principle
- Linear rank tests
- Sequential monitoring
- Stratified tests
- Determination of information
- Example
- Conditional tests

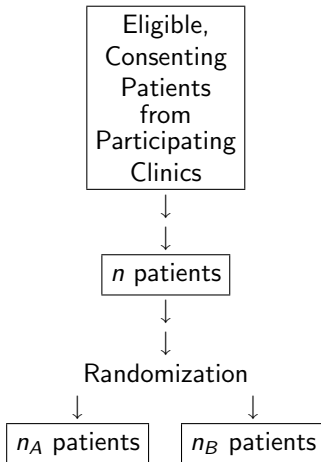
The population model



The randomization model

Randomization
Tests

Rosenberger,
W. F.



The importance of randomization:

Cornfield:

- 1. It controls the probability that the treated and control groups differ more than a calculable amount in their exposure to disease, in immune history, or with respect to any other variable, known or unknown to the experimenter, that may have a bearing on the outcome of the trial. This calculable difference tends to zero as the size of the two groups increase.*
- 2. It makes possible, at the end of the trial, the answer to the question "In how many experiments could a difference of this magnitude have arisen by chance alone if the treatment truly has no effect?" It may seem mysterious that a mathematician could actually predict the course of future experiments. All you have to do is compute what would happen if a given set of numbers were randomly allocated in all possible ways between the two groups. Randomization allows this.*

Randomization as a basis for inference

- Under H_0 , patient responses are a deterministic sequence unaffected by treatment.
- The observed treatment group difference depends only on the way in which the n patients were assigned.
- One chooses an appropriate metric for the treatment effect and computes this for all possible permutations of the randomization sequence.
- The sum of probabilities of sequences yielding a more extreme result than the observed treatment effect is then the p -value.

Unconditional and conditional tests

- What does all possible randomization sequences mean?
- One extreme: all possible sequences including $AAAAA\dots AAA$ and $BBBBB\dots BBBB$, which give no information on the treatment effect. This is an *unconditional test*.
- Other extreme: all possible sequences with the same number of A 's and B 's as the observed sequence. This is a *conditional test*.
- Something in between. Subjective. Must be defined in protocol: (e.g., within $\pm\delta n$ of N_A).

Randomization procedures

The most commonly used randomization procedures in clinical trials are

- Permuted Block Design
- Stratified Block Design
- Wei's Urn Design
- Stratified Urn Design
- Efron's Biased Coin Design

We cannot use the methodology developed in this talk for Efron's biased coin design, so we focus only on the first four.

Wei's urn design

After j patients we have $N_A(j)$ patients on treatment A and $N_B(j)$ patients on treatment B . Assign patient $j + 1$ to treatment A with probability $N_B(j)/j$.

We can do this within strata, and the strata will be independent. Unlike the permuted block design, it may not force balance within strata. The permuted block design can be unbalanced only if the last block is partially filled.

Outline

- The randomization principle
- Linear rank tests
- Sequential monitoring
- Stratified tests
- Determination of information
- Example
- Conditional tests

Family of linear rank tests

Randomization
Tests

Rosenberger,
W. F.

$$S = \sum_{i=1}^n (a_{in} - \bar{a}_n) T_i,$$

where a_{in} is a score associated with the outcome of the i th subject out of n subjects and $\bar{a}_n = \sum_{i=1}^n a_{in}/n$. In the following discussion, we will assume for convenience that $\bar{a}_n = 0$.

Randomization version of Mantel-Haenszel test

$$a_{in} = 1 \text{ or } 0.$$

Randomization version of Wilcoxon test

$$a_{in} = \frac{r_{in}}{n+1} - \frac{1}{2},$$

where r_{in} are the integer ranks.

Normal scores test

$$a_{in} = \Phi^{-1} \left(\frac{r_{in}}{n+1} \right),$$

where Φ is the standard normal distribution function.

Randomization version of logrank test

n ordered survival times $\tau_{(1)}, \dots, \tau_{(n)}$.

$$a_{in} = E(X_{(i)}) - 1,$$

where $X_{(1)}, \dots, X_{(n)}$ are order statistics from unit exponential random variables.

Logrank test with censoring

Let $\tau_{(1)} < \dots < \tau_{(M)}$ denote the M ordered distinct event times and let $\delta_i = 1$ if patient i has an event; $\delta_i = 0$ if patient i is censored. Let R_m denote the number of patients at risk immediately before $\tau_{(m)}$.

$$a_{in} = \delta_i - \sum_{m=1}^i \frac{\delta_m}{R_m}.$$

Randomization procedures

- We will investigate the permuted block design and Wei's urn design in the context of the linear rank test under a randomization model, and also the stratified version of the randomization procedures.
- A stratified randomization procedure will lead to a stratified linear rank test.

Using randomization tests

- Exact tests: computationally intensive.
- Monte Carlo (Rerandomization): generate K randomization sequences and compute an approximate p -value based on those sequences. Difficult for conditional tests. Does not fit into the framework of sequential monitoring, which is based on finding the joint asymptotic distribution of sequentially computed statistics.
- Use asymptotic normality of the linear rank test. This is the approach we take.
- Will mention the Monte Carlo approach again later.

Why not use them?

- Assumption free; analyze as you randomize; follows naturally from the randomization model; lauded as one of the purposes of randomization: to allow a basis for inference.
- Even the simple t -test violates the randomization principle.
- History; difficulty of computation; evolution of SAS; black-box software culture.
- Often gives same result as population-based test.

The Lindeberg condition

$$\lim_{n \rightarrow \infty} \frac{\max_{1 \leq i \leq n} a_{in}^2}{\sum_{i=1}^n a_{in}^2} = 0,$$

ensures that the linear rank test, suitably normalized, is asymptotically standard normal. For most randomization procedures, the Lindeberg condition holds for the

- Mantel-Haenszel test
- Wilcoxon test
- Normal scores test
- Logrank test
- Censored logrank test, provided the censoring indicator is independent of the treatment (Zhang and Rosenberger, *J. Nonpar. Statist.*, 2005). (No counting processes needed!)

Outline

- The randomization principle
- Linear rank tests
- Sequential monitoring
- Stratified tests
- Determination of information
- Example
- Conditional tests

Lan-DeMets Procedure

There is an increasing continuous function $\alpha(t)$, $t \in [0, 1]$, such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$, then we find constants d_1, d_2, \dots, d_K , such that

$$P(S^{(1)} > d_1) = \alpha(t_1),$$

$$P(S^{(2)} > d_2, S^{(1)} \leq d_1) = \alpha(t_2) - \alpha(t_1),$$

\vdots

$$P(S^{(K)} > d_K, S^{(K-1)} \leq d_{K-1}, \dots, S^{(1)} \leq d_1) = \alpha(t_K) - \alpha(t_{K-1}).$$

where t_k represents the “information fraction” at the k th inspection.

Lan-DeMets Procedure

To develop a sequential monitoring plan we need to establish:

- The joint asymptotic distribution of $(S^{(1)}, S^{(2)}, \dots, S^{(K)})$.
- The *information fractions* t_1, \dots, t_K .
- The boundary values d_1, \dots, d_K by numerical integration.

We focus on the first two items.

Outline

- The randomization principle
- Linear rank tests
- Sequential monitoring
- Stratified tests
- Determination of information
- Example
- Conditional tests

Permuted block randomization

Let

$$S_B^{(k)} = \sum_{i=1}^{M_k} w_i \sum_{j=1}^m a_{j(i)} T_{j(i)}, \quad 1 \leq k \leq K,$$

with $N_k = mM_k$, where w_i is the weight for block i , $a_{j(i)}$ is the centered score function for the j th patient in block i and $T_{j(i)}$ is the treatment allocation for the j th patient in block i .

THEOREM 1. For the permuted block design, as $M_1, \dots, M_K \rightarrow \infty$, $\Sigma^{-1/2}(S_B^{(1)}, S_B^{(2)}, \dots, S_B^{(K)})$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$, and identity covariance matrix, where $\Sigma_{ii} = \text{Var}(S_B^{(i)})$, $\Sigma_{ij} = \text{Var}(S_B^{(k)})$, with $k = \min(i, j)$, and

$$\text{Var}(S_B^{(k)}) = \frac{m}{m-1} \sum_{i=1}^{M_k} w_i^2 \sum_{j=1}^m a_{j(i)}^2.$$

Stratified block design

Let

$$S_{TB}^{(k)} = \sum_{l=1}^L \sum_{i=1}^{M_{lk}} w_{il} \sum_{j=1}^m a_{j(il)} T_{j(il)}, \quad 1 \leq k \leq K, \quad 1 \leq l \leq L,$$

with $N_k = \sum_{l=1}^L m M_{lk}$, where w_{il} is the weight block i in stratum l , M_{lk} is the number of blocks at the k th inspection in stratum l , $a_{j(il)}$ is the centered score function for j th patient in i th block of stratum l and $T_{j(il)}$ is the treatment allocation for j th patient in i th block i of stratum l .

COROLLARY 1. For each stratum l , $1 \leq l \leq L$, as $M_{l1}, \dots, M_{lK} \rightarrow \infty$, $\Sigma^{-1/2}(S_{TB}^{(1)}, S_{TB}^{(2)}, \dots, S_{TB}^{(K)})$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$, and identity covariance matrix, where $\Sigma_{ii} = \text{Var}(S_{TB}^{(i)})$, $\Sigma_{ij} = \text{Var}(S_{TB}^{(k)})$, with $k = \min(i, j)$, and

$$\text{Var}(S_{TB}^{(k)}) = \frac{m}{m-1} \sum_{l=1}^L \sum_{i=1}^{M_{lk}} w_{il}^2 \sum_{j=1}^m a_{j(il)}^2.$$

Stratified urn design

THEOREM 2. Assume some regularity conditions (e.g., the Lindeberg condition holds in each increment between the looks). Let

$$b_{ik} = a_{ik} - r \sum_{s=i+1}^{N_k} \left(\frac{a_{sk}}{s-1} \prod_{j=i}^{s-2} \left(1 - \frac{r}{j} \right) \right), \quad i = 1, \dots, N_k,$$

and let $\lambda_k = (b_{1k}, \dots, b_{N_k k}, 0, \dots, 0)'$, where a_{ik} is the score function for the i th patient at the k th inspection; b_{ik} is the corresponding modified score function; λ_k is a $N_k \times 1$ vector with $b_{ik} \neq 0$ for some $i \leq N_k$. Let $(\Lambda)_{km} = \lambda_k' \lambda_m$ be a $K \times K$ matrix. Then as

$N_k - N_{k-1} \rightarrow \infty$, $1 \leq k \leq K$, $\Lambda^{-1/2}(S^{(1)}, \dots, S^{(K)})'$ converges in distribution to a multivariate normal random vector with mean 0 and identity covariance matrix.

Stratified urn design

- The regularity conditions on the scores are generally mild for practical use.
- They hold for Mantel-Haenszel, Wilcoxon, and normal scores tests.
- For the logrank test, the theorem requires the number of events between the $(k - 1)$ th and k th inspection is asymptotically the same order as $N_k - N_{k-1}$. This may not be true in some survival trials where patients are accrued in larger numbers at the beginning of the trial.
- In most settings, if $N_k - N_{k-1}$ is of moderate size, the multivariate normal approximation should be reasonable.

Outline

- The randomization principle
- Linear rank tests
- Sequential monitoring
- Stratified tests
- Determination of information
- Example
- Conditional tests

Determination of information fraction

- Information is a population-based concept from the likelihood principle. How do we interpret information in a randomization context?
- The most natural approach is to treat information as the inverse of the variance of the test statistic (or an approximation) and thus define it as

$$t_k = \frac{\text{Var}(S^{(k)})}{\text{Var}(S^{(K)})}.$$

- We can compute this information fraction based on the variance of the randomization procedure, which is often not a simple task.

Determination of information fraction

- For the permuted block design, t_k is simply the proportion of the number of blocks observed, or the proportion of the number of patients observed at the interim point, if the same weights are used in each block.
- This extends easily to the same conclusion for stratified blocks.
- For the stratified urn design, it is considerably more complicated. The information fraction at the k th inspection must be estimated as

$$t_k = \frac{\sum_{l=1}^L w_l^2 \sum_{j=1}^{N_{lk}} b_{j(lk)}^2}{\sum_{l=1}^L w_l^2 \sum_{j=1}^{N_{lK}} a_{j(lK)}^2}.$$

Outline

- The randomization principle
- Linear rank tests
- Sequential monitoring
- Stratified tests
- Determination of information
- Example
- Conditional tests

Example

- Supplemental Therapeutic Oxygen for Prethreshold Retinopathy of Prematurity (STOP-ROP) trial
- 649 infants in 30 clinical centers from February 1994 to March 1999
- Stratified urn design was used within clinical center and ROP severity strata.
- We analyze data on 156 infants from five strata, taking October 31, 1996, as our interim time point. At this time, 83 responses had been observed, 44 on conventional and 39 on experimental treatment. Response was binary, so binary scores are used. We use equal weights within strata.

Example continued

$$S_T^{(1)} = \sum_{l=1}^5 \sum_{j=1}^{N_{l1}} a_{j(l1)} T_{j(l1)} = 4.5,$$

$$\text{Var}(S_T^{(1)}) = \sum_{l=1}^5 \sum_{j=1}^{N_{l1}} b_{j(l1)}^2 = 20.491;$$

thus $W_T^{(1)} = 0.994$. We estimate the information as

$$t_1 = \frac{\sum_{l=1}^5 \sum_{j=1}^{N_{l1}} b_{j(l1)}^2}{\sum_{l=1}^5 \sum_{j=1}^{N_{l2}} a_{j(l2)}^2} = \frac{20.491}{39} = 0.525.$$

If the O'Brien-Fleming (1979) spending function is used, $\alpha(0.525) = 0.007$, $d_1 = 2.457 * \sqrt{20.491} = 11.122$.

Example continued

For the final inspection,

$$S_T^{(2)} = \sum_{l=1}^5 \sum_{j=1}^{N_{l2}} a_{j(l2)} T_{j(l2)} = -3,$$

$$\text{Var}(S_T^{(2)}) = \sum_{l=1}^5 \sum_{j=1}^{N_{l2}} b_{j(l2)}^2 = 39.684;$$

thus $W_T^{(2)} = -0.476$. As $\alpha(1) - \alpha(0.525) = 0.043$, from the variance-covariance matrix of $(W_T^{(1)}, W_T^{(2)})$, we obtain $d_2 = 1.702 * \sqrt{39.684} = 10.722$.

Outline

Randomization
Tests

Rosenberger,
W. F.

- The randomization principle
- Linear rank tests
- Sequential monitoring
- Stratified tests
- Determination of information
- Example
- **Conditional tests**

Conditional tests

This is a much more complicated problem as we must find the joint asymptotic distribution of sequentially computed test statistics, conditional on $N_A = n_A$. Zhang does this in her doctoral thesis, but only for $K = 2$ for Wei's urn design. The information fraction must be redefined as

$$t_k = \frac{\text{Var}(S^{(k)} | N_A = n_A)}{\text{Var}(S^{(K)} | N_A = n_A)}.$$

Monte Carlo sequential conditional tests

For the conditional test, we generate K sequences and order only those sequences that have the same or close to the same number of A 's and B 's as were observed. For the k th inspection, this means that we consider only the subset of sequences such that $N_{Al} \in n_{Al,obs.} \pm \gamma n, l = 1, \dots, k - 1$. In order to have sufficient sequences to estimate the conditional probability effectively, we need on average

$$\frac{1}{4\epsilon \Pr\{(N_{Al} \in n_{Al,obs.}) \pm \gamma n, l = 1, \dots, k - 1\}}.$$

This number can be quite large.

Conclusions

- Randomization provides a basis for inference; while this is well-known in the culture of biostatistics, randomization tests are rarely used in clinical trials.
- It seems reasonable, given the limited assumptions of the test, and the natural randomized structure of clinical trials, that they should at least be employed for the primary outcome analysis of simple treatment effect.
- The linear rank test formulation can accomplish this by providing asymptotically normal test statistics for binary, continuous, and survival data.
- We now provide a method to perform sequential monitoring of a randomization test, a critical component in most clinical trials.