Sequential Monitoring of Randomization Tests

William F. Rosenberger and Victoria Plamadeala

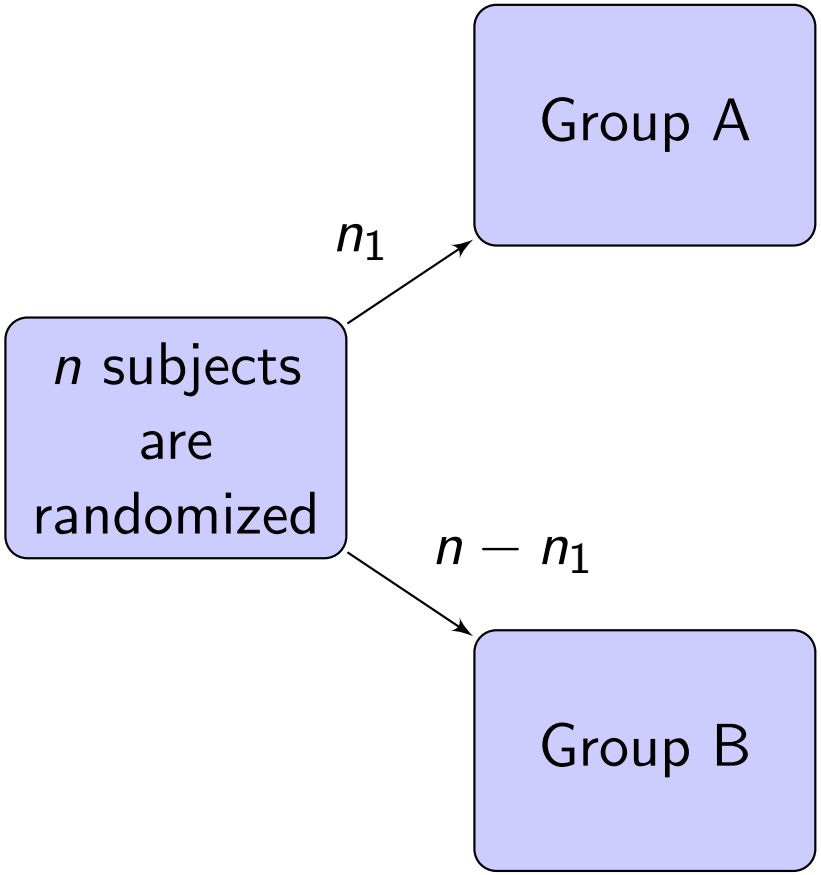
Department of Statistics
George Mason University
Fairfax, VA
and
Precision Therapeutics
Pittsburgh, PA

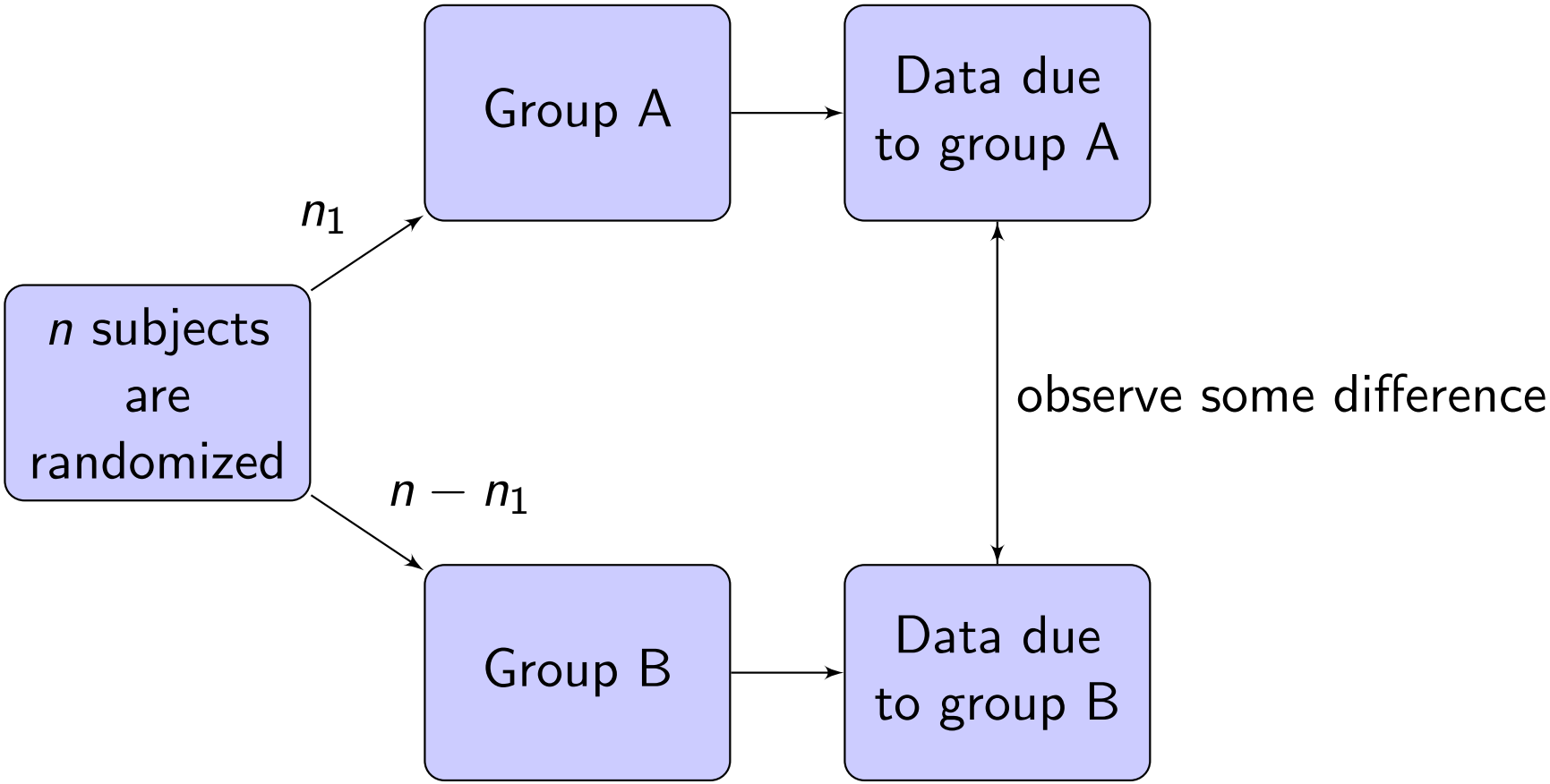
June 6, 2011

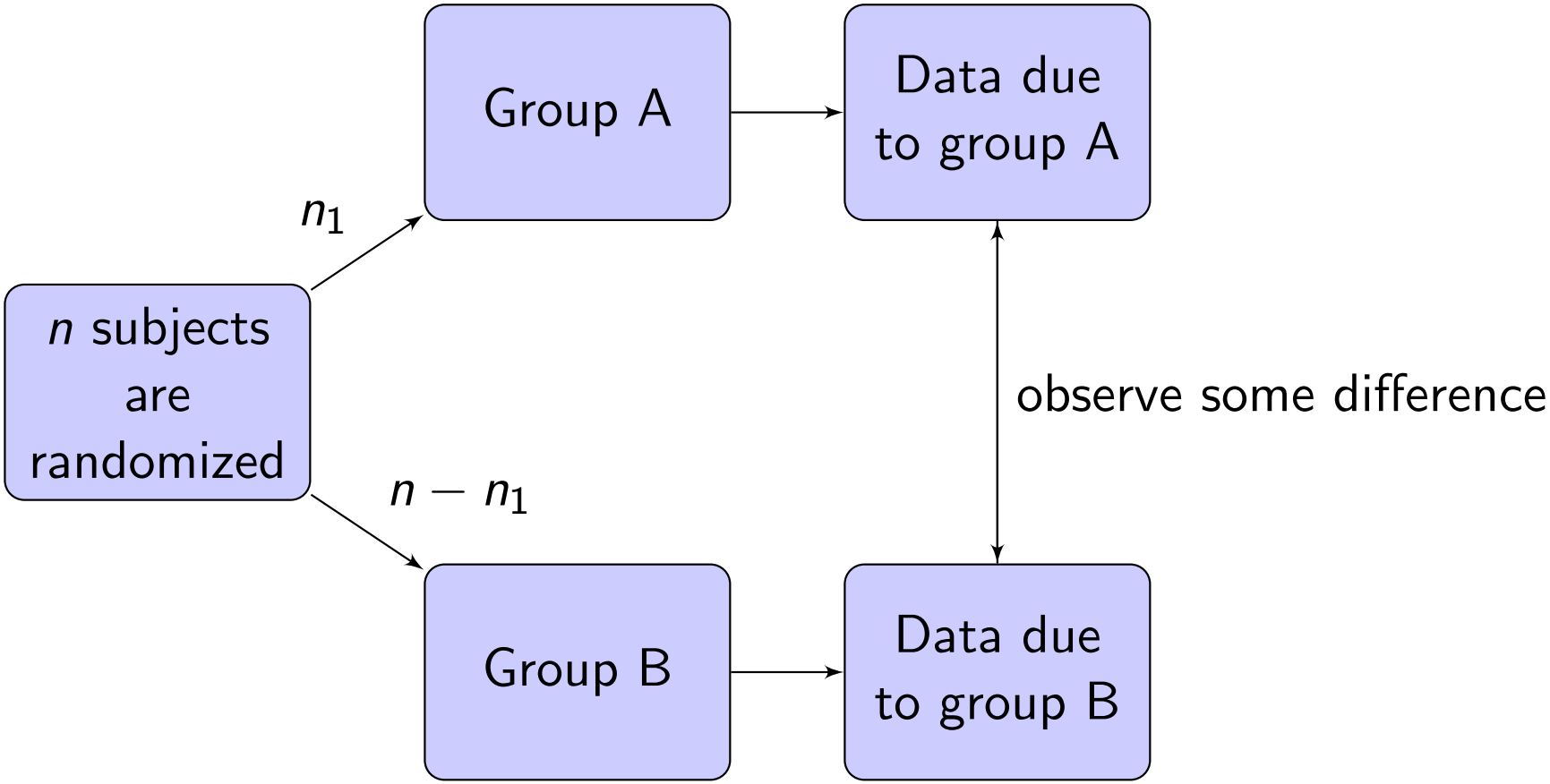
Everyone knows that clinical trials do not follow a population model with random sampling. In fact the only random mechanism is the randomization itself.

For this reason, the FDA often requires a randomization-based inference analysis by applying a randomization test (which they often call "re-randomization test") to the primary outcome analysis. Such analyses are typically done by using Monte Carlo simulation to regenerate randomization procedures and then compute the test statistics p-value as the proportion of times the simulated statistic exceeds the observed.

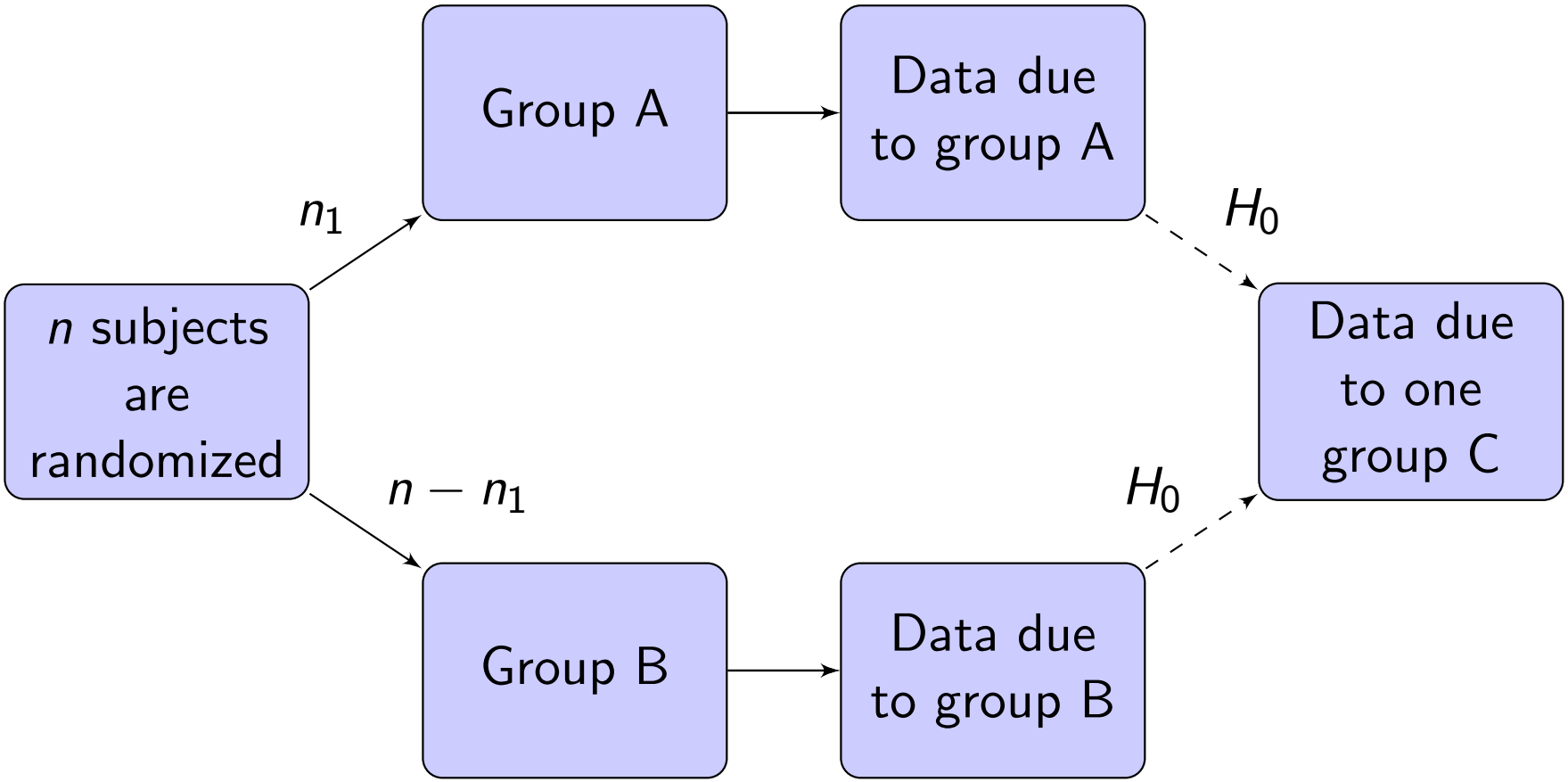
However, this is an "unconditional" test, which may not be desirable. It also ignores the possibility that there was sequential monitoring in the trial. In this talk we describe how to compute conditional tests and incorporate them in the sequential monitoring plan.


Group A
n_1
n subjects are randomized
$n - n_1$
Group B

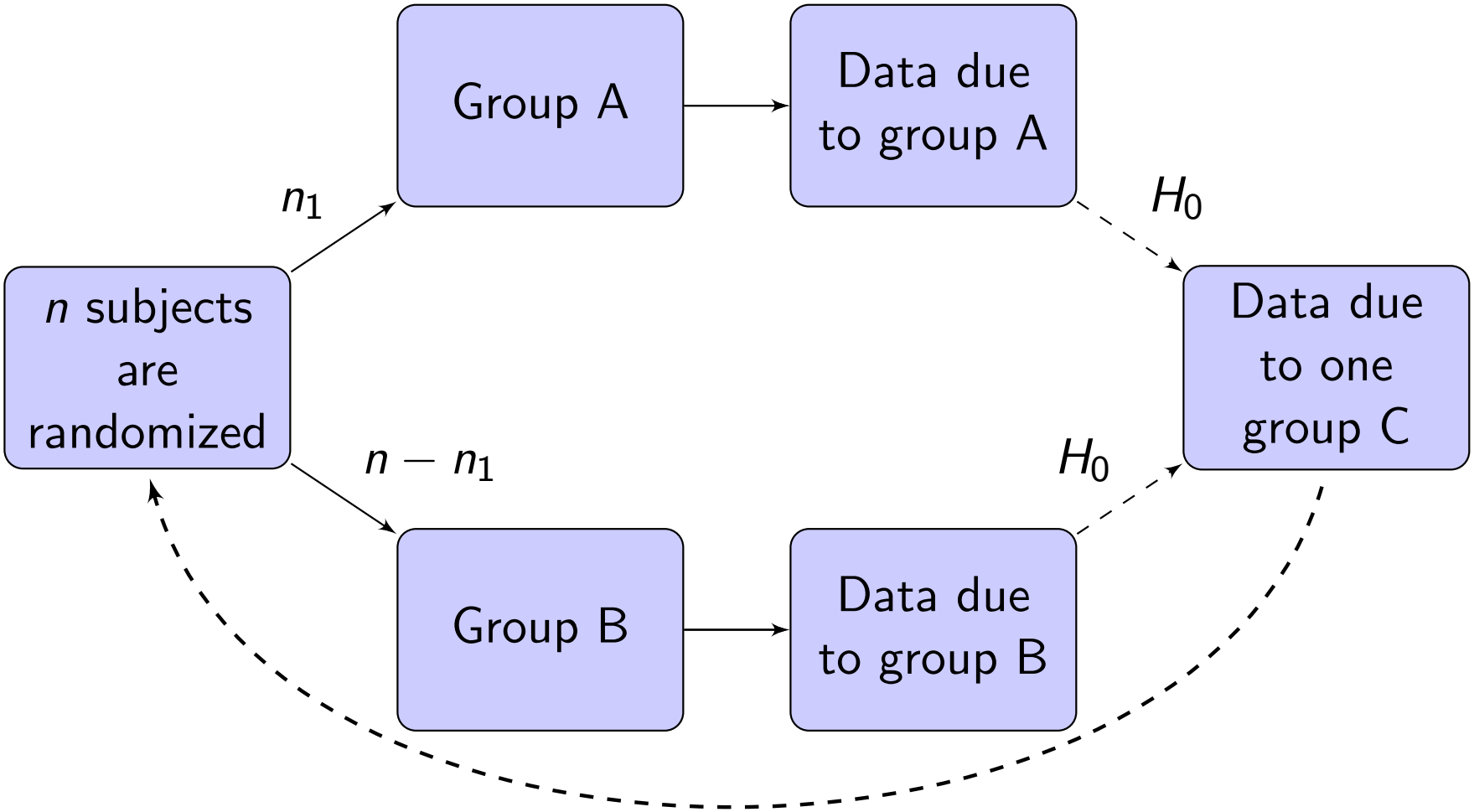

Group A
Data due to group A
n_1
n subjects are randomized
observe some difference
$n - n_1$
Group B
Data due to group B



Test H_0 : A and B are not different.



Test H_0 : A and B are not different.



Any observed difference is entirely due to the randomization

Test H_0 : A and B are not different.

The p–value of the randomization test depends on the randomization procedure, $\mathbf{T}_n = (T_1, ..., T_n)'$, used:

- **complete randomization:**

$$\phi_{j+1} = P(T_{j+1} = 1) = (T_{j+1} = 0) = 1/2,\ j = 0, ..., n-1.$$

Typically we use some form of restricted randomization to promote balance between treatment assignments both at the end and throughout the middle of the trial:

- **permuted block design**
- **biased coin design** and its generalizations and extensions

Restricted randomization procedures:
Efron's (1971) biased coin design (BCD)

$$\phi_{j+1}(m_j) = P(T_{j+1} = 1 | \sum_{i=1}^{j} T_i = m_j) = \begin{cases} 1/2, & \text{if} \quad m_j = j/2 \\ p, & \text{if} \quad m_j < j/2 \\ 1-p, & \text{if} \quad m_j > j/2, \end{cases}$$

$j = 0, ..., n-1$, m_j is the number assigned to treatment 1 when j subjects have already been randomized.

The p–value of the randomization test depends on

- the type of reference set used:
  - Ω_u – unconditional set with cardinality 2^n;
  - Ω_c – conditional set with cardinality $\binom{n}{n_1}$, where n_1 is the final number assigned to treatment 1.

The p–value of the randomization test depends on

- the type of reference set used:
  - Ω_u – unconditional set with cardinality 2^n;
  - Ω_c – conditional set with cardinality $\binom{n}{n_1}$, where n_1 is the final number assigned to treatment 1.
- the metric of the treatment effect.

Example

- Suppose BCD($p = 3/4$) was used. The observed responses are $\{2.3, 1.9, 2.2, 2.1, 2.0\}$ on treatments $\{1, 0, 0, 1, 0\}$.
- The analog of the Wilcoxon rank sum test gives:

$$\sum_{i=1}^{5} a_{jn} T_i = (2 \ -2 \ 1 \ 0 \ -1)(1 \ 0 \ 0 \ 1 \ 0)' = 2.$$

- Choose the type of reference set:
  - Conditional: use only $\binom{5}{2}$ sequences that satisfy $n_1 = 2$.
  - Obtain the randomization distribution by computing all possible inner products on the reference set.

In the earlier example:

The unconditional set Ω_u contains 2^5 sequences:

0, 0, 0, 0, 0	0, 1, 0, 0, 0	1, 0, 0, 0, 0	1, 1, 0, 0, 0
0, 0, 0, 0, 1	0, 1, 0, 0, 1	1, 0, 0, 0, 1	1, 1, 0, 0, 1
0, 0, 0, 1, 0	0, 1, 0, 1, 0	1, 0, 0, 1, 0	1, 1, 0, 1, 0
0, 0, 0, 1, 1	0, 1, 0, 1, 1	1, 0, 0, 1, 1	1, 1, 0, 1, 1
0, 0, 1, 0, 0	0, 1, 1, 0, 0	1, 0, 1, 0, 0	1, 1, 1, 0, 0
0, 0, 1, 0, 1	0, 1, 1, 0, 1	1, 0, 1, 0, 1	1, 1, 1, 0, 1
0, 0, 1, 1, 0	0, 1, 1, 1, 0	1, 0, 1, 1, 0	1, 1, 1, 1, 0
0, 0, 1, 1, 1	0, 1, 1, 1, 1	1, 0, 1, 1, 1	1, 1, 1, 1, 1

In the earlier example:

The conditional set Ω_c that satisfies $\{\sum_{i=1}^{5} T_i = N_1(5) = 2\}$:

$1,1,0,0,0$

$0,1,0,0,1$ $\quad$ $1,0,0,0,1$

$0,1,0,1,0$ $\quad$ $1,0,0,1,0$

$0,0,0,1,1$

$0,1,1,0,0$ $\quad$ $1,0,1,0,0$

$0,0,1,0,1$

$0,0,1,1,0$

In the earlier example:

The conditional distribution of $S(\mathbf{T}) = (2, -2, 1, 0, -1)\mathbf{T}$ under the BCD(p):

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \\ 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \\ -3 \\ -2 \\ -1 \\ 1 \\ 2 \\ 3 \\ 0 \end{bmatrix} \text{ with prob. } \begin{bmatrix} 0.5(1-p)^2p^2/C \\ 0.5(1-p)^2p^2/C \\ 0.5^2(1-p)p^2/C \\ 0.5^2(1-p)p^2/C \\ 0.5^3p^2/C \\ 0.5^3p^2/C \\ 0.5^2(1-p)p^2/C \\ 0.5^3p^2/C \\ 0.5^3p^2/C \\ 0.5^2(1-p)p^2/C \end{bmatrix}$$

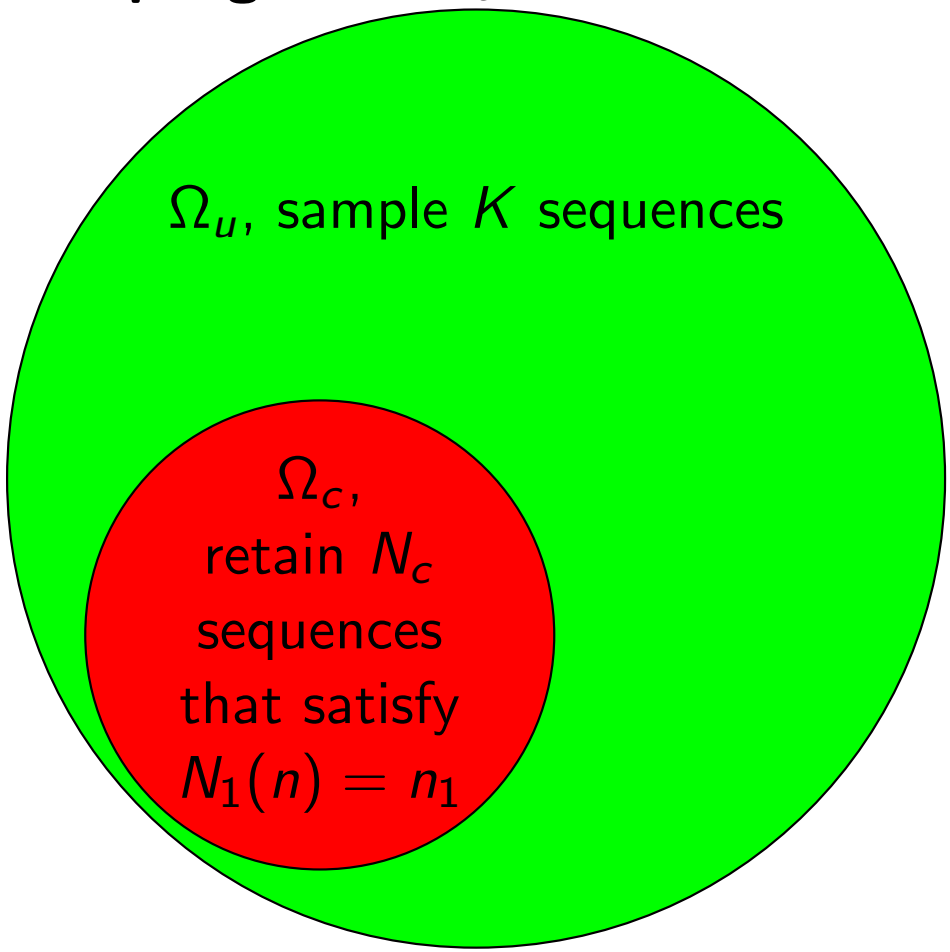
where $C = p^2(2.5 - 3p + p^2)$.

Conditional tests under the BCD(p): computational techniques

- Exact tests (Hollander and Peña, 1988; Mehta, Patel and Wei, 1988a) for modern clinical trials are computationally infeasible.
- No asymptotic approximation exists for the randomization test under Efron's BCD.

Conditional Tests Comparing Two Groups

Method 1: Sampling from Ω_u



Method 1: Sampling from Ω_u

- Use the randomization rule ϕ_{j+1} to generate these sequences.
- The main problem is deciding on K, the overall MC sample size.

Method 1: Sampling from Ω_u

- $K \sim NB(P(N_1(n) = n_1), N_c)$
- One can compute $P(N_1(n) = n_1)$ exactly under the BCD(p) (Markaryan and Rosenberger, 2010).

Method 1: Sampling from Ω_u

Table: *Approximate 95th percentile of K for various n, n_1, $N_c = 2500$; BCD (2/3).*

n	$n_1 = 0.45n$	$n_1 = 0.48n$	$n_1 = 0.50n$
100	3,531,344	55,060	5,117
200	3,611,280,266	881,557	5,117
500	$3,877,310 \times 10^{12}$	3,611,026,232	5,117

Method 2: Sampling directly from Ω_c

- Perhaps some computation efficiency could be gained if the N_c sequences were sampled directly from Ω_c.
- This requires adjusting the randomization rules ϕ_{j+1} to generate sequences directly from Ω_c.

Method 2: Sampling directly from Ω_c

The appropriate sampling rule is

$$p_{j+1} = P\left(T_{j+1} = 1 | \sum_{i=1}^{j} T_i = m_j, \sum_{i=1}^{n} T_i = n_1\right)$$

$$= \begin{cases} \phi_{j+1}(m_j) \dfrac{P\left(N_1(n) = n_1 | N_1(j+1) = m_j + 1\right)}{P\left(N_1(n) = n_1 | N_1(j) = m_j\right)}, \ 1 \leq j \leq n-1, \\ 1/2 \dfrac{P\left(N_1(n) = n_1 | T_{j+1} = 1\right)}{P\left(N_1(n) = n_1\right)}, \ j = 0. \end{cases}$$

Method 2: Sampling directly from Ω_c

Closed form solutions for p_{j+1} were derived for Efron's BCD(p) using combinatorics.

1. Use p_{j+1} to sample N_c sequences directly from Ω_c.

1. Use p_{j+1} to sample N_c sequences directly from Ω_c.
2. Evaluate the statistic for each sequence.

1. Use p_{j+1} to sample N_c sequences directly from Ω_c.
2. Evaluate the statistic for each sequence.
3. Use the sample proportion, $\widehat{p}$, as the estimator of p_c, the conditional test p–value.

1. Use p_{j+1} to sample N_c sequences directly from Ω_c.
2. Evaluate the statistic for each sequence.
3. Use the sample proportion, $\widehat{p}$, as the estimator of p_c, the conditional test p–value.
4. $\text{MSE}(\widehat{p}) = p_c(1-p_c)/N_c \leq 1/4N_c \leq \epsilon$. For $\epsilon = 0.0001$, $N_c \geq 2500$.

1. Use p_{j+1} to sample N_c sequences directly from Ω_c.
2. Evaluate the statistic for each sequence.
3. Use the sample proportion, $\widehat{p}$, as the estimator of p_c, the conditional test p–value.
4. $\text{MSE}(\widehat{p}) = p_c(1-p_c)/N_c \leq 1/4N_c \leq \epsilon$. For $\epsilon = 0.0001$, $N_c \geq 2500$.
5. For higher precision, one can use $P(|\widehat{p} - p_c| \leq 0.1p_c) = 0.99$.

1. Use p_{j+1} to sample N_c sequences directly from Ω_c.
2. Evaluate the statistic for each sequence.
3. Use the sample proportion, $\widehat{p}$, as the estimator of p_c, the conditional test p–value.
4. $\text{MSE}(\widehat{p}) = p_c(1-p_c)/N_c \leq 1/4N_c \leq \epsilon$. For $\epsilon = 0.0001$, $N_c \geq 2500$.
5. For higher precision, one can use $P(|\widehat{p} - p_c| \leq 0.1p_c) = 0.99$. If $p_c = 0.04$, $N_c \approx 16,000$, by the CLT.

Sequential Randomization Tests Comparing Two Groups

Suppose an α level test is planned for some clinical trial comparing two treatments after n subjects have responded.

But want to allow for $L-1$ interim inspections of the data after $1 \leq r_1 < r_2 < \ldots < r_{L-1} < r_L = n$ patients have responded.

A Type I error spending function (Lan and DeMets, 1983), a nondecreasing function

$$\alpha^* : [0, 1] \mapsto [0, \alpha].$$

Information fraction: $0 < t_1 < t_2 < ... < t_{L-1} < t_L = 1$.

A typical sequential monitoring plan with L interim looks:

- $P(V_{r_1} > d_1) = \alpha^*(t_1)$

A typical sequential monitoring plan with L interim looks:

- $P(V_{r_1} > d_1) = \alpha^*(t_1)$
- $P(V_{r_1} \leq d_1, V_{r_2} > d_2) = \alpha^*(t_2) - \alpha^*(t_1)$

A typical sequential monitoring plan with L interim looks:

- $P(V_{r_1} > d_1) = \alpha^*(t_1)$
- $P(V_{r_1} \leq d_1, V_{r_2} > d_2) = \alpha^*(t_2) - \alpha^*(t_1)$
- $P(V_{r_1} \leq d_1, V_{r_2} \leq d_2, V_{r_3} > d_3) = \alpha^*(t_3) - \alpha^*(t_2)$

A typical sequential monitoring plan with L interim looks:

- $P(V_{r_1} > d_1) = \alpha^*(t_1)$
- $P(V_{r_1} \leq d_1, V_{r_2} > d_2) = \alpha^*(t_2) - \alpha^*(t_1)$
- $P(V_{r_1} \leq d_1, V_{r_2} \leq d_2, V_{r_3} > d_3) = \alpha^*(t_3) - \alpha^*(t_2)$
- $\vdots$
- $P(V_{r_1} \leq d_1, V_{r_2} \leq d_2, ..., V_{r_L} > d_L) = \alpha - \alpha^*(t_{L-1})$.

The randomization version of the sequential testing plan:

$$\begin{cases} P(V_{r_1} > d_1 | N_1(r_1) = n_{11}) & = \alpha^*(t_1), \\ P(V_{r_1} \leq d_1, V_{r_2} > d_2 | N_1(r_1) = n_{11}, N_1(r_2) = n_{12}) & = \alpha^*(t_2) - \alpha^*(t_1), \\ P(V_{r_1} \leq d_1, V_{r_2} \leq d_2, V_{r_3} > d_3 | \bigcap_{j=1}^{3} N_1(r_j) = n_{1j}) & = \alpha^*(t_3) - \alpha^*(t_2), \\ & \vdots \\ P(V_{r_1} \leq d_1, ..., V_{r_{L-1}} \leq d_{L-1}, V_n > d_L | \bigcap_{j=1}^{L} N_1(r_j) = n_{1j}) & = \alpha - \alpha^*(t_{L-1}). \end{cases}$$

(Zhang and Rosenberger, 2008)

Equivalent to

$$\begin{cases} P\left(V_{r_1} > d_1 | N_1(r_1) = n_{11}\right) & = \alpha^*(t_1), \\ P\left(V_{r_2} > d_2 | V_{r_1} \leq d_1, \bigcap_{j=1}^{2}\{N_1(r_j) = n_{1j}\}\right) & = \frac{\alpha^*(t_2)-\alpha^*(t_1)}{1-\alpha^*(t_1)}, \\ P\left(V_{r_3} > d_3 | \bigcap_{j=1}^{2}\{V_{r_j} \leq d_j\}, \bigcap_{j=1}^{3}\{N_1(r_j) = n_{1j}\}\right) & = \frac{\alpha^*(t_3)-\alpha^*(t_2)}{1-\alpha^*(t_2)}, \\ & \vdots \\ P\left(V_n > d_L | \bigcap_{j=1}^{L-1}\{V_{r_j} \leq d_j\}, \bigcap_{j=1}^{L}\{N_1(r_j) = n_{1j}\}\right) & = \frac{\alpha-\alpha^*(t_{L-1})}{1-\alpha^*(t_{L-1})}. \end{cases}$$

1. At each look l, sample sequences that satisfy $\bigcap_{j=1}^{l}\{N_1(r_j) = n_{1j}\}$.

1. At each look l, sample sequences that satisfy $\bigcap_{j=1}^{l}\{N_1(r_j) = n_{1j}\}$.
2. Retain those that satisfy $\bigcap_{j=1}^{l-1}\{V_{r_j} \leq d_j\}$.

1. At each look l, sample sequences that satisfy $\bigcap_{j=1}^{l}\{N_1(r_j) = n_{1j}\}$.
2. Retain those that satisfy $\bigcap_{j=1}^{l-1}\{V_{r_j} \leq d_j\}$.
3. Evaluate the current stage statistic V_{r_l} for each retained sequence.

1. At each look l, sample sequences that satisfy $\bigcap_{j=1}^{l}\{N_1(r_j) = n_{1j}\}$.
2. Retain those that satisfy $\bigcap_{j=1}^{l-1}\{V_{r_j} \leq d_j\}$.
3. Evaluate the current stage statistic V_{r_l} for each retained sequence.

1. At each look l, sample sequences that satisfy $\bigcap_{j=1}^{l}\{N_1(r_j) = n_{1j}\}$.
2. Retain those that satisfy $\bigcap_{j=1}^{l-1}\{V_{r_j} \leq d_j\}$.
3. Evaluate the current stage statistic V_{r_l} for each retained sequence.
4. The resulting values approximate the required distribution.

1. At each look l, sample sequences that satisfy $\bigcap_{j=1}^{l}\{N_1(r_j) = n_{1j}\}$.
2. Retain those that satisfy $\bigcap_{j=1}^{l-1}\{V_{r_j} \leq d_j\}$.
3. Evaluate the current stage statistic V_{r_l} for each retained sequence.
4. The resulting values approximate the required distribution.
5. Use this distribution to estimate d_l.

The main problem is to sample sequences that satisfy $\bigcap_{j=1}^{l}\{N_1(r_j) = n_{1j}\}$.

The randomization–based information fraction is defined as the ratio of the conditional variances (Rosenberger and Lachin, 2002):

$$t_l = \frac{\mathbf{a}'_{r_l} \boldsymbol{\Sigma}_{|r_l} \mathbf{a}_{r_l}}{\mathbf{a}'_n \boldsymbol{\Sigma}_{|n} \mathbf{a}_n}.$$

where $\boldsymbol{\Sigma}_{|r_l} = Var(\mathbf{T}^{(r_l)} | N_1(r_1) = n_{11}, ..., N_1(r_l) = n_{1l})$, for which we have derived exact computational expressions under the BCD(p) and complete randomization.

We generate an allocation sequence under the BCD(3/4) and observations from two normal distributions to simulate the sequential analysis of the randomization test (3 inspections).

At each inspection the goal is to estimate the information by extrapolating the unknown data using three methods:

1. Sampling from two different normal distributions with estimated unknown parameters.

We generate an allocation sequence under the BCD(3/4) and observations from two normal distributions to simulate the sequential analysis of the randomization test (3 inspections).

At each inspection the goal is to estimate the information by extrapolating the unknown data using three methods:

1. Sampling from two different normal distributions with estimated unknown parameters.
2. Sampling from one normal distribution with estimated unknown parameters.

We generate an allocation sequence under the BCD(3/4) and observations from two normal distributions to simulate the sequential analysis of the randomization test (3 inspections).

At each inspection the goal is to estimate the information by extrapolating the unknown data using three methods:

1. Sampling from two different normal distributions with estimated unknown parameters.
2. Sampling from one normal distribution with estimated unknown parameters.
3. Bootstrapping the unknown data by groups.

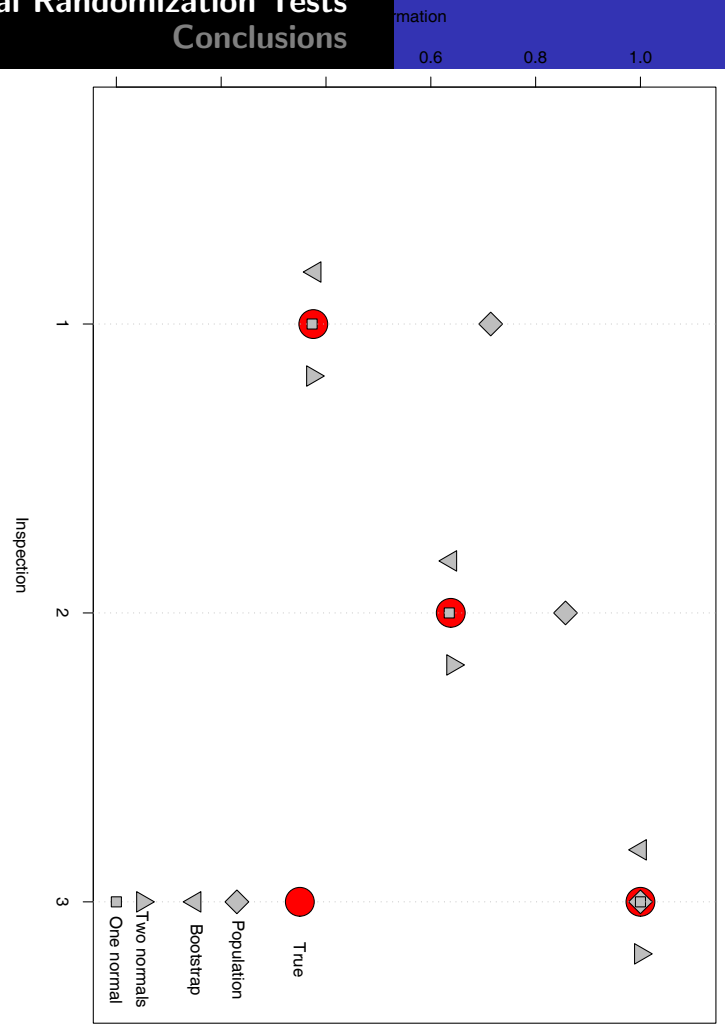


Figure: *Information estimation with three interim inspections: BCD(3/4),* $L = 3$; $r_1 = 250$ *and* $n_{11} = 126$, $r_2 = 300$ *and* $n_{12} = 148$, $n = 350$ *and* $n_1 = 174$.

Conclusions

1. Under Efron's BCD(p), the conditional randomization test is not asymptotically normal and therefore the technique of sampling from the conditional reference provides a reasonable method of approximation.

2. We have answered a question posed by Zelen at a recent JSM: "How does one compute conditional randomization tests?" Sampling from the conditional reference set facilitated the approximation of stratified conditional randomization tests under Efron's BCD(p).

3. To our knowledge, no one has ever established a method for sequential monitoring of conditional randomization tests We now have a format in which this can be done.