Selected Topics in Phase II and III Clinical Trials

Tatsuki Koyama, PhD
Department of Biostatistics, Vanderbilt University Medical Center
tatsuki.koyama@vumc.org

Osaka City University, Biostatistics.

November 10 ∼ 18, 2021

Last updated: 2021-11-08 13:10
R version: 4.1.1

Contents

Chapter 1

Overview

1.1 Observational Study: CEASAR

Prostate cancer is the second leading cause of cancer death among American men behind lung cancer. The common treatment choices for localized disease are surgery, radiation, and observation (active surveillance). For localized prostate cancer, 5-year survival is nearly 100%, and in comparative effectiveness studies, patient-reported disease-specific functional outcomes are often used as the primary endpoint. The Comparative Effectiveness Analysis of Surgery and Radiation (CEASAR) study[1] assessed patient-reported functional outcomes and health-related quality of life at 3 years after treatment.

Suppose we are interested in comparing the Sexual Functional Score (QOL) after 3 years from treatment.

```
groupSum(d$Epic36, d$Treatment, Combined = FALSE)

             N Min    Q1  Med Q3 Max Mean   SD   SE
Surgery   1222   0 10.00 33.3 70 100 41.0 33.4 0.96
Radiation  691   0  6.67 38.3 70 100 40.4 33.5 1.27
```

```
with(d, t.test(Epic36 ~ Treatment))


Welch Two Sample t-test

data:  Epic36 by Treatment
```

[1] Barocas et al., “Association between radiation therapy, surgery, or observation for localized prostate cancer and patient-reported outcomes after 3 years” *JAMA*. 2017. **317**(11):1126-1140.

```
t = 0.4, df = 1432, p-value = 0.7
alternative hypothesis: true difference in means between group Surgery and group Radiation is
95 percent confidence interval:
 -2.51  3.74
sample estimates:
  mean in group Surgery mean in group Radiation
                   41.0                    40.4
```

This shows there is no statistically significant difference in QOL. However, it is well-known that the patient populations for Surgery and Radiation are very different. The baseline QOL is quite different between the groups as shown below.

```
groupSum(d$Epic00, d$Treatment, Combined = FALSE)

             N Min   Q1 Med Q3 Max Mean   SD   SE
Surgery   1388   0 41.7  80 95 100 65.9 32.8 0.88
Radiation  853   0 23.3  60 85 100 54.5 33.1 1.13
```

```
with(d, t.test(Epic00 ~ Treatment))


Welch Two Sample t-test

data:  Epic00 by Treatment
t = 8, df = 1793, p-value = 3e-15
alternative hypothesis: true difference in means between group Surgery and group Radiation is
95 percent confidence interval:
  8.61 14.24
sample estimates:
  mean in group Surgery mean in group Radiation
                   65.9                    54.4
```

As Table below shows, some of other baseline characteristics that are probably associated with the post-treatment QOL are highly different between groups.

Table 1.1: CEASAR baseline characteristics

	N	Surgery	Radiation	Test Statistic
		$N = 1455$	$N = 908$	
Age at diagnosis	2363	57 62 66	63 68 73	$F_{1,2361}$=417, P<0.001[1]
Race : White	2363	75% (1088)	73% (662)	χ^2_3=10.4, P=0.016[2]
Black		12% (175)	16% (148)	
Hispanic		8% (113)	6% (57)	
Others		5% (79)	5% (41)	
TIBI cat : 0-2	2279	33% (468)	20% (175)	χ^2_3=98.6, P<0.001[2]
3-5		55% (766)	55% (478)	
6-8		10% (147)	19% (169)	
9-15		2% (22)	6% (54)	
DAmico Prostate Cancer Risk : Low Risk	2357	41% (597)	34% (307)	χ^2_2=17, P<0.001[2]
Intermediate Risk		42% (613)	44% (395)	
High Risk		17% (243)	22% (202)	
PSA at diagnosis, corrected	2363	4.2 5.1 6.9	4.5 5.9 8.5	$F_{1,2361}$=49, P<0.001[1]
Marital Status : Not married	2262	17% (234)	25% (215)	χ^2_1=21.2, P<0.001[2]
Married		83% (1159)	75% (654)	
Education : Less than high school	2266	9% (120)	13% (111)	χ^2_4=15, P=0.005[2]
High school graduate		21% (294)	21% (186)	
Some college		22% (306)	24% (208)	
College graduate		24% (336)	21% (185)	
Graduate/professional school		24% (340)	21% (180)	
Income : Less than $30,000	2132	17% (228)	28% (224)	χ^2_3=67.7, P<0.001[2]
$30,001 - $50,000		17% (223)	23% (187)	
$50,001 - $100,000		33% (439)	29% (238)	
More than $100,000		33% (432)	20% (161)	
SF36 Physical Score	2287	85 100 100	75 90 100	$F_{1,2285}$=87.4, P<0.001[1]
EPIC Sexual Function -Baseline	2241	41.7 80.0 95.0	23.3 60.0 85.0	$F_{1,2239}$=76.9, P<0.001[1]
EPIC Sexual Function -3 years	1913	10.00 33.33 70.00	6.67 38.33 70.00	$F_{1,1911}$=0.4, P=0.528[1]

$_a$ b $_c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables. N is the number of non–missing values. Numbers after percents are frequencies. Tests used: [1]Wilcoxon test; [2]Pearson test

In general, establishing a cause-and-effect association from an **observational** study is difficult due to **confounders**.

Confounder A prognostic factor that is associated with both response (e.g., Quality of Life) and explanatory variable (e.g., treatment choice).

We can analyze the data with a method that accounts for the baseline difference in the treatment groups.

$$\text{QOL} \sim \text{Treatment} \times (\text{Baseline QOL} + \text{Age} + \text{Race} + \text{TIBI} + \text{Risk} + \text{PSA})$$

How about comorbidities? sex? smoking?
Many statistical methods exist to establish causal relationships from an observation study such as propensity scores and instrumental variables.

Can observational studies establish a cause-and-effect association?

Philip Morris International
`https://www.pmi.com/our-business/about-us/our-views/health-effects-of-smoking-tobacco`

- "Cigarette smoking causes serious disease and is addictive."
- "All cigarettes are harmful and addictive."
- "Public health authorities have concluded that secondhand smoke causes diseases, including lung cancer and heart disease, ..."

JT
`https://www.jti.co.jp/tobacco/responsibilities/guidelines/responsibility/index.html`
`https://www.jti.com/about-us/our-business/our-six-core-principles`

- "Smoking is a cause of serious diseases including lung cancer, coronary heart disease, emphysema and chronic bronchitis."
- "All relevant risk factors need to be taken into consideration when investigating the cause or causes of a disease in any smoker."

1.2 Experiment

Observational study A study design in which the investigator does not control the assignment of treatment of individual study subjects (Piantadosi[2])

[2] Clinical Trials: A Methodologic Perspective

Experiment A study in which the investigator makes a series of observations under controlled/arranged conditions. In particular, the investigator controls the treatment applied to the subjects by design. (Piantadosi)

> **Clinical trial** A prospective study comparing the effect and value of an intervention against a control in human subjects (Friedman[3])

Advantages of observational studies include:

- Lower cost.
- Greater timeliness.
- A broad range of patients.
- Greater application where experiments would be impossible or unethical.

The advantage of clinical trials is that they can establish a cause-and-effect association.

1.2.1 Example: PIVOT

In Prostate Cancer Intervention Versus Observation Trial[4] (PIVOT), prostate cancer patients who were good candidates for radical prostatectomy were enrolled from 1994 to 2002. The last observation was made in 2010. The results were presented at American Urological Association Annual Meeting in May, 2011. The *inclusion criteria* for the study were:

- 75 years or younger.
- Localized disease.
- PSA $\leq 50mg/mg$.
- Diagnosed with 12 months.
- Radical prostatectomy candidate.

With the all-cause mortality as the primary endpoint, the primary objective was to answer the following question:

> Among men with clinically localized prostate cancer detected during the early PSA era, does the intent to treat with radical prostatectomy reduce all-cause & prostate cancer mortality compared to observation?

[3] Fundamentals of Clinical Trials

[4] Wilt et al. (PIVOT Study Group). “Radical prostatectomy versus observation for localized prostate cancer”. *N Engl J Med*. 2012. **367**(3):203–213.

- 13,022 men entered into screening registry.
- 5,023 were eligible.
- 4,292 declined to participate.
- 731 were randomized. (364 prostatectomy, 367 observation)
- Radical prostatectomy was performed on 281 (77%) of the prostatectomy group and 35 (10%) of the observation group. The following table summarized the assigned and received treatments.

	Actual Treatment			
Assigned Treatment	Surgery	Observation	Other	
Surgery	281 (77%)	53 (15%)	30 (8%)	364
Observation	36 (10%)	292 (80%)	39 (11%)	367
	317 (43%)	345 (47%)	69 (9%)	731

Intention-to-treat analysis compares 364 surgery patients and 367 observation patients based on their assigned treatments.

As-treated analysis compares 317 surgery patients and 345 observation patients based on their received treatments.

Per-protocol analysis compares 281 surgery patients and 292 observation patients who adhered to the protocol.

Conclusions: “Among men with localized prostate cancer detected during the early era of PSA testing, radical prostatectomy did not significantly reduce all-cause or prostate-cancer mortality, as compared with observation, through at least 12 years of follow-up. Absolute differences were less than 3 percentage points.”

Chapter 2

Multiplicity in Clinical Trials -FDA's Guidance-

2.1 Background

The problem When K hypothesis tests are conducted with type I error rate of α, the overall type I error rate becomes higher than α.
The overall type I error rate $=$ P[At least one type I error] $=$ Family-wise error.

Suppose there are 8 hypothesis tests, and each is conducted at 5% level. Then

$$\begin{aligned} \text{P[At least one type I error]} &= 1 - \text{P[no type I error]} \\ &= 1-(1-0.05)^8 \\ &= 0.337 \end{aligned}$$

And to control the family-wise type I error rate at 5%, each test must be conducted at $\alpha = 0.00639$ because

$$1-(1-0.00639)^8 = 0.05$$

Recent development

2016.12 EMA: Guideline on multiplicity issues in clinical trials

2017.1 FDA: Draft guidance. Multiple endpoints in clinical trials

2017.8 Stat in Med: A Dmitrienko and RB D'Agostino. "Editorial: Multiplicity issues in clinical trials"

2018.1 Journal of Biopharm Stat: Special issue on multiplicity issues in clinical trials

2018.5 NEJM: A Dmitrienko and RB D'Agostino. "Multiplicity considerations in clinical trials"

Guidelines

- EMA: Guideline on multiplicity issues in clinical trials (Draft)
  - 15 pages
  - Draft published on 12/15/2016
  - Replaces "Points to consider on multiplicity issues in clinical trials" (Adopted 2002)
- FDA: Multiple endpoints in clinical trials: Guidance for industry
  - 50 pages
  - Draft published in January 2017
  - Details on statistical methods in addition to general principles

Prespecification is necessary.
"An important principle for controlling multiplicity is to prospectively specify all planned endpoints, time points, analysis populations, and analyses."

Multiplicity topics / sources of multiplicity

- Multiple endpoints
  - Primary endpoint family
  - Secondary endpoint family
  - (Exploratory endpoints)
  - Co-primary endpoints
  - Composite endpoints/multi-component
- Multiple looks
  - Interim analyses ("outside the scope" (FDA))
- Multiple analyses
  - Subgroup analyses
  - Multiple analyses methods

- Superiority/Non-inferiority (FDA)
- Safety variables (EMA)
- Multiple treatment arms (EMA)
- Dose-response studies (EMA)
- Estimation (EMA)

Endpoints

- Primary endpoints
  "Success on any one alone could be considered sufficient to demonstrate the drug's effectiveness"
- Secondary endpoints
  Provide additional evidence of efficacy.
- Exploratory endpoints
  "All other endpoints" (Is adjustment necessary?)
  "endpoints that are thought to be less likely to show an effect but are included to explore new hypotheses"

  Endpoints are frequently ordered by

  - clinical importance (Mortality as primary)
  - the likelihood of demonstrating an effect (PFS as primary, OS as secondary)

Composite endpoint

Combine clinical outcomes into a single variable.

- Cardiovascular death OR heart attack OR stroke. (coronary artery disease)
- MAKE 30 (Major Adverse Kidney Events: impaired renal function OR hemodialysis OR death (AKI)

"Analyses of the components of the composite endpoint are important and can influence interpretation of the overall study results."

Co-primary endpoints

Demonstration of treatment effects on more than one endpoint is necessary to conclude efficacy. FDA requires each test be done at 5% level.

> Relaxation of alpha ... would undermine the assurance of an effect on each disease aspect considered essential to showing that the drug is effective.

$\alpha = 0.22$ $(0.22^2 \approx 0.05)$ if independent.
Type II error rate inflation may be severe. $(0.05^2 = 0.0025)$

- Co-endpoints are likely to be correlated, but the correlation is unknown.
- Contrast this to a group sequential design where the correlation of accumulated data can be computed.
  $Cov(Z_j, Z_k) = \sqrt{N_j/N_k}$.
  $\alpha_1 = 0.030$ and $\alpha_2 = 0.030$ for the Pocock boundary.

2.2 Statistical methods

Statistical methods to control family-wise type I error rate (FWER)

- “Two-arm trials that examine treatment versus control on multiple endpoints”
- “Similar considerations: different time points, different doses.

Two types

- Single-step procedures
- Multistep procedures (step-down, step-up, sequential procedures)
  - Generally more efficient (power)
  - Confidence interval not readily available

1. The Bonferroni Method
2. The Holm Method
3. The Hochberg Method
4. Prospective Alpha Allocation Scheme
5. The Fixed-Sequence Method
6. The Fallback Method
7. Gatekeeping Testing Strategies
8. The Truncated Holm and Hochberg Procedures for Parallel Gatekeeping
9. Multi-Branched Gatekeeping Procedures

10. Resampling-Based, Multiple-Testing Procedures

Common Statistical Method

1. Bonferroni (Single step; assumption free)
   Each hypothesis test is conducted at α/K level. $\alpha/K \approx$ the solution, a, to $\alpha = 1-(1-a)^K$.

2. Holm (Multi-step step down; assumption free)
   $H_1, \cdots, H_m$ is a family of m null hypotheses, and $P_1, \cdots, P_m$ are the corresponding P-values. The ordered P-values, $P_{[i]}$ are compared to $\alpha/(m+1-k)$, and let k be the smallest index such that $P_{[i]} > \alpha/(m+1-i)$. Reject the null hypotheses $H_{(1)}, \cdots, H_{(k-1)}$.

Example:
$\alpha = 0.05$. $p_1 = 0.015$, $p_2 = 0.03$, $p_3 = 0.04$, $p_4 = 0.01$.

Bonferroni method:
Each p-value is compared to $\alpha/4 = 0.013$. Only H_4 is rejected.

Holm method:
$p_{[1]} = 0.01$, $p_{[2]} = 0.015$, $p_{[3]} = 0.03$, $p_{[4]} = 0.04$.
The corresponding critical values are
$0.05/4 = 0.013$, $0.05/3 = 0.017$, $0.05/2 = 0.025$, and 0.05.

3. Hochberg (Multi-step step up; Positive correlation)
   Similar to Holm but backwards. Start comparing the largest p-value to α and work the way down to the smallest p-value. Reject all H_0 once $p_{[k]} < \alpha/(m+1-k)$.

4. Prospective Alpha Allocation Scheme (Single step; Positive correlation)

   Similar to Bonferroni, but use α_1, α_2, $\cdots$, α_m such that

$$(1-\alpha_1) \times \cdots \times (1-\alpha_m) = (1-\alpha).$$

$$\text{Example: } (1-0.00639)^8 = 1-0.05$$

FDA on Hochberg:

> $\cdots$ beyond the aforementioned cases where the Hochberg procedure is known to be valid, its use is generally not recommended for the primary comparisons of confirmatory clinical trials unless it can be shown that adequate control of Type I error rate is provided.

Sequential method

5. Fixed-Sequence Method
   Tests endpoints in a predefined order, all at $\alpha = 0.05$, moving to the next endpoint only after a success on the previous endpoint.

6. Fallback Method
   Fixed-sequence method with some α "saved" for later use. (e.g., 0.03 on the first test, 0.02 on the second.)

Suppose $p_A = 0.045$, $p_B = 0.015$, $p_C = 0.065$, and the sequence is C, B, A.

	A (0.045)	B (0.015)	C (0.065)
Bonferroni	not Reject	Reject	not Reject
Holm	not Reject	Reject	not Reject
FSM (C,B,A)	not Reject	not Reject	not Reject
FSM (A,B,C)	Reject	Reject	not Reject
Fallback (C,B,A)	not Reject	Reject	not Reject

Gatekeeping testing strategy

The gatekeeping testing strategy tests the primary and secondary families sequentially with $\alpha = 0.05$ for the primary family and with some α passed on to the secondary family.

Serial gatekeeping strategy The primary family are tested as co-primary endpoints ($\alpha = 0.05$). The secondary family is tested only if all primary null hypotheses are rejected (at $\alpha = 0.05$).

Parallel gatekeeping strategy The primary family uses a strategy that allows passing of an individual α, and the secondary family allocates the passed-on (accumulated) amount.

Parallel gatekeeping strategy

- Primary endpoints (A, B)
  $\alpha = 0.05$
  Bonferroni with 0.04 for A and 0.01 for B.

- Secondary endpoints (C, D, E)
  $\alpha = 0$
  Holm.

Suppose $p_A = 0.035$, $p_B = 0.055$, $p_C = 0.010$, $p_D = 0.045$, $p_E = 0.019$.

A (0.035)	B (0.055)	C (0.010)	D (0.045)	E (0.019)
Reject	not Reject			
0.04 is passed to (C, D, E) Critical values: 0.0133, 0.02, 0.04				
		Reject	not Reject	Reject

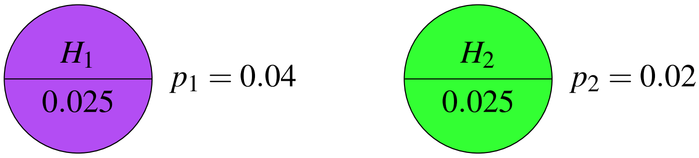
2.3 Graphical approach

Using **gMCP** package in R.

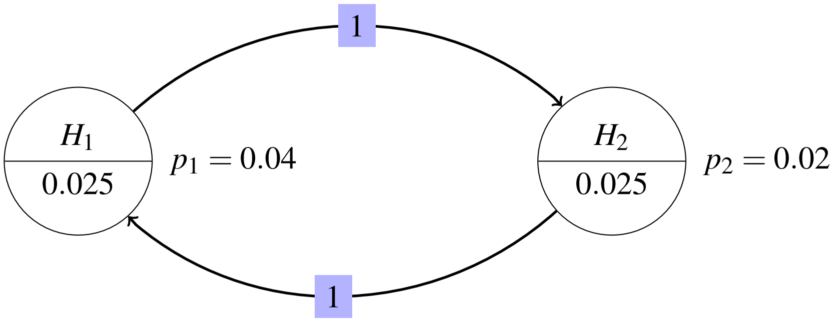
- K Rohmeyer, F Klinglmueller (2018). gMCP: Graph Based Multiple Test Procedures. R package version 0.8-14.

- F Bretz, M Posch, EGlimm, FKlinglmueller, W Maurer, K Rohmeyer (2011), Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes or parametric tests. Biometrical Journal 53(6), pages:894-913.

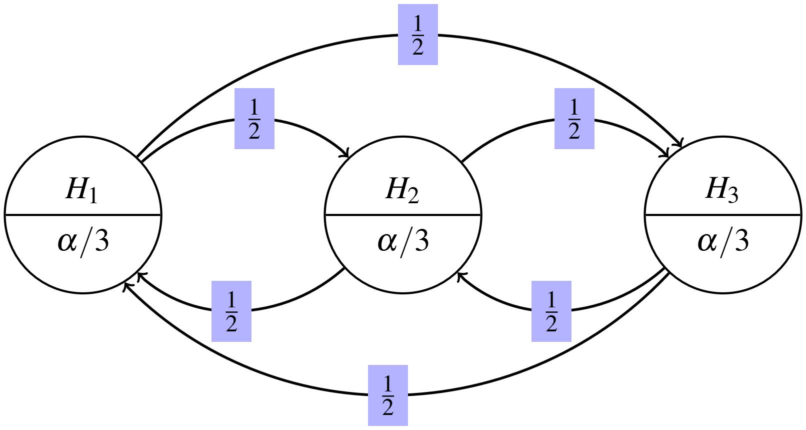
Bonferroni



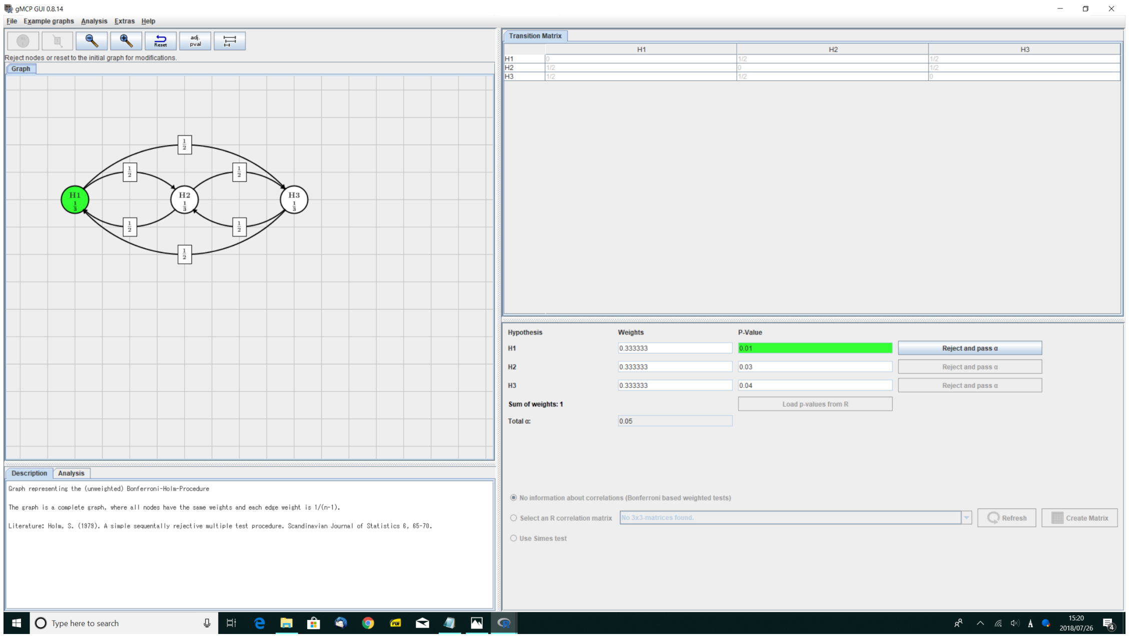
Holm



Holm procedure with 3 hypotheses



Holm using gMCP Package: Step 1



Reject nodes or reset to the initial graph for modifications.

Graph representing the (unweighted) Bonferroni-Holm-Procedure

The graph is a complete graph, where all nodes have the same weights and each edge weight is 1/(n-1).

Literature: Holm, S. (1979). A simple sequentally rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65-70.

	H1	H2	H3
H1	0	1/2	1/2
H2	1/2	0	1/2
H3	1/2	1/2	0

Hypothesis	Weights	P-Value
H1	0.333333	0.01
H2	0.333333	0.03
H3	0.333333	0.04

Sum of weights: 1

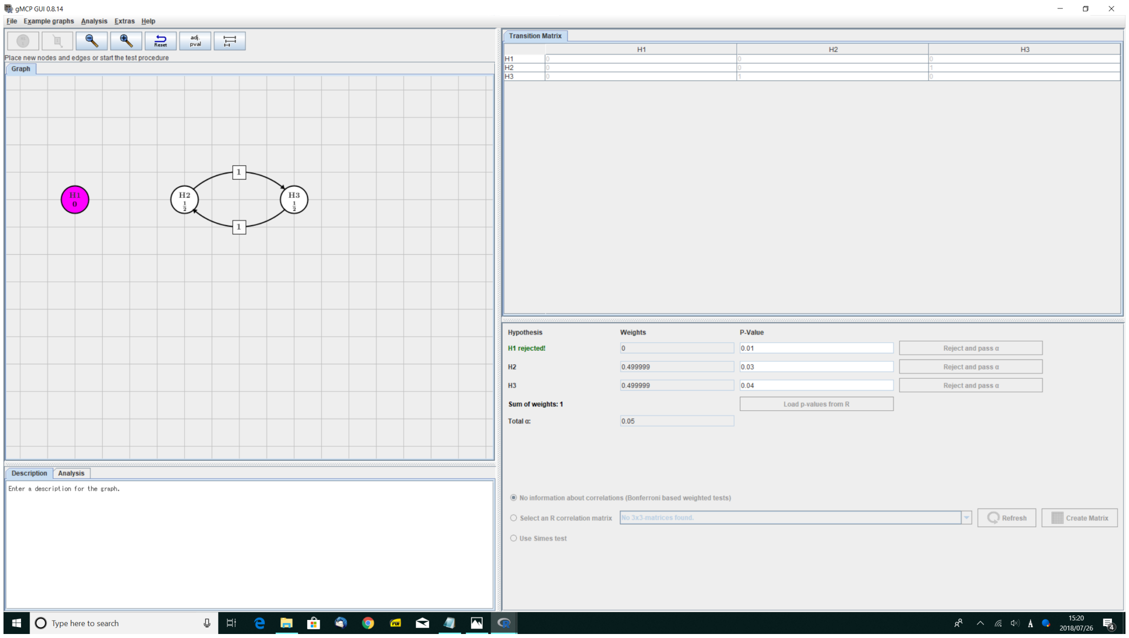
Total α: 0.05

● No information about correlations (Bonferroni based weighted tests)

○ Select an R correlation matrix No 3x3-matrices found.

○ Use Simes test

Holm using gMCP Package: Step 2



Place new nodes and edges or start the test procedure

Enter a description for the graph.

Transition Matrix

	H1	H2	H3
H1	0	0	0
H2	0	0	1
H3	0	1	0

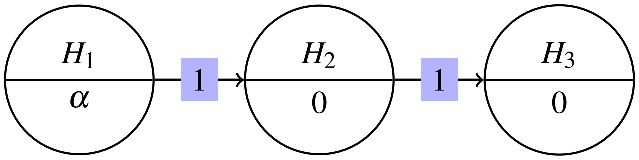
Hypothesis	Weights	P-Value	
H1 rejected!	0	0.01	Reject and pass α
H2	0.499999	0.03	Reject and pass α
H3	0.499999	0.04	Reject and pass α
Sum of weights: 1		Load p-values from R	
Total α:	0.05		

● No information about correlations (Bonferroni based weighted tests)

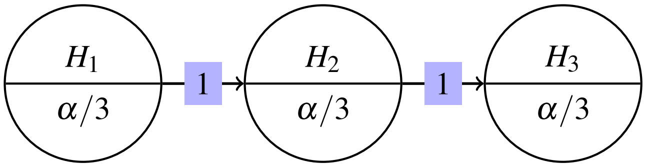
○ Select an R correlation matrix No 3x3-matrices found. Refresh Create Matrix

○ Use Simes test

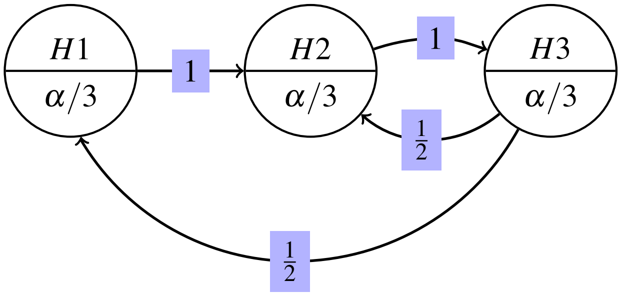
Fixed sequence method



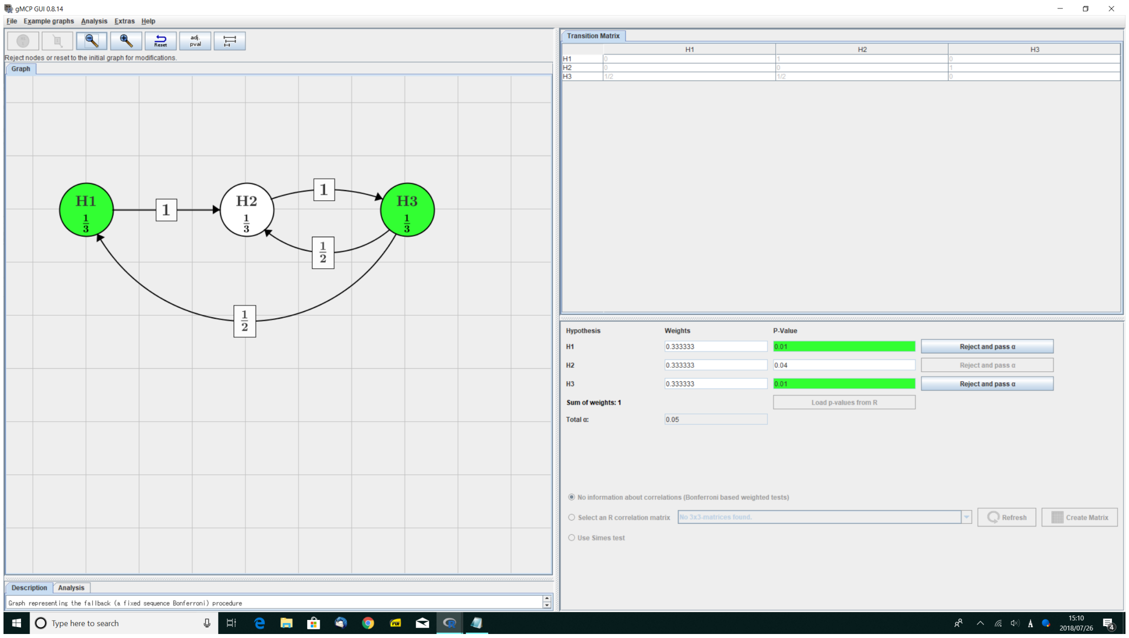
Fall back method



Improved fall back method



Improved fall back with gMCP package: Step 1



Reject nodes or reset to the initial graph for modifications.

Graph representing the fallback (a fixed sequence Bonferroni) procedure

	H1	H2	H3
H1	0	1	0
H2	0	0	1
H3	1/2	1/2	0

Hypothesis	Weights	P-Value
H1	0.333333	0.01
H2	0.333333	0.04
H3	0.333333	0.01

Sum of weights: 1

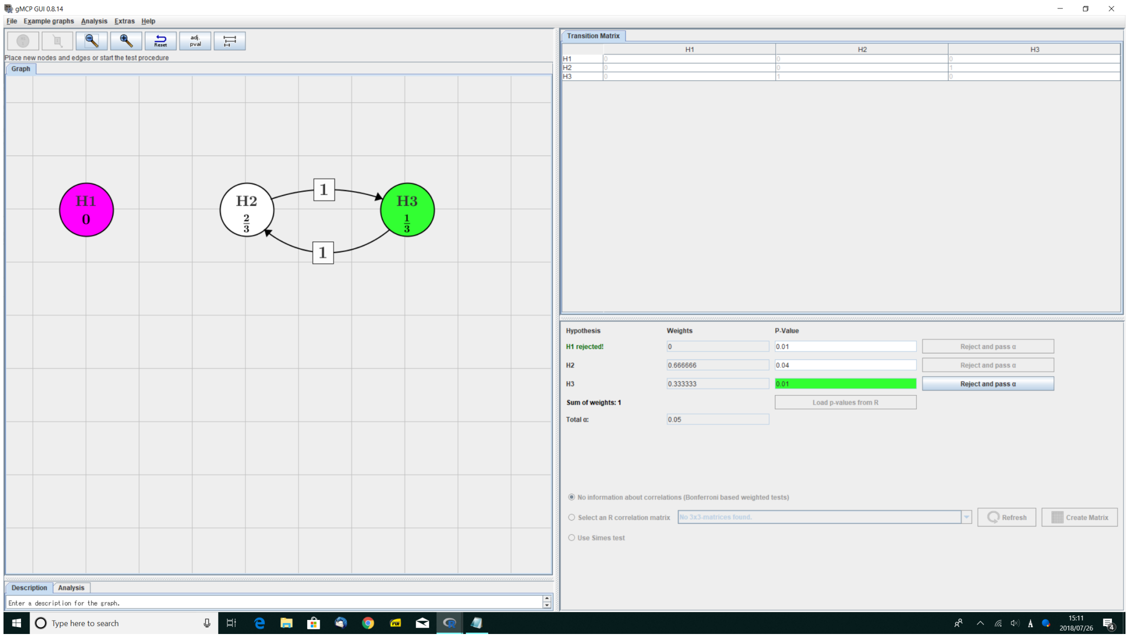
Total α: 0.05

● No information about correlations (Bonferroni based weighted tests)

○ Select an R correlation matrix No 3x3-matrices found.

○ Use Simes test

Improved fall back with gMCP package: Step 2



Place new nodes and edges or start the test procedure

Enter a description for the graph.

	H1	H2	H3
H1	0	0	0
H2	0	0	1
H3	0	1	0

Hypothesis	Weights	P-Value
H1 rejected!	0	0.01
H2	0.666666	0.04
H3	0.333333	0.01

Sum of weights: 1

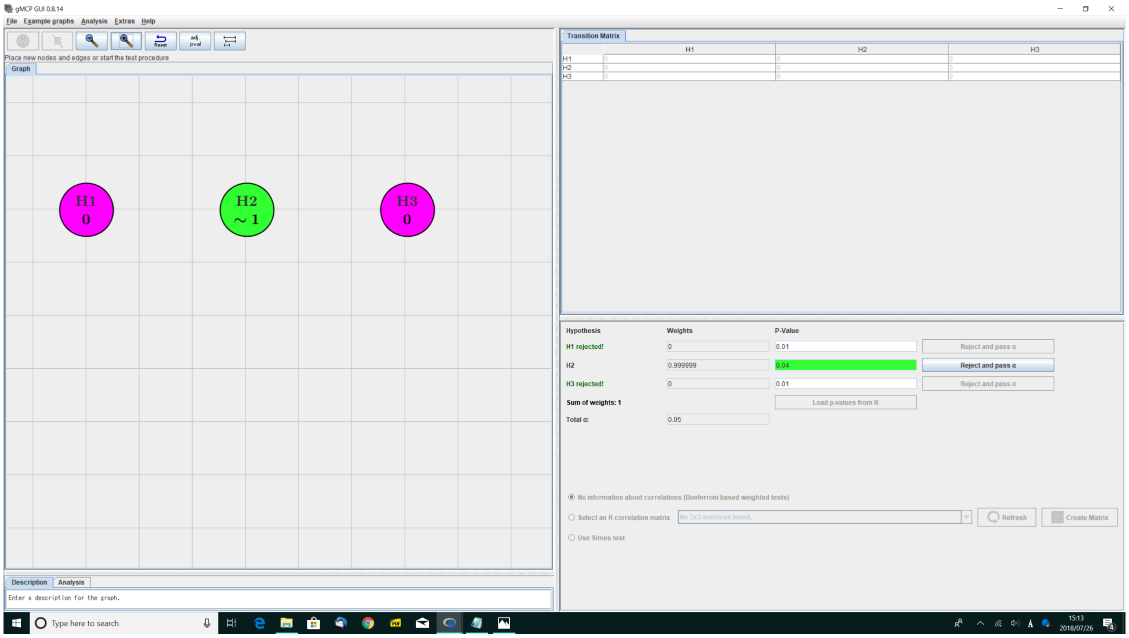
Total α: 0.05

● No information about correlations (Bonferroni based weighted tests)

○ Select an R correlation matrix No 3x3-matrices found.

○ Use Simes test

Improved fall back with gMCP package: Step 3



Place new nodes and edges or start the test procedure

Transition Matrix

	H1	H2	H3
H1	0	0	0
H2	0	0	0
H3	0	0	0

Hypothesis	Weights	P-Value	
H1 rejected!	0	0.01	Reject and pass α
H2	0.999999	0.04	Reject and pass α
H3 rejected!	0	0.01	Reject and pass α
Sum of weights: 1		Load p-values from R	
Total α:	0.05		

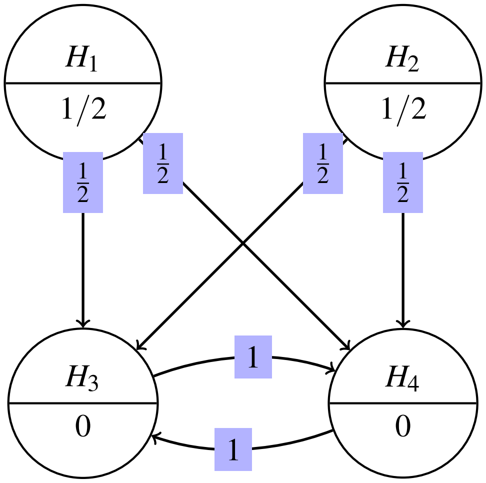
● No information about correlations (Bonferroni based weighted tests)

○ Select an R correlation matrix No 3x3-matrices found. Refresh Create Matrix

○ Use Simes test

Enter a description for the graph.

A Parallel gatekeeping procedure



Primary

Secondary

Chapter 3

Randomization

3.1 Example: Polio vaccine trial (1954)

In 1954, 1.8 million children participated in the largest clinical trial to date to assess the effectiveness of the vaccine developed by Jonas Salk in preventing paralysis or death from poliomyelitis.

- 1.8 million children in selected school districts throughout the US were involved in this placebo-controlled trial.
  - Why was placebo necessary?
  - $60,000$ cases in 1952; about half of that in 1953.
- Randomized trial.
  - $750,000$ children participated.
  - They required parents' consent.
  - Half of the children with consent were randomized into the vaccine group.
- NFIP design.
  - The National Foundation for Infantile Paralysis (NFIP) conducted a study in which all 2nd graders with consent received the vaccine with 1st and 3rd graders acting as control.
  - $1,125,000$ children participated.
  - The control children did not require consent.
    Systematic difference between groups.
  - No blinding.
  - Polio is a contagious disease!

The results of the SVF trial are tabulated below[1].

The randomized double-blind design

	Size	Rate*
Treatment	$200,000$	26
Control	$200,000$	71
No consent	$350,000$	46

(*Rate of polio cases per $100,000$)

The NFIP design

	Size	Rate*
Treatment	$225,000$	25
Control	$725,000$	54
No consent	$125,000$	44

(*Rate of polio cases per $100,000$)

3.2 Introduction

Randomization Assignment of patients or experimental subjects to two or more treatments by chance alone.

Main advantages of randomization

- It removes the potential of bias in the allocation of participants to the intervention group or to the control group (allocation bias).
- It tends to produce similar (compatible) groups in terms of measured as well as unmeasured confounders
  confounding by indication in observational studies.

Randomization is considered so important that the intention-to-treat principle considered sacrosanct: "Analyze by assigned treatments irrespective of actual treatment received."
Perceived disadvantages of randomization are often about emotional and ethical issues.
→ randomization before consent
Predecessors to randomization:

- Alternating assignments (TCTCTCTC...).

[1] Freedman et al. Statistics, second edition

- Treatment assignment based on birthday / day of the week.

The primary problems with these non-random assignment are the lack of assurance of comparability (baseline balance). An additional issue with the "alternating assignments" is that if one is unblinded, all the rest are unblinded, too.

3.3 Simple randomization

For each subject, flip a coin to determine treatment assignment. P[treatment 1] $= \cdots =$ P[treatment k] $= 1/k$.
Problems with simple randomization and how to deal with them.

- Imbalance in treatment allocation.
  - Replacement randomization.
  - Block randomization.
  - Adaptive randomization. (Biased coin / Urn model etc.)
- Imbalance in baseline patient characteristics.
  - Stratified randomization. (Stratified permuted block randomization)
  - Covariate adaptive randomization. (Minimization randomization)

3.4 Imbalance in treatment allocation

If the number of patients, N is 20, $P[10$ and $10] = 0.18$. The probability of 7 to 13 split or worse is 26%. The treatment effect variance for $7-13$ split relative to $10-10$ split is

$$\left(\frac{1}{7}+\frac{1}{13}\right) / \left(\frac{1}{10}+\frac{1}{10}\right) = 1.098.$$

7-13 split is only $1/1.098 = 0.92$ as efficient as 10-10 split.
Even if treatment allocation is balanced at the end of trial, there may be a (severe) imbalance at some point. Because we may monitor trials over time, we prefer to have balance over time.

3.4.1 Block randomization

To ensure a better balance (in terms of number of patients) across groups over time, consider a block randomization (random permuted blocks).

Block randomization ensures approximate balance between treatments by forcing balance after a small number of patients (say 4 or 6). For example, the first 4 patients are allocated to treatment A or B sequentially based on $AABB$.
There are 6 sequences of A, A, B, B, and let each sequence have 1/6 chance of being selected.

$$AABB \qquad ABAB \qquad ABBA \qquad BAAB \qquad BABA \qquad BBAA$$

```
for (i in 1:5) {
    cat(i, sample(rep(LETTERS[1:2], each = 2), 4, replace = FALSE),
        "\n")
}

1 B A A B
2 B B A A
3 A A B B
4 A B A B
5 B A A B
```

What's wrong with block size of 2? Block size of 200?
Easily applicable to more than 2 groups (A, B, C)

```
for (i in 1:5) {
    cat(i, sample(rep(LETTERS[1:3], each = 2), 6, replace = FALSE),
        "\n")
}

1 C C B B A A
2 B C A C A B
3 B B A C A C
4 A C C B A B
5 A C C A B B
```

Easily applicable to unequal group sizes ($N_a = 40$ and $N_b = 20$).

```
for (i in 1:5) {
    cat(i, sample(rep(LETTERS[1:2], c(4, 2)), 6, replace = FALSE),
        "\n")
}

1 A B A B A A
```

```
2 B A A A B A
3 A B A A B A
4 B A A A B A
5 A B A A A B
```

Why might we want unequal group sizes?

- We may want to have a better estimate of the effect for the new treatment.

- Treatment costs may be very different.
  Given the total sample size and the relative cost of treatment 2 to treatment 1, we can find the optimal allocation ratio to minimize the total cost.

- Variances may be different.
  Suppose the means, μ_1 and μ_2, of treatment groups are being compared using

  $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

  For a given $N = n_1 + n_2$, the test statistic is maximized when the denominator is minimized. Solving

  $$\frac{\partial}{\partial n_1}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{N - n_1}\right) = 0$$

  we get

  $$\frac{n_1}{N} = \frac{\sigma_1}{\sigma_1 + \sigma_2}.$$

  Therefore, the optimal allocation ratio is $r = n_1/n_2 = \sigma_1/\sigma_2$.

Analysis should account for the randomization scheme but often does not. Matts and McHugh (1978 *J Chronic Dis*) point out that

- because blocking guarantees balance between groups and increases the power of a study, blocked randomization with the appropriate analysis is more powerful than not blocking at all or blocking and ignoring it in the analysis.

- not accounting for blocking in analysis is conservative.

3.4.2 Biased coin and urn model

These techniques are sometimes classified as “adaptive randomization”.
Allocation of i-th patient depends on how many have been randomized to group A (n_a) and group B (n_b).
Any given time, the probability of allocation to group A may be

$$P[A] = \frac{n_b}{n_a + n_b}.$$

Or the rule may be to use $P[A] = 2/3$ when $n_b - n_a > 5$, and $P[B] = 2/3$ when $n_a - n_b > 5$. Characteristics of such a randomization scheme are usually studied by simulations.
An urn model is one type of biased coin randomization.

- Prepare an urn with one Amber ball and one Blue ball.
- Pick one ball and make the corresponding treatment assignment (A/B).
- Put a ball of the opposite color in the urn.

```
urn1 <- function(n) {
    # randomize n patients into A or B.  At any time P[A] =
    # (#B so far + 1) / (#A so far + 1 + #B so far + 1).
    out <- data.frame(matrix(0, ncol = 4, nrow = n + 1))
    out[1, 1] <- 1
    out[1, 2] <- 1
    for (i in 1:n) {
        out[i, 3] <- out[i, 2]/(out[i, 1] + out[i, 2])
        out[i, 4] <- sample(c("A", "B"), 1, prob = out[i, 2:1])
        out[i + 1, 1] <- out[i, 1] + (out[i, 4] == "A")
        out[i + 1, 2] <- out[i, 2] + (out[i, 4] == "B")
    }
    out[, 1] <- out[, 1] - 1
    out[, 2] <- out[, 2] - 1
    names(out) <- c("A so far", "B so far", "P[A next]", "Next")
    out[1:n, ]
}

urn1(n = 10)

   A so far B so far P[A next] Next
1         0        0     0.500    A
```

```
2          1         0      0.333    A
3          2         0      0.250    B
4          2         1      0.400    B
5          2         2      0.500    A
6          3         2      0.429    A
7          4         2      0.375    B
8          4         3      0.444    B
9          4         4      0.500    B
10         4         5      0.545    A
```

3.5 Imbalance in baseline patient characteristics

Block randomization and biased coin model ensure that the group sizes are reasonably balanced. In order to facilitate the comparison of treatment effects, balance on important baseline variables is sometimes desired.

- Randomization does not guarantee all the measured variables will be balanced. And imbalance does not mean randomization did not work.

Senn (1994) It is argued that this practice [testing baseline homogeneity] is philosophically unsound, of no practical value and potentially misleading. Instead it is recommended that prognostic variables be identified in the trial plan and fitted in an analysis of covariance regardless of their baseline distribution (statistical significance).

Piantadosi These methods, while theoretically unnecessary, encourage covariate balance in the treatment groups, which tends to enhance the credibility of trial results.

An annonymous reviewer Since this is a randomized controlled trial, comparison of baseline characteristics (Table 1) is not necessary. The problem with this approach is that when comparing baseline characteristics we already know that the null hypothesis is true if the randomization was done correctly. Thus, we would expect 1 test in 20 to give a ‘significant’ result with $p < 0.05$ by chance alone. The best approach is to specify key prognostic factors to include in multivariable models irrespective of their significance between treatment groups.

3.5.1 Stratified randomization

Stratified randomization is applied to ensure that the groups are balanced on baseline variables that are thought to be significant.

- Create strata based on the variables for which balance is sought.
  e.g., (Male, 65 or younger), (Male, older), (Female, younger), (Female older)

- Randomize to treatments within each stratum. **Use block randomization!**
  What's wrong with

  - using simple randomization within a stratum?
  - using too many strata?

- Stratification should be accounted for in analysis.

  - Pre-randomization stratification and post-randomization stratification (at time of analysis) has no clear winner.

- If trial is large, stratification may not be necessary

- Stratification by center is a good idea from practical viewpoints.

  - Allows randomization to be hosted at each site
  - Allows sites to be removed and still maintains balance

- Block randomization is a special type of stratified randomization where strata are defined by ... .

- If each stratum has a target size, plans need to be in place to close down recruitment based on the baseline characteristics. e.g., "We do not need any more (Male, older)".

3.5.2 Adaptive and minimization randomization

Adaptive randomization can be used to reduce baseline imbalance:

- Define an imbalance function based on factors thought to be important

- Then use a rule to define P[treatment A] so that the next assignment is more likely to reduce imbalance.

For example, the factors to balance are sex (male/female) and hypertension (yes/no), and let the imbalance function be

$$I = 2 \times (\text{sex imbalance}) + 3 \times (\text{hypertension imbalance}).$$

The patients randomized so far are

	Sex		Hypertension	
	Male	Female	Yes	No
Group 1	10	3	8	5
Group 2	8	3	6	5

The next patient is male-non hypertensive. The imbalance will be

$$I = 2 \times (11-8) + 3 \times (6-5) = 9 \text{ if Group 1,}$$
$$I = 2 \times (10-9) + 3 \times (6-5) = 5 \text{ if Group 2.}$$

Thus let P[Group 2] $= 2/3$.
Minimization randomization uses the same idea but use P[Group 2] = 1, to eliminate randomness when there is some imbalance. Randomize only when to assign the next patient to either group gives the same value of I.

3.6 Response adaptive randomization

As the name suggests, response adaptive randomization methods use the information about the response so far to allocate the next patient.
Play the winner: The idea is to allocate more patients in the treatment that seems to be working better. To apply these methods, it is necessary to have a response quickly. Urn model can be used to make treatment assignment imbalance based on the results (success/failure) of each treatment so far. (e.g., put one blue ball if the treatment B yields success.)
Instead of updating the probabilities of treatment assignment after each patient, we can update them after a group of patients' results are available to reduce administrative burden. In a phase II clinical trial, play the winner design may be used to reduce the number of treatments in consideration. (e.g., Only retain the treatment arms that have P[positive response] > 0.4.)

3.6.1 Example: ECMO

Bartlett et al.[2] conducted a randomized study of the use of extracorporeal membrane oxygenation (ECMO) to treat newborns with respiratory failure. A play-the-winner design[3] was used because

- the outcome is known soon after randomization.
- most ECMO patients were expected to survive and most control patients were expected to die.
  - Ethically, the investigators felt obligated not to withhold the lifesaving treatment.

[2] "Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study". (1985) *Pediatrics*
[3] Zelen (1969) *JASA*; Wei and Durham (1978) *JASA*

- – Scientifically, they felt obligated to perform a randomized study.

The randomization plan:

- The first patient will be randomized to ECMO or the conventional treatment (CT) with equal probability.
- For each patient who survives on ECMO or dies on CT, one ECMO ball is added to the urn.
- For each patient who survives on CT or dies on ECMO, one CT ball is added to the urn.
- The trial will be terminated when 10 balls of one kind have been added, and that treatment will be chosen as the winner.

What actually happened:

P(ECMO)=1/2 Patient 1 was randomized to ECMO and survived.

P(ECMO)=2/3 Patient 2 was randomized to CT and died.

P(ECMO)=3/4 Patient 3 was randomized to ECMO and survived

P(ECMO)=4/5 Patient 4 was randomized to ECMO and survived

P(ECMO)=5/6 Patient 5 was randomized to ECMO and survived

P(ECMO)=6/7 Patient 6 was randomized to ECMO and survived

P(ECMO)=7/8 Patient 7 was randomized to ECMO and survived

P(ECMO)=8/9 Patient 8 was randomized to ECMO and survived

P(ECMO)=9/10 Patient 9 was randomized to ECMO and survived

P(ECMO)=10/11 Patient 10 was randomized to ECMO and survived

Randomization was stopped when there were 11 ECMO patients who survived and 1 CT patient who died.
Controversies followed because ...

```r
fisher.test(cbind(c(11, 0), c(0, 1)))
```

```
Fisher's Exact Test for Count Data

data:  cbind(c(11, 0), c(0, 1))
p-value = 0.08
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.282   Inf
sample estimates:
odds ratio
       Inf
```

“In retrospect it would have been better to begin with two or three pairs of balls, which probably would have resulted in more than one control patient.”

3.7 Nonbipartite matching in clinical trials

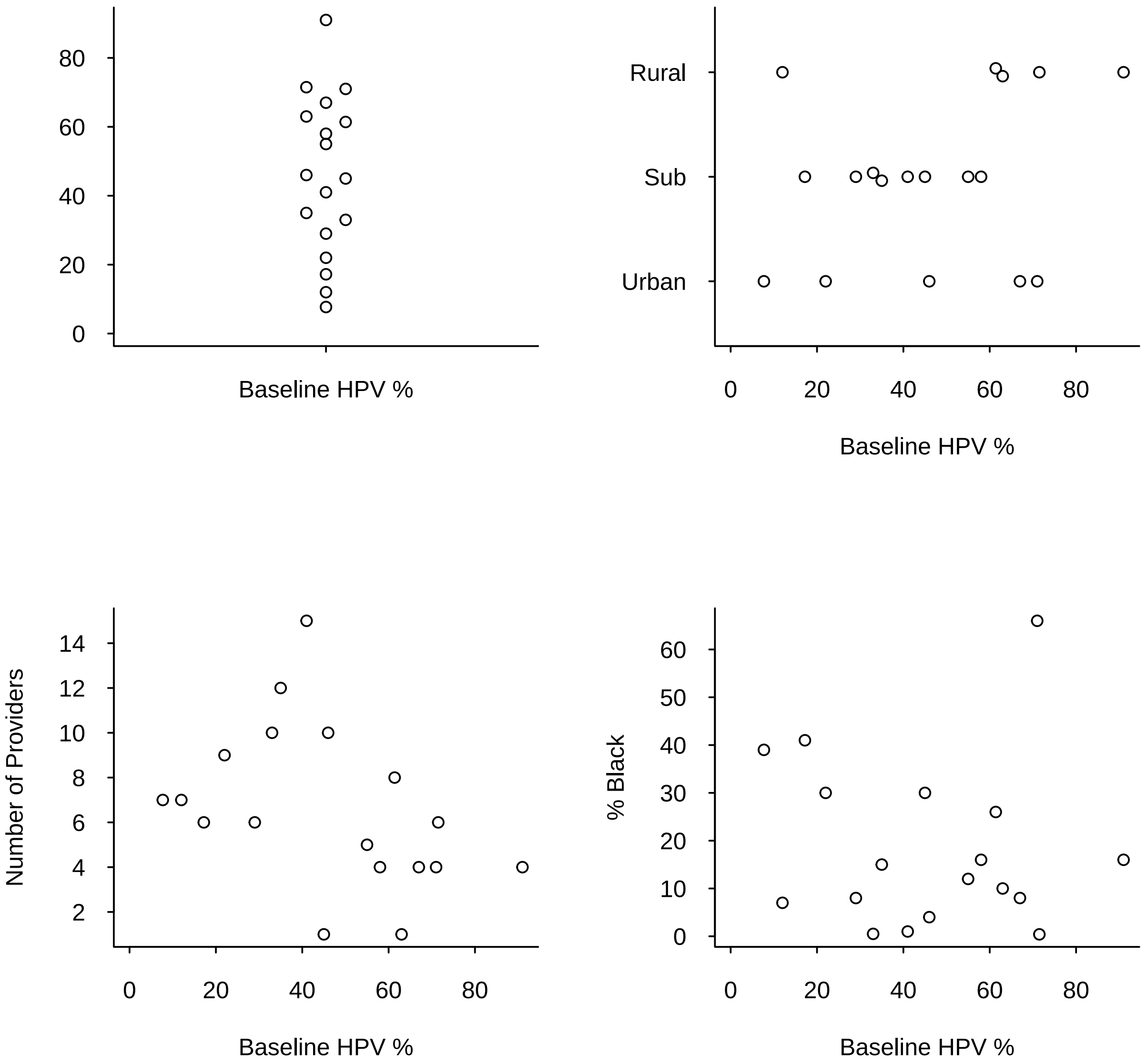
- Reference: Lu B, Greevy R, Xu X, Beck C. “Optimal nonbipartite matching and its statistical applications”. 2011. 65(1):21-30.

- R package “nbpMatching”
  `http://biostat.mc.vanderbilt.edu/wiki/Main/MatchedRandomization`

When baseline data on all the subjects are available, the subjects can be matched, and treatments are randomly assigned within each pair. Matching is done to minimize some form of multidimensional (multivariate) distance, e.g., Mahalanobis distance.

3.7.1 Example

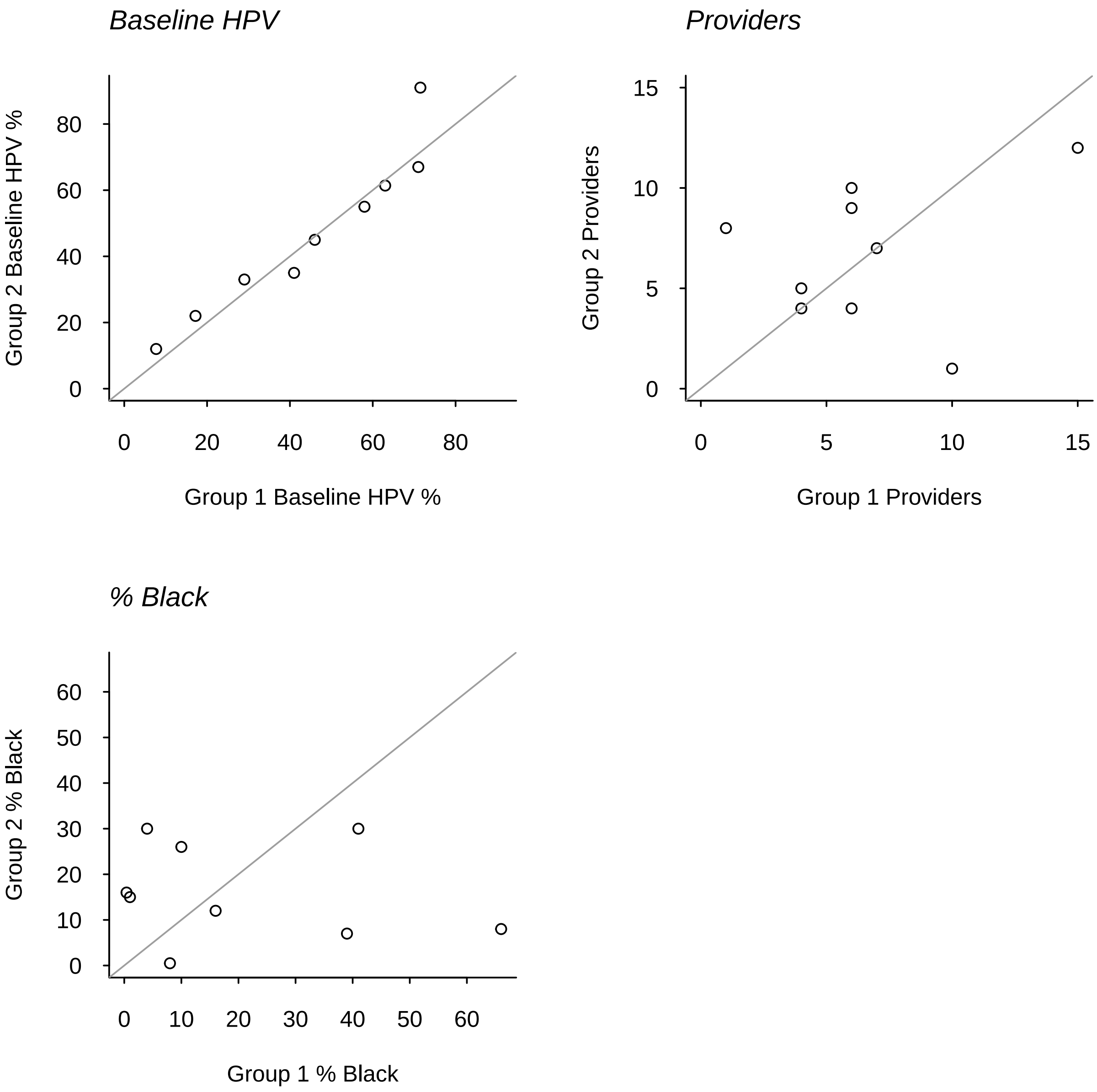
The human papillomavirus (HPV) vaccine will prevent a high proportion of vaginal, oropharyngeal, vulvar, and penile cancers. Yet the proportion of 11- and 12-year old girls who receive this vaccine is not very high. The investigators would like to test effectiveness of the tailored coaching intervention to educate the health-care providers. For this study, 18 community-based, private pediatric practices have been recruited.



By Baseline HPV % Only

```
  Group1 Group2
1   17.2   22.0
2   41.0   35.0
3   29.0   33.0
4   58.0   55.0
```

```
5   46.0   45.0
6    7.7   12.0
7   71.5   91.0
8   63.0   61.4
9   71.0   67.0
```



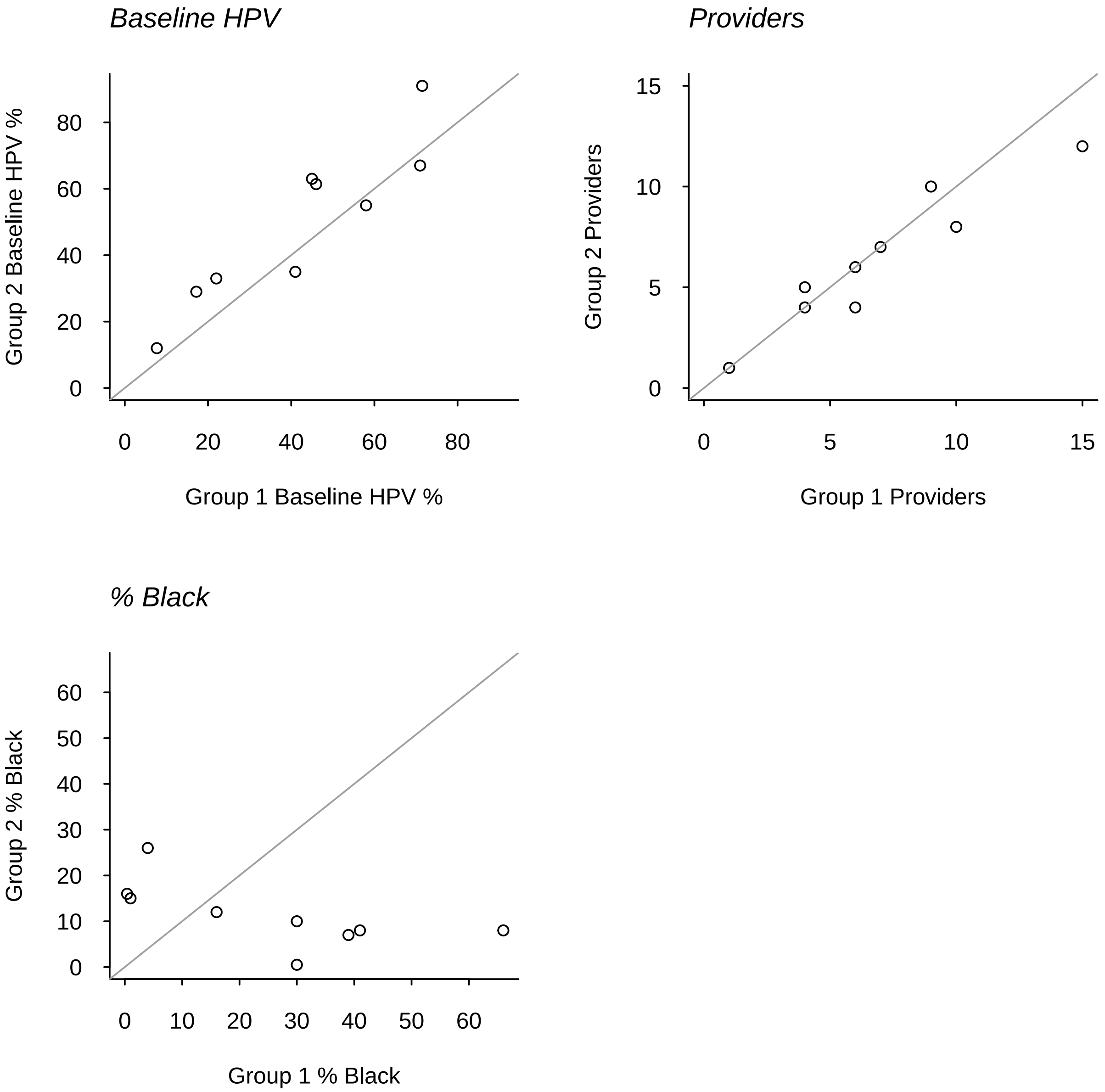
Baseline HPV

Providers

% Black

By Baseline HPV % and Number of Providers

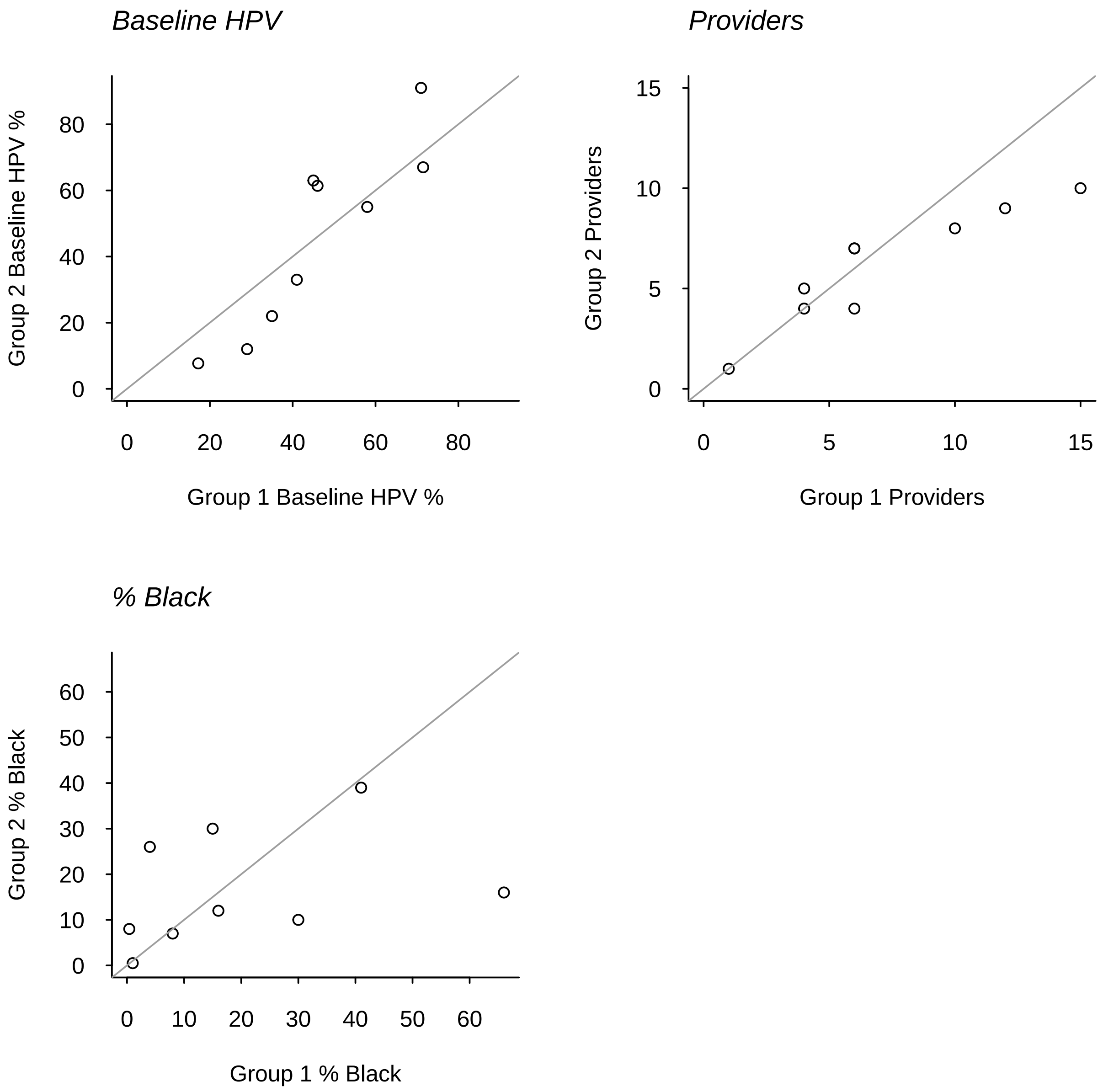
```
  Group1 Group2
1   17.2   29.0
2   41.0   35.0
3   22.0   33.0
4   58.0   55.0
5   46.0   61.4
6    7.7   12.0
7   45.0   63.0
8   71.5   91.0
9   71.0   67.0
```



By Baseline HPV % and Number of Providers and % Black

```
  Group1 Group2
1   17.2    7.7
2   41.0   33.0
3   35.0   22.0
4   29.0   12.0
```

```
5   58.0   55.0
6   46.0   61.4
7   45.0   63.0
8   71.5   67.0
9   71.0   91.0
```



Chapter 4

Superiority, Non-inferiority, and equivalence

4.1 Superiority and non-inferiority

In a phase 3 clinical trial, the objective is often to show the new treatment is better than the conventional treatment. This is called a superiority clinical trial, in which the following hypotheses are tested:

$$H_{S0} : \delta = 0$$
$$H_{S1} : \delta > 0$$

where δ is the difference of the treatment effects. Here we assume larger values of δ indicate a favorable result. We generally do not conduct a two-sided hypothesis test in a clinical trial. Type I error rate is usually set at 2.5%, and power is usually set at 80% or 90% at some clinically meaningful value, δ_S.

When there is a conventional treatment that is known to be effective, it may be of interest to show non-inferiority of the new treatment to the control.

$$H_{N0} : \delta = -\delta_I$$
$$H_{N1} : \delta > -\delta_I$$

Here, δ_I is a pre-specified positive number, which is referred to as the non-inferiority margin. It is customary to set the power of non-inferiority test at 0. Mathematically, superiority testing and non-inferiority testing are very similar; one is a location shift of the other.

After observing the data, $\hat{\delta}$, it is always possible to compute δ_I such that H_0 for the non-inferiority test is rejected. Therefore, it is necessary to define the non-inferiority margin a priori.

A non-inferiority trial usually requires a bigger sample size than a superiority trial does. That is, $\delta_I < \delta_S$. δ_I needs to be small enough to be clinically indifferent, and δ_S needs to be large enough to be clinically meaningful.

Because when the data from the control and treatment groups are similar, it biases towards no difference, the intention-to-treat analysis biases towards positive results in a non-inferiority trial. This can be seen in the following small simulation study.

```
sig <- 4
del <- 1
alp <- 0.025
bet <- 0.1

(n <- ceiling(2 * (qnorm(alp) + qnorm(bet))^2 * sig^2/del^2))

[1] 337

supSim <- function(B, n, mu0, mu1, sig, pSwitch = 0) {
    # pSwitch is the proportion of patients switching the
    # group (T -> C and C -> T are the same.)
    nSwitch <- ceiling(n * pSwitch)
    toSwitch <- sample(1:n, nSwitch)
    X0 <- Y0 <- matrix(rnorm(B * n, mu0, sig), ncol = B)
    X1 <- Y1 <- matrix(rnorm(B * n, mu1, sig), ncol = B)

    for (i in toSwitch) {
        X0[i, ] <- Y1[i, ]
        X1[i, ] <- Y0[i, ]
    }
    XBar0 <- apply(X0, 2, mean)
    XBar1 <- apply(X1, 2, mean)

    zVal <- sqrt(n) * (XBar1 - XBar0)/(sqrt(2) * sig)
    pVal <- 1 - pnorm(zVal)
    table(pVal < 0.025)
}
(supSim0.0 <- supSim(B = 10000, n = n, mu0 = 0, mu1 = 0, sig = 4))


FALSE  TRUE
 9786   214

(supSim1.0 <- supSim(B = 10000, n = n, mu0 = 0, mu1 = 1, sig = 4))


FALSE  TRUE
 1077  8923
```

```
(supSim0.1 <- supSim(B = 10000, n = n, mu0 = 0, mu1 = 0, sig = 4,
    pSwitch = 0.2))


FALSE  TRUE
 9743   257

(supSim1.1 <- supSim(B = 10000, n = n, mu0 = 0, mu1 = 1, sig = 4,
    pSwitch = 0.2))


FALSE  TRUE
 5163  4837

niSim <- function(B, n, del, niMargin = 1, sig, pSwitch = 0) {
    # del is the true difference (>0).  Under null,
    # del=-niMargin; under alternative, del=0.  pSwitch is
    # the proportion of patients switching the group (T ->
    # C and C -> T are the same.)
    nSwitch <- ceiling(n * pSwitch)
    toSwitch <- sample(1:n, nSwitch)
    X0 <- Y0 <- matrix(rnorm(B * n, 0, sig), ncol = B)  # control
    X1 <- Y1 <- matrix(rnorm(B * n, -del, sig), ncol = B)  # new treatment

    for (i in toSwitch) {
        X0[i, ] <- Y1[i, ]
        X1[i, ] <- Y0[i, ]
    }
    XBar0 <- apply(X0, 2, mean)
    XBar1 <- apply(X1, 2, mean)

    zVal <- sqrt(n) * (XBar1 - XBar0 + niMargin)/(sqrt(2) * sig)
    pVal <- 1 - pnorm(zVal)
    table(pVal < 0.025)
}
(niSim0.0 <- niSim(B = 10000, n = n, del = 1, niMargin = 1, sig = 4))


FALSE  TRUE
 9764   236
```

```
(niSim1.0 <- niSim(B = 10000, n = n, del = 0, niMargin = 1, sig = 4))


FALSE  TRUE
  984  9016

(niSim0.1 <- niSim(B = 10000, n = n, del = 1, niMargin = 1, sig = 4,
    pSwitch = 0.2))


FALSE  TRUE
 7397  2603

(niSim1.1 <- niSim(B = 10000, n = n, del = 0, niMargin = 1, sig = 4,
    pSwitch = 0.2))


FALSE  TRUE
 1021  8979
```

In a superiority trial, subjects' switching treatment groups does not cause type I error inflation even though power reduces. In a non-inferiority trial, when 20% of the subjects switch groups, type I error rate was inflated to about 25%; however, the power is remained at 90% because, the centers of distributions coincide under the alternative.

There is no multiplicity penalty for testing superiority and non-inferiority in the same clinical trial. It is because these hypotheses are nested in the sense that if H_{S0} is rejected, H_{N0} is always rejected, and if H_{N0} is not rejected, H_{S0} is not rejected. We can test for both sets of hypotheses with one confidence interval.

4.2 Equivalence

In the statistical hypothesis testing paradigm, no conclusion can be reached by failing to reject H_0, and equivalence can not be concluded by failing to reject a superiority null hypothesis. In an equivalence trial, the following hypotheses are tested:

$$H_{E0} : |\delta| \leq \delta_e$$
$$H_{E1} : |\delta| > \delta_e$$

In the clinical trial literature, non-inferiority trials are often referred to as equivalence trials. There are seldom any therapeutic equivalence trial; most of the equivalence trials are early phase bioequivalence trials in the pharmacokinetics/pharmacodinamics arena. In bioequivalence trials, several pharmacokinetic (PK) parameters, such as, C_{max}, C_{min}, and AUC for a generic drug are compared to those for the marketed drug.

An example of clinical equivalence trial
Pri et. al "Leukotriene antagonists as first-line or add-on asthma-controller therapy". New England Journal of Medicine (2011). In this pragmatic clinical trial, the investigators aimed to show leukotrine-receptor antagonist (LTRA) is equivalent to either an inhaled glucocorticoid for first-line asthma-controller therapy or a long-acting beta agonist (LABA) as add-on therapy in patients already receiving inhaled glucocorticoid therapy. This is a *p*ragmatic trial, as well. As with non-inferiority trials, the intention-to-treat analysis biases towards equivalence, making it challenging to handle dropouts and non-compliances (switching treatment arms).

Chapter 5

Phase II Oncology Clinical Trials

5.1 Introduction

Phase II clinical trial A clinical trial designed to test the feasibility of, and level of activity of, a new agent or procedure. (safety and activity)

Some typical characteristics of a typical phase II clinical trial include:

- It includes a placebo and two to four doses of the test drug.
- When the response is observed quickly, adaptive designs may be beneficial and used because they may
  - improve quality of estimation of the MED (minimum effective dose (lowest dose of a drug that produces the desired clinical effect).
  - increase number of patients allocated to MED.
  - allow for early stopping for futility.

The primary objectives of phase II trials are:

- To determine whether the drug is worthy of further study in phase III trial. Significant treatment effect? / dose-response relationship?
- To gather information to help design phase III trial.
  - Determine dose(s) to carry forward
  - Determine the primary and secondary endpoints
  - Estimate treatment effects for power/sample size analysis
  - Estimate recruitment rate

- Examine feasibility of treatment (logistics of administration and cost)
- Learn about side effects and toxicity

In phase II clinical trials, parallel group designs, crossover designs, and factorial designs are often used.

5.2 Phase II trials in oncology

A phase II clinical trial in oncology generally uses a fixed dose chosen in a phase I trial. The primary objective is to assess therapeutic response to treatment. In the simplest case, a single treatment arm is compared to a historical control. In other cases, a control group and/or multiple doses are included.
The treatment efficacy is often evaluated on surrogate markers for a timely (quick) evaluation of efficacy.

Surrogate outcome An outcome measurement in a clinical trial that substitutes for a definitive clinical outcome or disease status.

- CD4 counts in AIDS study.
- PSA (prostatic specific antigen) in prostate cancer study.
- Blood pressure in cardiovascular disease.
- 3 months survival (binary) for survival.
- Tumor shrinkage for survival.

Tumor response to treatment is evaluated according to Response Evaluation Criteria in Solid Tumors (RECIST)

Complete response (CR) Disappearance of all target lesions.

Partial response (PR) At least a 30% decrease in the sum of the longest diameter (LD) of target lesions, taking as reference the baseline sum LD.

Stable disease (SD) Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD, taking as reference the smallest sum LD since the treatment started.

Progressive disease (PD) At least a 20% increase in the sum of the LD of target lesions, taking as reference the smallest sum LD recorded since the treatment started or the appearance of one or more new lesions.

Generally, objective tumor response is defined as CR or PR in RECIST so that the response variable has a binary endpoint. In the rest of chapter, we will consider a single arm trial with a binary response. The hypothesis of interest is one-sided $H_1 : p > p_0$, and the type I error rate is usually 5 to 10%. The power is usually 80 to 90%.

5.3 Classical (old) two-stage designs

It is crucial that these phase II studies have an opportunity to stop early for toxicity, and that is accomplished by Data Monitoring Committee (DMC), aka, Data and Safety Monitoring Board (DSMB). It is also desired to discard ineffective treatment early, and two-stage designs with a futility stop has been popular.
We will discuss the designs proposed by Gehan (1961), Fleming (1982), and Simon (1989), using the following unified notation:

- stage I sample size $\cdots n_1$.
- stage I data $\cdots X_1 \sim Binomial(n_1, p)$.
- stage I critical value $\cdots r_1$ so that if $X_1 \le r_1$ then terminate the study for futility.
- stage II sample size $\cdots n_2$.
- stage II data $\cdots X_2 \sim Binomial(n_2, p)$.
- total sample size $\cdots n_t = n_1 + n_2$.
- total data $\cdots X_t \equiv X_1 + X_2$.
- stage II critical value $\cdots r_t$ so that if $X_t \le r_t$ then terminate the study for futility, otherwise conclude efficacy.

5.3.1 Gehan's design

It is old (1961) and outdated but may be ok to use in limited situations. The design calls for the first stage with $n_1 = 14$ and $r_1 = 0$, i.e., if no positive response is observed in 14, then stop for futility. The rational is that if true response rate is at least 20%, then $X_1 = 0$ is unlikely. In fact, it is 0.044. The second stage sample size depends on the desired precision for estimating p, and it ranges between 1 and 86. A typical n_2 is 14 so that $n_t = 28$.

5.3.2 Fleming's design

Fleming (1982) proposed a multistage design for phase II clinical trials. One of its key characteristics is stopping early for efficacy.
Example
$H_0 : p = 0.15$, $H_1 : p = 0.30$. (powered at 0.30)
$\alpha = .05$, $\beta = .2$
(Reject H_0 in stage 1 if $X_1 \ge s_1$.

n_1	r_1	s_1	n_t	r_t	α	$1-\beta$	$E_0[N]$	$E_1[N]$
29	4	9	47	10	0.0490	0.8013	36.6	36.9

5.4 Simon's design

In his 1989 paper, Simon introduced two criteria to choose a 2 stage design for single arm and one sided test. The optimal design has the smallest expected sample size under H_0 ($n_1 + E_{p_0}[n_2]$), and the minimax design has the smallest total sample size ($n_1 + n_2$). For $p_0 = 0.15$ and $p_1 = 0.30$,

	n_1	r_1	n_t	r_t	α	$1-\beta$	$E_0[N]$	pet_0	$E_1[N]$	pet_1
optimal	19	3	55	12	0.048	0.801	30.4	0.68	50.2	0.13
minimax	23	3	48	11	0.046	0.804	34.5	0.54	46.7	0.05
single stage	$--$	$--$	48	11	0.048	0.819	48.0	0.00	48.0	0.00

5.4.1 Conditional power

To find a good design (sample sizes and critical values), we need to understand the *conditional* power of a design. The conditional power is the probability of rejecting H_0 (in stage 2) given the stage 1 result, i.e., conditioned on $X_1 = x_1$. Clearly, when $X_1 > r_t$, conditional power is 1, and when $X_1 \leq r_1$ (futility stop), conditional power is 0.

$$\begin{aligned} CP(x_1) &= P[\text{Reject in stage 2}|x_1] = P[x_1 + X_2 > r_t|x_1] \\ &= P[X_2 > r_t - x_1|x_1] \\ &= \sum_{x_2=r_t-x_1+1}^{n2} \binom{n_2}{x_2} p^{x_2}(1-p)^{n_2-x_2} \end{aligned}$$

Conditional power is a function of p, x_1 and n_2 as well as r_t

To obtain the unconditional power, we need to integrate (sum) the conditional power over all possible x_1 values.

$$\begin{aligned} \rho(p) &= \sum_{x_1=0}^{n_1} CP(x_1) P_p[X_1 = x_1] \\ \rho(p) &= \sum_{x_1=r_1+1}^{n_1} CP(x_1) \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1}. \end{aligned}$$

Given α and β a design is chosen so that $\rho(p_0) \leq \alpha$ and $\rho(p_1) \geq 1-\beta$.

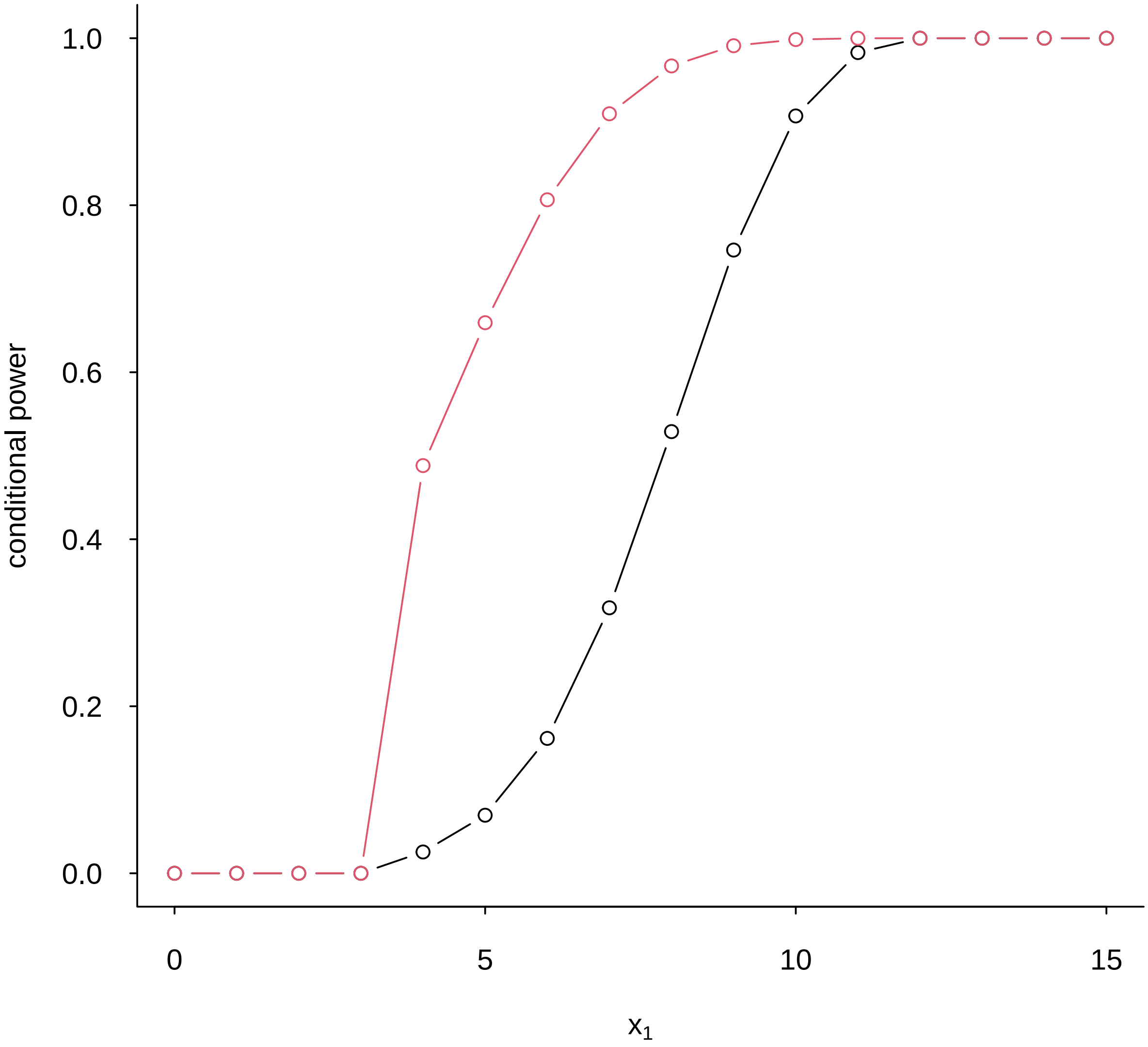
Unlike in a single-stage situation, there may be more than one *good* design. Simon used the *optimal* and *minimax* to choose two reasonable designs among many satisfying the type I error rate and power constraints.
Expected sample size under the null can be written as

$$\begin{aligned}E_{p_0}[n_t] &= n_1 + n_2 P[\text{continue to stage 2}|p_0] \\ &= n_1 + n_2 \times P[X_1 > r_1|p_0] \\ &= n_1 + n_2 \times \sum_{x_1=r_1+1}^{n_1} \binom{n_1}{x_1} p_0^{x_1}(1-p_0)^{n_1-x_1}.\end{aligned}$$

5.4.2 Computing design characteristics

```
simon.d <- function(n1, r1, nt, rt, p0, p1, pl = TRUE, simple = FALSE) {
    # x1 <= r1 stop for futility xt <= rt conclude futility
    R4 <- function(x) {
        round(x, 4)
    }
    R1 <- function(x) {
        round(x, 1)
    }

    x1 <- 0:n1
    pst1.0 <- dbinom(x1, n1, p0)
    pst1.1 <- dbinom(x1, n1, p1)

    cp0 <- 1 - pbinom(rt - x1, nt - n1, p0)
    cp1 <- 1 - pbinom(rt - x1, nt - n1, p1)
    cp0[x1 <= r1] <- 0
    cp1[x1 <= r1] <- 0
    cp0[x1 > rt] <- 1
    cp1[x1 > rt] <- 1

    pow0 <- sum(pst1.0 * cp0)
    pow1 <- sum(pst1.1 * cp1)

    keep <- pmax(pst1.0, pst1.1) > 0.00009
    out1 <- data.frame(x1 = R4(x1), pst1.0 = R4(pst1.0), pst1.1 = R4(pst1.1),
        cp0 = R4(cp0), cp1 = R4(cp1))[keep, ]
    pet0 <- pbinom(r1, n1, p0)
    pet1 <- pbinom(r1, n1, p1)
    en0 <- n1 + (1 - pet0) * (nt - n1)
    en1 <- n1 + (1 - pet1) * (nt - n1)

    out2 <- data.frame(n1, r1, nt, rt, p0 = formatC(p0, digit = 2,
        format = "f"), p1 = formatC(p1, digit = 2, format = "f"))
    out3 <- data.frame(pow0 = R4(pow0), pow1 = R4(pow1), en0 = R1(en0),
        en1 = R1(en1), pet0 = R4(pet0), pet1 = R4(pet1))
    if (pl) {
        plot(out1$x1, out1$x1, type = "n", las = 1, ylim = c(0,
            1), bty = "L", xlab = expression(x[1]), ylab = "conditional power")
        lines(out1$x1, out1$cp0, col = 1, type = "b")
        lines(out1$x1, out1$cp1, col = 2, type = "b")
    }
    out <- list(out1, out2, out3)
    if (simple)
        out <- list(out2, out3)

    out
}

simon.d(n1 = 23, r1 = 3, nt = 48, rt = 11, p0 = 0.15, p1 = 0.3)

[[1]]
   x1 pst1.0 pst1.1    cp0   cp1
1   0 0.0238 0.0003 0.0000 0.000
2   1 0.0966 0.0027 0.0000 0.000
3   2 0.1875 0.0127 0.0000 0.000
4   3 0.2317 0.0382 0.0000 0.000
5   4 0.2044 0.0818 0.0255 0.488
6   5 0.1371 0.1332 0.0695 0.659
7   6 0.0726 0.1712 0.1615 0.806
8   7 0.0311 0.1782 0.3179 0.909
9   8 0.0110 0.1527 0.5289 0.967
10  9 0.0032 0.1091 0.7463 0.991
11 10 0.0008 0.0655 0.9069 0.998
12 11 0.0002 0.0332 0.9828 1.000
13 12 0.0000 0.0142 1.0000 1.000
14 13 0.0000 0.0052 1.0000 1.000
```

```
15 14 0.0000 0.0016 1.0000 1.000
16 15 0.0000 0.0004 1.0000 1.000

[[2]]
  n1 r1 nt rt   p0   p1
1 23  3 48 11 0.15 0.30

[[3]]
    pow0  pow1  en0  en1 pet0   pet1
1 0.0455 0.803 34.5 46.7 0.54 0.0538
```



Given a design, computing operational characteristics such as type I error rate, power, expected sample size is not difficult; however, solving for the optimal, minimax, and other preferable designs is not trivial. Simon's original papers show how to do this.

A very good webpage by Anastasia Ivanova at UNC is at `http://cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign.aspx`.

5.4.3 Something in between

The two criteria, optimal and minimax, give two designs that are extreme, and neither may fit the investigators' needs. For example, for testing $H_0 : p = 0.3$ with $\alpha = 0.05$ and $\beta = 0.10$ at $p_1 = 0.45$, the optimal design and minimax designs are:

	n_1	r_1	n_t	r_t	α	$1-\beta$	$E_0[N]$	pet_0
optimal	40	13	110	40	0.048	0.901	60.8	0.70
balanced	53	18	106	39	0.043	0.903	64.4	0.78
minimax	77	27	88	33	0.050	0.901	78.5	0.86

The optimal design tends to have a small n_1 and the minimax design tends to have a large n_1. Therefore, a simple approach to find a good alternative design is to force $n_1 = n_2$. (balanced design of Ye and Shyr, 2007)

```
simon.d(n1 = 40, r1 = 13, nt = 110, rt = 40, p0 = 0.3, p1 = 0.45,
    pl = FALSE, simple = TRUE)

[[1]]
  n1 r1  nt rt   p0   p1
1 40 13 110 40 0.30 0.45

[[2]]
    pow0  pow1  en0 en1  pet0   pet1
1 0.0482 0.901 60.8 105 0.703 0.0751

simon.d(n1 = 53, r1 = 18, nt = 106, rt = 39, p0 = 0.3, p1 = 0.45,
    pl = FALSE, simple = TRUE)

[[1]]
  n1 r1  nt rt   p0   p1
1 53 18 106 39 0.30 0.45

[[2]]
    pow0  pow1  en0 en1  pet0   pet1
1 0.0431 0.903 64.4 102 0.784 0.0687
```

A more systematic approach is to express the criteria for optimization as

$$q(w) = w \times (n_t) + (1-w) \times E_0[N],$$

where $0 \leq w \leq 1$. $q(0)$ and $q(1)$ correspond to the optimal and minimax designs, respectively. Computation shows that the minimax design is the best design with respect to $q(w)$ for $w \in (0.827, 1]$. In between the optimal and minimax designs, the following "admissible" designs exist that optimize $q(w)$ for certain ranges of w. (Jung, Lee, Kim, George, 2004)

	n_1	r_1	n_t	r_t	α	$1-\beta$	$E_0[N]$	pet_0	w
optimal	40	13	110	40	0.048	0.901	60.8	0.70	$(0, 0.006)$
admissible 1	43	14	104	38	0.050	0.903	60.8	0.70	$(0.006, 0.136)$
admissible 2	48	16	101	37	0.050	0.901	61.3	0.75	$(0.136, 0.182)$
admissible 3	40	12	94	35	0.048	0.902	62.8	0.58	$(0.182, 0.303)$
admissible 4	46	14	91	34	0.049	0.902	64.1	0.60	$(0.304, 0.827)$
minimax	77	27	88	33	0.050	0.901	78.5	0.86	$(0.827, 1)$
single stage	$--$	$--$	90	34	0.045	0.900	90.0	0.00	

5.5 Data analysis following a two-stage design in phase II clinical trials

The primary objective of a (cancer) phase II clinical trial is to make a correct "go/no-go" decision; however, making a good inference for p is advantageous for planning the following phase III trial. We have seen before that when we terminate a study based on an interim summary of the data, a usual statistic that we often compute may be biased. In this section, we will look at the issue of bias in two-stage design in phase II clinical trial in detail. Simon's design will be our focus, but many general discussions can be applied to other designs as well.

5.5.1 p-value

If we ignore the fact that the data were gathered in a two-stage design and compute a p-value as if $X \sim Binomial(n_t, p)$, it is bigger than the true p-value with the following definition/interpretation.

p-value the probability under the null hypothesis that we would observe the data *as or more extreme* than what we have observed

The term "as or more extreme" can be interpreted as "as big or bigger evidence against H_0". In a simple single-stage design, the meaning of this is usually straightforward. We can all agree that $Z = 2.0$ is more extreme (more evidence against H_0) than $Z = 1.9$. However, in two-stage designs, understanding the definition of p-value sometimes gets tricky.

Example:

$H_0 : p = 0.3$, $H_1 : p > 0.3$; $\alpha = 0.05$ and the power is 0.80 at $p = 0.5$. Then the optimal design is: $n_1 = 15$, $r_1 = 5$, $n_t = 46$, $r_t = 18$.

```
simon.d(n1 = 15, r1 = 5, nt = 46, rt = 18, p0 = 0.3, p1 = 0.5,
    pl = FALSE, simple = TRUE)
```

```
[[1]]
  n1 r1 nt rt   p0   p1
1 15  5 46 18 0.30 0.50

[[2]]
    pow0  pow1  en0  en1  pet0  pet1
1 0.0499 0.803 23.6 41.3 0.722 0.151
```

Now suppose we observe $X_1 = 7$ in stage 1 so that we move on to the second stage. And in stage 2, we observe additional 12 positive responses in $n_2 = 31$ patients $(19$ in 46 total) so that H_0 is rejected because $X_t = 19 > r_t$.

If we compute a p-value without taking into account the study design, we might use $X \sim Binomial(46, 0.3)$ and compute

$$p_{\mathsf{c}} = P_0[X \geq 19] = \sum_{i=19}^{46} \binom{46}{i} 0.3^i (1-0.3)^{46-i}$$

where p_{c} is a *conventional* p-value. H_0 is rejected but this p-value is greater than α as shown below:

```
1 - pbinom(18, 46, 0.3)

[1] 0.0681
```

To see this inconsistency clearly, we will rewrite above as

$$\begin{aligned} p_{\mathsf{c}} &= P_0[X \geq 19] \\ &= \sum_{x_1=0}^{15} P_0[X_2 \geq 19 - x_1 | X_1 = x_1] P_0[X_1 = x_1]. \end{aligned}$$

From this expression we see that in computing p_{c}, we include sample paths that can not be realized with this Simon's design, namely, $X_1 = 0,\ X_2 \geq 19;\ X_1 = 1,\ X_2 \geq 18;\ \cdots;\ X_2 = 5,\ X_2 \geq 14.$ A *proper* p-value that takes into account the actual sampling scheme used may be

$$p_{\mathsf{p}} = \sum_{x_1=6}^{15} P_0[X_2 \geq 19 - x_1 | X_1 = x_1] P_0[X_1 = x_1].$$

In general, for Simon-like two-stage designs, p-value should be calculated

$$p_{\mathrm{p}} = \sum_{x_1=r_1+1}^{n_1} P_0[X_2 \geq x_t - x_1 | X_1 = x_1] P_0[X_1 = x_1],$$

if $x_1 > r_1$ (i.e., if there is a second stage).
The following simple R script computes this p-value:

```
pp <- function(n1, r1, nt, rt, x1, xt, p0) {
    x1v <- (r1 + 1):n1
    p.val <- sum((1 - pbinom(xt - x1v - 1, (nt - n1), p0)) *
        dbinom(x1v, n1, p0))
    pc <- 1 - pbinom(xt - 1, nt, p0)
    if (x1 <= r1) {
        p.val <- pc <- 1 - pbinom(x1 - 1, n1, p0)
    }
    c(p.val = p.val, pc = pc)
}

pp(n1 = 15, r1 = 5, nt = 46, rt = 18, x1 = 7, xt = 19, p0 = 0.3)

 p.val     pc
0.0499 0.0681
```

When $x_1 \leq r_1$ so that the trial is terminated in stage 1, we can define

$$p_{\mathrm{p}} = P_0[X_1 \geq x_1].$$

Thus we think that "moving on to the second stage" has more evidence against H_0 than "terminating in the first stage for futility", which makes sense.
The *proper* p-value (p_{p}) has the following characteristics:

- It is always smaller than or equal to p_{c}.
- It is consistent with the hypothesis testing, i.e., $p_{\mathrm{p}} \leq \alpha$ if and only if H_0 is rejected.
- If $X_t = r_t + 1$, then p_{p} is equal to the level of the test (so-called the *actual* type I error rate).
- It does not distinguish different sample paths that lead to the same X_t. That is, evidence against H_0 is identical if x_t is the same regardless of x_1.
  For example, $X_1 = 8$, $X_2 = 12$ and $X_1 = 10$, $X_2 = 10$ yield the same p-values.

When does this (p_{p}) break down?
It breaks down when we allow n_2 to be different for various values of X_1. In some modifications of Simon's design (e.g., Banerjee A, Tsiatis AA. Stat Med 2006), the stage 2 sample size varies with x_1. Then, p_{p} can not be computed because we cannot order the sample paths simply based on X_t.
A bigger concern is that this p_{p} cannot be used when n_2 is changed from that planned. An even bigger concern is if the actual n_2 is different from that planned, how can we re-compute the critical value, r_t, to control type I error rate? The answer is not simple!

5.5.2 Point estimate

Because the results from a phase II clinical trial are often used in planning a phase III clinical trial, a good estimate of p is often of interest.

MLE

In a single stage design, the MLE of p is $\hat{p} = x/n$. For a Simon's design, we can write the likelihood, letting Y_i denote the individual datum from a $Bernoulli(p)$ population, as follows:

$$L(p|\boldsymbol{Y}) = \begin{cases} \Pi_{i=1}^{n_1} p^{y_i}(1-p)^{1-y_i} & \text{if } \sum_i^{n_1} y_i \le r_1 \\ \Pi_{i=1}^{n_t} p^{y_i}(1-p)^{1-y_i} & \text{if } \sum_i^{n_1} y_1 > r_1 \end{cases}$$

$$l(p|\boldsymbol{X}) = \begin{cases} x_1 log(p) + (n_1 - x_1) log(1-p) & \text{if } x_1 \le r_1 \\ x_t log(p) + (n_t - x_t) log(1-p) & \text{if } x_1 > r_1 \end{cases}$$

Therefore, the MLE for π is

$$\hat{p}(x) = \begin{cases} x_1/n_1 & \text{if } x_1 \le r_1 \\ x_t/n_t & \text{if } x_1 > r_1 \end{cases}$$

We have seen before that this $\hat{p}(x)$ has a downward bias, i.e., $E_p[\hat{p}(x)] \le p$. A simple explanation is that when $\hat{p}$ is small at the end of stage 1, we tend to terminate the study, and this downward bias tends to remain; however when $\hat{p}$ is large at the end of stage 1, more data are gathered and the upward bias of stage 1 tends to be corrected.
Example: $p_0 = 0.3$, $p_1 = 0.5$, $\alpha = 0.05$, $\beta = 0.2$. Then the minimax design is ($n_1 = 19$, $r_1 = 6$, $n_t = 39$, $r_t = 16$). Further suppose $X_1 = 8$ and $X_2 = 12$ so that $X_t = 20$.

$$\hat{p} = \frac{20}{39} = 0.513.$$

Whitehead

We can write the bias of the MLE estimator as:

$$B(p) = E_p[\hat{p}(x)] - p.$$

So a good estimator would be

$$\breve{p} = \hat{p} - B(p).$$

However, $B(p)$ is unknown, so we need to estimate it. Let's use the current estimate of p in $B(p)$. That is

$$\hat{p}_w = \hat{p} - B(\hat{p}_w).$$

This is Whitehead's estimator (1986 Biometrika). We can write

$$\hat{p}_w = \hat{p} - E_{\hat{p}_w}[\hat{p}(x)] + \hat{p}_w,$$

which leads to

$$E_{\hat{p}_w}[\hat{p}(x)] = \hat{p}.$$

To find $\hat{p}_w$, we need to numerically solve for $\hat{p}_w$ that satisfies

$$\begin{aligned} E_{\hat{p}_w}[\hat{p}(x)] &= \sum_{x_1=0}^{r_1} \frac{x_1}{n_1} P[X_1 = x_1 | p = \hat{p}_w] + \sum_{x_1=r_1+1}^{n_1} \sum_{x_2=0}^{n_2} \frac{x_1 + x_2}{n_t} P[X_1 = x_1 | p = \hat{p}_w] P[X_2 = x_2 | p = \hat{p}_w] \\ &= \hat{p} \end{aligned}$$

In the current example, $\hat{p}_w = 0.520$.

Koyama

We can write the bias of the MLE estimator as:

$$B(p) = E_p[\hat{p}(x)] - p.$$

So a good estimator would be

$$\breve{p} = \hat{p} - B(p).$$

However, $B(p)$ is unknown, so let's use $B(\hat{p})$, that is

$$\hat{p}_k = \hat{p} - B(\hat{p}).$$

This is simpler and more straightforward than Whitehead's estimator. We can write

$$\begin{aligned}\hat{p}_k &= \hat{p} - E_{\hat{p}}[\hat{p}(x)] + \hat{p} \\ &= 2\hat{p} - E_{\hat{p}}[\hat{p}(x)].\end{aligned}$$

Solving for $\hat{p}_k$ is considerably easier. First compute

$$E_{\hat{p}}[\hat{p}(x)] = \sum_{x_1=0}^{r_1} \frac{x_1}{n_1} P[X_1 = x_1 | p = \hat{p}] + \sum_{x_1=r_1+1}^{n_1} \sum_{x_2=0}^{n_2} \frac{x_1+x_2}{n_t} P[X_1 = x_1 | p = \hat{p}] P[X_2 = x_2 | p = \hat{p}],$$

then subtract it from $2\hat{p}$. In the current example, $\hat{p}_k = 0.521$.

Unbiased estimator

For a general multistage design with early stopping for futility and efficacy, Jung and Kim (2004 Stat Med) found the unbiased estimator of p. They showed that the pair (M,S), where M is the number of stage (when terminated) and S the number of successes, is complete and sufficient for p. And clearly x_1/n_1 is unbiased for p, the uniformly minimum variance unbiased estimator (UMVUE) is found through Rao-Blackwell theorem.

The expression for $\hat{p}_{ub}$ is complex, but for Simon's two-stage design (two-stage with only futility stop), it can be written as

$$\hat{p}_{ub} = \frac{\sum_{x_1=(r_1+1)\vee(x_t-n_2)}^{n_1\wedge x_t} \binom{n_1-1}{x_1-1}\binom{n_2}{x_t-x_1}}{\sum_{x_1=(r_1+1)\vee(x_t-n_2)}^{n_1\wedge x_t} \binom{n_1}{x_1}\binom{n_2}{x_t-x_1}},$$

where $a \wedge b = min(a,b)$ and $a \vee b = max(a,b)$.

For the current example, $max(r_1+1, x_t - n_2) = max(6+1, 20-20) = 7$, and $min(n_1, x_t) = min(19, 20) = 19$, and

$$\begin{aligned}\hat{p}_{ub} &= \frac{\sum_{x_1=7}^{19} \binom{18}{x_1-1}\binom{20}{20-x_1}}{\sum_{x_1=7}^{19} \binom{19}{x_1}\binom{20}{20-x_1}} \\ &= 0.517.\end{aligned}$$

Median estimator

Another simple estimator is the value, p_0^* such that the p-value for testing $H_0 : p = p_0^*$ is 0.5 by the realized sample path. Many adaptive designs for phase II clinical trials were originally motivated as a hypothesis testing procedure, and computing this estimator should be fairly simple in many designs. If the test statistic is continuous, this estimator is known as the median unbiased estimator (Cox and Hinkley 1974). It is unbiased for the true median. The proof uses the fact that the p-value is distributed $Unif(0,1)$ under H_0.
We need to find p_0^* such that

$$\begin{aligned} p_{\mathsf{p}} &= \sum_{x_1=r_1+1}^{n_1} P_{p_0^*}[X_2 \geq x_t - x_1 | X_1 = x_1] P_{p_0^*}[X_1 = x_1] \\ &= \sum_{x_1=7}^{19} P_{p_0^*}[X_2 \geq 20 - x_1 | X_1 = x_1] P_{p_0^*}[X_1 = x_1] \\ &= 0.5. \end{aligned}$$

$p_0^* = 0.500.$

Comparisons

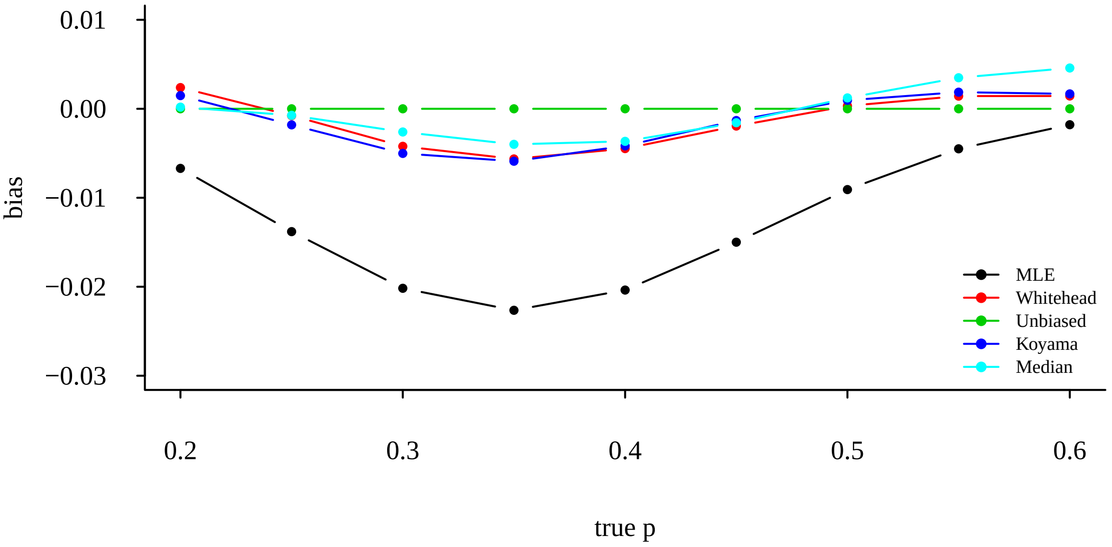
To compare these methods, we compute the bias of each estimator for various true values of p. Use bias and mean squared error = var + bias^2 to compare them. For each estimator, compute $\hat{p}(X)$ for every sample path (defined by X in $[0, n_t]$) and compute

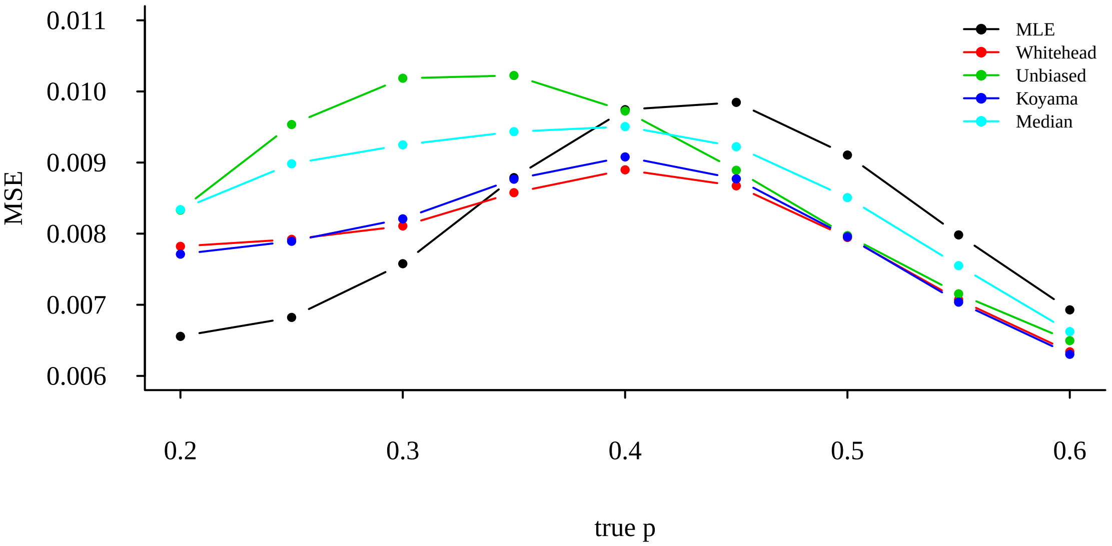
$$E_p[\hat{p}(X)] = \sum_{x=0}^{n_t} \hat{p}(x) P_p[X = x].$$

Mean squared errors can be computed by

$$\begin{aligned} MSE_p[\hat{p}(X)] &= E_p[(\hat{p}(X) - p)^2] \\ &= \sum_{x=0}^{n_t} (\hat{p}(x) - p)^2 P_p[X = x]. \end{aligned}$$

The following two plots show bias and MSE for the current example.


0.01
0.00
−0.01
−0.02
−0.03
bias
0.2
0.3
0.4
0.5
0.6
true p
MLE
Whitehead
Unbiased
Koyama
Median


0.011
0.010
0.009
0.008
0.007
0.006
MSE
0.2
0.3
0.4
0.5
0.6
true p
MLE
Whitehead
Unbiased
Koyama
Median

Chapter 6

Factorial design

6.1 Introduction

Factorial clinical trials (Piantadosi) Experiments that test the effect of more than one treatment using a design that permits an assessment of interactions among the treatments

The simplest example of a factorial design is 2 treatment, 2 treatment groups (2 by 2) designs. With this design, one group receives both treatment, a second group receives neither, and the other two groups receive one of A or B.

	Treatment B		
Treatment A	No	Yes	Total
No	n	n	$2n$
Yes	n	n	$2n$
Total	$2n$	$2n$	$4n$

Four treatment groups and sample sizes in a 2×2 balanced factorial design.
Alternatives to a 2×2 factorial design

- Two separate trials (for A and for B)
- Three arm trial (A, B, neither)

Two major advantages of factorial design (but not simultaneously):

- Allows investigation of interactions (drug synergy).

  Drug synergy occurs when drugs interact in ways that enhance effects or side-effects of those drugs.

- Reduces the cost (sample size) if the drugs do not interact.

Some requirements for conducting a clinical trial with factorial design:

- The side effects of two drugs are not cumulative to make the combination unsafe to administer.
- The treatments need to be administered in combination without changing dosage of the individual drugs.
- It is ethical not to administer the individual drugs. A and B may be given *in addition* to a standard drug so all groups receive some treatment.
- We need to be genuinely interested in studying drug *combination*, otherwise some treatment combinations are unnecessary.

Some terminology

- Factors (how many different treatments are in consideration)
- Levels (2 if yes/no)
- 2^k factorial studies have k factors, each with two levels (presence/absence)
- Full factorial design has no empty cells.

- Unreplicated study has one sample per cell (obviously not very common in clinical studies)
- Fractional factorial designs (some cells are left empty by design)
- Complete block designs / Incomplete block designs
- Latin squares

6.2 Notation and assumptions

	Treatment B	
Treatment A	No	Yes
No	μ	$\mu+\beta$
Yes	$\mu+\alpha$	$\mu+\alpha+\beta+\gamma$

With this formulation, α is the effect of treatment A, β is the effect of treatment B, and γ is the interaction effect. (If the effects of A and B are additive with no interaction, then $\gamma=0$.)

For a continuous outcome and large sample sizes (may be different for each group), we have the following for the observed sample cell means.

$$\begin{aligned}\overline{Y}_0 &\sim Normal(\mu, \sigma^2/n_0)\\ \overline{Y}_A &\sim Normal(\mu+\alpha, \sigma^2/n_A)\\ \overline{Y}_B &\sim Normal(\mu+\beta, \sigma^2/n_B)\\ \overline{Y}_{AB} &\sim Normal(\mu+\alpha+\beta+\gamma, \sigma^2/n_{AB})\end{aligned}$$

We assume $n = n_0 = n_A = n_B = n_{AB}$.

6.3 Test for the interaction effect

In a factorial design, we usually test the presence of interaction effect first.

$$\begin{aligned}H_0 &: \gamma = 0\\ H_1 &: \gamma \neq 0\end{aligned}$$

The observed mean responses are:

	Treatment B	
Treatment A	No	Yes
No	$\overline{Y}_0$	$\overline{Y}_B$
Yes	$\overline{Y}_A$	$\overline{Y}_{AB}$

The interaction effect may be estimated by

$$\hat{\gamma} = (\bar{Y}_{AB} - \bar{Y}_B) - (\bar{Y}_A - \bar{Y}_0),$$

and

$$Var(\hat{\gamma}) = \frac{4\sigma^2}{n}. \qquad \text{(Why is this problematic?)}$$

Thus, if we assume σ^2 is known, then

$$Z = \frac{\hat{\gamma}}{2\sigma/\sqrt{n}}$$

has $Normal(0,1)$ distribution under H_0.

If we have to estimate σ^2 and assume within-group variances are equal, we use a pooled sample variance, $s_p^2 = (s_0^2 + s_A^2 + s_B^2 + s_{AB}^2)/4$.

The test statistic

$$t = \frac{\hat{\gamma}}{2s_p/\sqrt{n}}$$

has a t distribution with $df = 4(n-1)$ under H_0.

6.4 Treatemnt effect

6.4.1 $\gamma \neq 0$

The treatment A effect can be estimated as

$$\hat{\alpha} = \bar{Y}_A - \bar{Y}_0,$$

and its variance is

$$Var(\hat{\alpha}) = \frac{2\sigma^2}{n}.$$

And we have

$$Z = \frac{\hat{\alpha}}{\sqrt{2}\sigma/\sqrt{n}} \sim Normal(0,1)$$

under H_0. We can estimate σ^2 by $s_p^2 = (s_A^2 + s_0^2)/2$. (Note: $2(n-1)s_p^2/\sigma^2 \sim \chi^2_{2(n-1)}$.) Then

$$t = \frac{\hat{\alpha}}{\sqrt{2}s_p/\sqrt{n}}$$

has a t distribution with $df = 2(n-1)$ under H_0. Constructing the test for β is exactly the same.

6.4.2 $\gamma = 0$

If no interaction is present then $\gamma = 0$, and $\tilde{\alpha} = \bar{Y}_{AB} - \bar{Y}_B$ can also be used to estimate α. If we use the average of $\hat{\alpha}$ and $\tilde{\alpha}$ to estimate α, this estimator has a smaller variance.

$$\breve{\alpha} = \frac{\hat{\alpha}+\tilde{\alpha}}{2} = \frac{(\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)}{2}$$

$$Var(\breve{\alpha}) = \frac{1}{4} Var\left(\bar{Y}_A - \bar{Y}_0 + \bar{Y}_{AB} - \bar{Y}_B\right) = \frac{\sigma^2}{n}$$

Similarly to before,

$$Z = \frac{\breve{\alpha}}{\sigma/\sqrt{n}} \sim Normal(0,1), \text{and}$$

$$t = \frac{\breve{\alpha}}{s_p/\sqrt{n}} \sim t_{df}$$

under H_0. (What's the df?) Here we use

$$s_p^2 = (s_0^2 + s_A^2 + s_B^2 + s_{AB}^2)/4,$$

and

$$\frac{4(n-1)}{\sigma^2} s_p^2 \sim \chi^2_{4(n-1)}$$

Constructing the test for β is exactly the same.
In order to have the same efficiency in a two-arm trial (A vs placebo), we would need $2n$ patients in each treatment arm.

$$var\left(\hat{\alpha}_1\right) = \frac{2\sigma^2}{2n} = \frac{\sigma^2}{n}.$$

So if we were to test A and B in two separate experiments we would need $2n$ per arm × 4 arms (A and placebo, B and placebo), totaling $8n$ subjects. Noticing we are repeating the placebo in these hypothetical experiments, we decide to use a 3-arm experiment with A, B, and placebo arms. Then we would require a total of $6n$ subjects for the same precision.

6.5 Examples

The group means are:

	Treatment B	
Treatment A	No	Yes
No	10	40
Yes	30	60

If there is a synergistic effect, then $\eta_{11} > 60$.

	Treatment B	
Treatment A	No	Yes
No	10	40
Yes	30	80

	Treatment B	
Treatment A	No	Yes
No	10	40
Yes	30	120

In the last situation, the treatment effects may be multiplicative.

	Treatment B	
Treatment A	No	Yes
No	$\log(10) = 1$	$\log(40) = 1.60$
Yes	$\log(30) = 1.48$	$\log(120) = 2.08$

Suppose the samples of size 20 yield the following estimates of the cell means.

	Treatment B	
Treatment A	No	Yes
No	9.83	40.05
Yes	28.94	59.76

Assuming no interaction, to estimate the drug A effect we compute either

$$\hat{\alpha}_1 = \bar{Y}_A - \bar{Y}_0 = 28.94 - 9.83 = 19.11$$

or

$$\tilde{\alpha}_1 = \bar{Y}_{AB} - \bar{Y}_B = 59.76 - 40.05 = 19.71$$

or their average $(19.11 + 19.71)/2 = 19.41$.

How bad is it to estimate α_1 this way when there is actually a significant interaction?

$$\begin{aligned} E[(\hat{\alpha}_1 + \tilde{\alpha}_1)/2] &= \frac{1}{2}E[(\bar{Y}_A - \bar{Y}_0) + (\bar{Y}_{AB} - \bar{Y}_B)] \\ &= \frac{1}{2}((\mu + \alpha_1) - \mu + (\mu + \alpha_1 + \beta_1 + \gamma_{11}) - (\mu + \beta_1)) \\ &= \alpha_1 + \frac{\gamma_{11}}{2} \end{aligned}$$

6.5.1 Example: the Physician's Health Study I (1989)

Read all about it on `http://phs.bwh.harvard.edu/`.
The Physician's Health Study was a randomized clinical trial designed to test the following two theories:

- Daily low-dose aspirin use reduces the risk of cardiovascular disease.
- Beta carotene reduces the risk of cancer.

Population hierarchy:

- 261,248 US male MDs aged 40 to 84.
- 112,528 responded to questionnaires.
- 59,285 willing to participate.
- 33,332 willing and eligible MDs enrolled in run-in (18 weeks of active aspirin and beta-carotene placebo).

  Run-in period Eligible patients are monitored for treatment compliance.
- $22,071$ randomized

	Beta-carotene		
Aspirin	Active	Placebo	Total
Active	$5,517$	$5,520$	$11,037$
Placebo	$5,519$	$5,515$	$11,034$
Total	$11,036$	$11,035$	$22,071$

Major findings:

- The trial's DSMB stopped the aspirin arm several years ahead of schedule on 1988/1/25 because it was clear that aspirin had a significant effect on the risk of a first myocardial infarction. (It reduced the risk by 44%.)

  - Did it change the sample sizes for the Beta-carotene components? (Next homework?)

- There were too few strokes or deaths to base sound clinical judgement regarding aspirin and stroke or mortality.

- The beta-carotene arm terminated as scheduled on 1995/12/12 with the conclusion that 13 years of supplementation with beta-carotene produced neither benefit nor harm. Beta-carotene alone was not responsible for the health benefit seen among people who ate plenty of fruits and vegetables.

- Over 300 other findings have emerged from the trial so far.

6.6 Treatment interactions

Factorial designs are the only way to study treatment interactions. Recall the interaction term is estimated by $\hat{\gamma} = (\bar{Y}_{AB} - \bar{Y}_B) - (\bar{Y}_A - \bar{Y}_0)$, and its variance is $Var(\hat{\gamma}) = 4\sigma^2/n$. This variance is 2 times as large as that of A and B main effects, and to have the same precision for an estimate of an interaction effect, the sample size has to be 4 times as large. This means, the two main advantages of the factorial designs (efficiency and interaction objectives) cannot be satisfied simultaneously.
When there is an AB interaction, we cannot use the estimators, $\check{\alpha}$ and $\check{\beta}$, which are only valid with no interaction effect. In fact, we cannot talk about an overall main effect in the presence of an interaction. Instead, we can talk about the effect of A in the absence of B,

$$\alpha = \bar{Y}_A - \bar{Y}_0,$$

or the effect of A in the presence of B

$$\alpha' = \alpha + \gamma = \bar{Y}_{AB} - \bar{Y}_B.$$

Some additional notes

- In the $2 \times 2 \times 2$ design (2^3 design), there are 3 main effects and 4 interactions possible. The number of high order interactions will grow quickly with k, but oftentimes, they are (assumed to be) 0.

- A “quantitative” interaction does not affect the direction of the treatment effect. For example when treatment B is effective either with or without treatment A, but the magnitude of its effectiveness changes.

- With a “qualitative” interaction, the effects of A are reversed with the presence of B. In this case, an overall treatment A effect does not make sense.

- The factorial design can be analyzed with linear models (analysis of variance models).

Limitations of factorial designs

- A higher level design can get complex quickly.
- Test for interaction requires a large sample size (or have a very low power if the study is powered for the main effects).
- Combination therapy may be considered as a treatment in its own right.

Of further interest...

- Partial (fractional) factorial designs have missing cells by design (especially when higher order interactions are assumed to be zero)

Chapter 7

Crossover design

Crossover trials are those in which each patient is given more than one treatment, each at different times in the study, with the intent of estimating differences between them.
In a simple 2×2 design (or AB/BA design), patients are randomized to either “A then B” group or “B then A” group.

	Period	
Group	I	II
AB	Treatment A	Treatment B
BA	Treatment B	Treatment A

2 Treatments / 2 Periods / 2 Sequences

	P_1	P_2	
S_1	A	B	n_1
S_2	B	A	n_2

2 Treatments / 2 Periods / 4 Sequences

	P_1	P_2	
S_1	A	B	n_1
S_2	B	A	n_2
S_3	A	A	n_3
S_4	B	B	n_4

2 Treatments / 4 Periods / 2 Sequences

	P_1	P_2	P_3	P_4	
S_1	A	B	A	B	n_1
S_2	B	A	B	A	n_2

7.1 Some characteristics of crossover design

- All subjects receive more than one treatment (not simultaneously).
- Each subject acts as own control. Therefore, the treatment groups are comparable without relying on randomization.
    - Treatment periods (order of A and B) are often randomly assigned.
    - Baseline characteristics are identical with regard to many patient characteristics, but not with regard to their recent history of exposure to other potentially effective treatments. **carryover effects**
    - The comparability of the treatment groups is not guaranteed by the structure of the trial alone. The investigators need to estimate the carryover effects.
- Crossover designs are not used ...
    - with any condition that treatment could effect considerable change.
    - for acute illness.
- Crossover designs are most suitable for treatments intended for rapid relief of symptoms in chronic diseases, where the long-term condition of the patient remains fairly stable.

Precision
The primary strength of crossover trials is increased efficiency. Suppose the treatment effects are

$$Y_t \sim Normal(\mu_t, \sigma^2),$$
$$Y_c \sim Normal(\mu_c, \sigma^2),$$

and we are interested in $\mu_t - \mu_c$. In a parallel design (with per group sample size of n), we have

$$\hat{\Delta} = \overline{Y}_t - \overline{Y}_c \sim Normal\left(\mu_t - \mu_c, \frac{2\sigma^2}{n}\right).$$

With a TC/CT crossover design with sample size of n,

$$\begin{aligned} var(\hat{\Delta}) &= \frac{2\sigma^2}{n} - 2cov(\overline{Y}_t, \overline{Y}_c) \\ &= \frac{2\sigma^2}{n}\left(1 - \rho_{tc}\right), \end{aligned}$$

where ρ_{tc} is the within-subject correlation of responses on treatments T and C. Therefore, a crossover design is more efficient than a parallel design given $\rho_{tc} > 0$.

Recruitment
Some patients may hesitate to participate in a clinical trial if there is a 50% probability of not receiving any effective treatment. With a crossover design, everyone is guaranteed to receive the test drug. On the other hand, the patients may hesitate to participate in a crossover trial because they will go through more than one treatment, especially when outcomes are assessed with diagnostic procedures such as X-ray, blood drawing, lengthy questionnaires.
Carryover effects
The biggest concern is the possibility that the treatment effect from one period might continue to be present during the following period. A sufficiently long “washout” period between the treatments may prevent significant carryover effects (but how long is sufficiently long?). If there are baseline measurements that represent patient’s disease status, this can be checked against their baseline levels.
If the treatment effects a permanent change or cure in the underlying condition, the treatment given after could look artificially superior.
Dropouts
In a crossover design, the trial duration tends to be longer than a comparable study using independent groups, which may cause more dropouts. Also because every patient take more than one treatment, dropouts due to severe side effects may also increase. The consequences of dropouts are more severe in crossover trial; a simple analysis cannot use only the data from the first period.

7.2 Analysis of 2×2 crossover design

	P_1	P_2
$S_1 = AB$	$\mu_{A1} = \beta_0$	$\mu_{B2} = \beta_0 + \beta_1 + \beta_2$
$S_2 = BA$	$\mu_{B1} = \beta_0 + \beta_1$	$\mu_{A2} = \beta_0 + \beta_2 + \beta_3$

$\beta_0 \cdots$ Treatment A effect
$\beta_1 \cdots$ Increment of treatment effect due to B.
$\beta_2 \cdots$ Carryover effect of treatment A
$\beta_3 \cdots$ Increment carryover effect of treatment B
Treatment B effect is $\beta_0 + \beta_1$, and the carryover effect due to treatment B is $\beta_2 + \beta_3$.
The primary hypotheses to test are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

And how to conduct this test (how to estimate β_1) depends on whether $\beta_3 = 0$. (We’ll see why later.)
Step 0: Assumptions

1. Sample size is n for S_1 and S_2.

2. $\overline{Y} \sim Normal\left(\mu, \sigma^2/n\right)$.

3. $Cor(Y_{A1}, Y_{B2}) = Cor(Y_{B1}, Y_{A2}) = \rho$.

Data:

	P_1	P_2
$S_1 = AB$	$\overline{Y}_{A1} = \hat{\beta}_0$	$\overline{Y}_{B2} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$
$S_2 = BA$	$\overline{Y}_{B1} = \hat{\beta}_0 + \hat{\beta}_1$	$\overline{Y}_{A2} = \hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_3$

Step 1: Test $\beta_3 = 0$
Note that under $H_0 : \beta_3 = 0$, we have $\mu_{A1} + \mu_{B2} = \mu_{B1} + \mu_{A2}$. Thus, we can use

$$Z_1 = \frac{(\overline{Y}_{B1} + \overline{Y}_{A2}) - (\overline{Y}_{A1} + \overline{Y}_{B2})}{\sqrt{Var(\overline{Y}_{B1} + \overline{Y}_{A2}) + Var(\overline{Y}_{A1} + \overline{Y}_{B2})}}$$

to test these hypotheses (assuming σ^2 is known).

Why is this good (convenient)? Because we don't have to worry about the correlations when computing Z_1. Instead, we can compute the within-subject ~~difference~~ sum. Let's say $\omega_{i1} = y_{A1i} + y_{B2i}$ and $\omega_{j2} = y_{B1j} + y_{A2j}$. Then we have

$$Z_1 = \frac{\overline{\omega}_2 - \overline{\omega}_1}{\sqrt{Var(\overline{\omega}_2) + Var(\overline{\omega}_1)}}$$

If we don't assume σ^2 is known, we can use a simple two-sample t-test.

$$t_1 = \frac{\overline{\omega}_2 - \overline{\omega}_1}{\sqrt{2} s_\omega / \sqrt{n}},$$

where $s^2_\omega = (s^2_{\omega_1} + s^2_{\omega_2})/2$. What's the degree of freedom?

Why is this bad?

Step 2a (If $\beta_3 = 0$)

We can estimate β_1 and test if $H_0 : \beta_1 = 0$. Let's say $\delta_{i1} = y_{B2i} - y_{A1i}$ and $\delta_{j2} = y_{A2j} - y_{B1j}$. Let's confirm that the following test statistic can be used to test this hypothesis.

$$t_2 = \frac{\overline{\delta}_1 - \overline{\delta}_2}{\sqrt{2} s_\delta / \sqrt{n}},$$

where $s_\delta^2 = (s_{\delta_1}^2 + s_{\delta_2}^2)/2$. What's the degree of freedom?
We are interested in estimating β_1. Note that $\hat{\beta}_1 = (\overline{\delta}_1 - \overline{\delta}_2)/2$. So a 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{0.975,2(n-1)} se(\hat{\beta}_1)$$
$$\frac{\overline{\delta}_1 - \overline{\delta}_2}{2} \pm t_{0.975,2(n-1)} \sqrt{s_\delta^2/(2n)}$$

Step 2b (If $\beta_3 = 0$)

We can estimate β_2 and test if $H_0 : \beta_2 = 0$. Note that $\hat{\beta}_2 = (\overline{\delta}_1 + \overline{\delta}_2)/2$ and $se(\hat{\beta}_2) = se(\hat{\beta}_1)$. Thus a 95% confidence interval for β_2 is

$$\frac{\overline{\delta}_1 + \overline{\delta}_2}{2} \pm t_{0.975,2(n-1)} \sqrt{s_\delta^2/(2n)}$$

But we are not that interested in estimating β_2.

Step 2c (If $\beta_3 = 0$)

Maybe we want to estimate β_0. We have two estimates of β_0, and we can take the average of them to get

$$\begin{aligned}\hat{\beta}_0 &= \frac{1}{2}\left(\overline{Y}_{A1} + (\overline{Y}_{B1} + \overline{Y}_{A2} - \overline{Y}_{B2})\right)\\ &= \frac{1}{2}\left(\overline{Y}_{B1} + \overline{Y}_{A2} - (\overline{Y}_{B2} - \overline{Y}_{A1})\right)\\ &= \frac{1}{2}(\overline{\varpi}_2 - \overline{\delta}_1)\end{aligned}$$

This means we can test to see if $\beta_0 = 0$ by testing

$$\begin{aligned} H_0 &: \mu_{B1} + \mu_{A2} = \mu_{B2} - \mu_{A1} \\ H_1 &: \mu_{B1} + \mu_{A2} \neq \mu_{B2} - \mu_{A1} \end{aligned}$$

Constructing the relevant t test statistic is not as straightforward as the previous steps because we cannot assume the true variances of ω_2 and δ_1 are equal. We can use

$$t = \frac{\overline{\omega}_2 - \overline{\delta}_1}{\sqrt{\frac{s^2_{\omega_2}}{n} + \frac{s^2_{\delta_1}}{n}}},$$

which follows a t distribution approximately with estimated degrees of freedom given by the Satterthwaite formula.
To estimate β_0 with a confidence interval, compute a confidence interval

$$(\overline{\omega}_2 - \overline{\delta}_1) \pm t_{1-\alpha/2,df} \sqrt{\frac{s^2_{\omega_2}}{n} + \frac{s^2_{\delta_1}}{n}}$$

first (applying the unequal-varince t-test), and divide it by 2.

Step 2d (If $\beta_3 = 0$)

Maybe we want to estimate $\beta_0 + \beta_1$. Again we can use the average of two estimates:

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 &= \frac{1}{2}\left(\overline{Y}_{B1} + (\overline{Y}_{B2} - \overline{Y}_{A2} + \overline{Y}_{A1})\right) \\ &= \frac{1}{2}\left(\overline{Y}_{A1} + \overline{Y}_{B2} - (\overline{Y}_{A2} - \overline{Y}_{B1})\right) \\ &= \frac{1}{2}\left(\overline{\omega}_1 - \overline{\delta}_2\right) \end{aligned}$$

Follow the same process as in step 2c.

Step 3 (If $\beta_3 \neq 0$)

Because the carryover affects S_1 and S_2 differently, we cannot eliminate β_2 as we did before.

$$\begin{aligned} \overline{Y}_{B2} - \overline{Y}_{A1} &= \hat{\beta}_1 + \hat{\beta}_2 \\ \overline{Y}_{B1} - \overline{Y}_{A2} &= \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3 \end{aligned}$$

Taking the within-individual difference is not going to help, so we cannot take an advantage of the correlated endpoint. In this case, we ignore the data from the second period.

$$\begin{aligned}\hat{\beta}_1 &= \overline{Y}_{B1} - \overline{Y}_{A1}\\ Var(\hat{\beta}_1) &= \frac{2\sigma^2}{n}\end{aligned}$$

Treat the study as a two sample test with sample size $= n$ per group.

Moreover, we can estimate the treatment effects with

$$\begin{aligned}\hat{\beta}_0 &= \overline{Y}_{A1}\\ \hat{\beta}_0 + \hat{\beta}_1 &= \overline{Y}_{B1}\end{aligned}$$

7.2.1 Variance of β

Step 1

$$\begin{aligned}\hat{\beta}_3 &= (\overline{Y}_{A2} - \overline{Y}_{A1}) - (\overline{Y}_{B2} - \overline{Y}_{B1})\\ Var(\hat{\beta}_3) &= \frac{4\sigma^2}{n}(1+\rho)\end{aligned}$$

Step 2 (If $\beta_3 = 0$)

$$\begin{aligned}
\hat{\beta}_1 &= \frac{1}{2}\left((\overline{Y}_{B2} - \overline{Y}_{A1}) + (\overline{Y}_{B1} - \overline{Y}_{A2})\right) \\
Var(\hat{\beta}_1) &= \frac{\sigma^2}{n}(1-\rho) \\
\hat{\beta}_2 &= \frac{1}{2}\left((\overline{Y}_{B2} - \overline{Y}_{A1}) - (\overline{Y}_{B1} - \overline{Y}_{A2})\right) \\
Var(\hat{\beta}_2) &= \frac{\sigma^2}{n}(1-\rho) \\
\hat{\beta}_0 &= \frac{1}{2}\left((\overline{Y}_{A1} - \overline{Y}_{B2}) + (\overline{Y}_{B1} + \overline{Y}_{A2})\right) \\
Var(\hat{\beta}_0) &= \frac{\sigma^2}{n} \\
\hat{\beta}_0 + \hat{\beta}_1 &= \frac{1}{2}\left((\overline{Y}_{B1} - \overline{Y}_{A2}) + (\overline{Y}_{A1} + \overline{Y}_{B2})\right) \\
Var(\hat{\beta}_0 + \hat{\beta}_1) &= \frac{\sigma^2}{n}
\end{aligned}$$

Step 3 (If $\beta_3 \neq 0$)

$$\begin{aligned}
\hat{\beta}_1 &= \overline{Y}_{B1} - \overline{Y}_{A1} \\
Var(\hat{\beta}_1) &= \frac{2\sigma^2}{n} \\
\hat{\beta}_0 &= \overline{Y}_{A1} \\
Var(\hat{\beta}_0) &= \frac{\sigma^2}{n} \\
\hat{\beta}_0 + \hat{\beta}_1 &= \overline{Y}_{B1} \\
Var(\hat{\beta}_0 + \hat{\beta}_1) &= \frac{\sigma^2}{n}
\end{aligned}$$

In step 2 and 3, variances of $\hat{\beta}_0$ and $\hat{\beta}_0 + \hat{\beta}_1$ are the same.
$Var\{\hat{\beta}_3\}$ is at least twice as large as $Var\{\hat{\beta}_2\}$ for $\rho \geq 0$. Therefore, any crossover trial designed to detect the differential treatment effects will have lower power for difference of the carryover effects, which is critical to detect the subsequent analysis and interpretation of the trial will be different. With the presence of a clinically important carryover effect difference, a crossover design is no more efficient than an independent-groups trial.
A two-stage procedure may be used: the difference of carryover effects is tested first with a type I

error rate of $10 \sim 20\%$ before moving on to the primary hypothesis testing of the treatment effects. Estimates will be different depending on the conclusion from the first stage.

7.3 Examples

7.3.1 From clinicalTrials.gov

Capecitabine/Erlotinib Followed of Gemcitabine Versus Gemcitabine/Erlotinib Followed of Capecitabine

`http://clinicaltrials.gov/ct2/show/NCT00440167`
This crossover trial is performed in advanced and metastatic pancreatic cancer not previously exposed to chemotherapy. The study compares a standard arm with gemcitabine plus erlotinib to an experimental arm with capecitabine plus erlotinib. It is the first trial of its kind to incorporate second-line treatment into the study design. Patient who fail on first-line therapy are switched to the comparator chemotherapy without erlotinib. The trial therefore not only compares two different regimens of first-line treatment, it also compares two sequential treatment strategies.

Colchicine Randomized Double-Blind Controlled Crossover Study in Behcet's Disease

`http://clinicaltrials.gov/ct2/show/study/NCT00700297`
Method: patients were randomized at the study entry to take either colchicine or placebo. At 4 months, they were crossed over. Those who were taking colchicine went on placebo and those on placebo went on colchicine. Each patient tried therefore, both colchicine and placebo. The primary outcome was the effect of colchicine on the disease activity index, the IBDDAM (16-17). To calculate the overall IBDDAM of the baseline, the IBDDAM of the last 12 months (prior to the study) of each manifestation was calculated and added together. The overall disease activity index was then divided to the number of months (12 months) to have the mean activity index per month. IBDDAM was then measured every 2 months (in the middle and at the end, in each arm of the study). The total IBDDAM of the 4 months was then divided by 4 to have the mean activity index per month. The secondary outcome was to see how the individual symptoms responded to colchicine (IBDDAM of each manifestation).

Statistical analysis: The analysis was done by the intention to treat method. As the difference between IBDDAM before and after treatment had normal distribution Student T test for paired samples were used to evaluate the outcome in the colchicine and the placebo group. As the Levene's test showed the homogeneity of variance, ANOVA (one way) was used to test the effect of treatment (colchicine and placebo) and gender on patients' outcome. The dependent variable was the difference between IBDDAM (before and after the treatment). The independent variables were the treatment, and the gender. SPSS 15 was used for all statistical calculations.

A Placebo-Controlled, Cross-Over Trial of Aripiprazole
`http://clinicaltrials.gov/ct2/show/record/NCT00351936`
Primary endpoint: Evaluate the effects of aripiprazole on weight, Body Mass Index (BMI), and waist/hip circumference.
This study is a ten-week, placebo-controlled, double-blind, cross-over, randomized trial of the novel antipsychotic agent, aripiprazole, added to 20 obese stable olanzapine-treated patients with schizophrenia or schizoaffective disorder. The advantage of the crossover design is that each subject will act as their own control and fewer subjects will be required.

The double-blind, placebo-controlled, crossover study will consist of two random order 4-week treatment arms (aripiprazole 15 mg or placebo) separated by a 2-week adjuvant treatment washout. Following baseline, subjects will be randomized, double-blind, to either aripiprazole or placebo for 4 weeks. After the initial 4 weeks of medication patients will be reassessed, have a 2-week washout period and then crossover to the other treatment for another 4 weeks.

Data management and statistical analysis will be provided by Dr. David Schoenfeld from the Massachusetts General Hospital, Biostatistics Center.

7.4 Examples

7.4.1 Hills and Armitage

Hills M, Armitage P (1979) "The two-period cross-over clinical trial". *Br J Clin Pharmacol.* **8**: 7-20.

- Children with enuresis were treated with a new drug or placebo for 14 days
- The primary data are number of dry nights out of 14.

An estimate of within-subject differences (treatment effects) is $\delta = Y_A - Y_B$. The carryover effects may be estimated by

$$Z_1 = \frac{\overline{\delta}_1 - \overline{\delta}_2}{\sqrt{var(\overline{\delta}_1) + var(\overline{\delta}_2)}},$$

and Z is approximately normally distributed under H_0. Similarly the overall treatment effect can be estimated by

$$Z_2 = \frac{\overline{\delta}_1 + \overline{\delta}_2}{\sqrt{var(\overline{\delta}_1) + var(\overline{\delta}_2)}},$$

and this is approximately normal under H_0.

```
d0 <- c(8, 5, 12, 11, 14, 10, 8, 0, 6, 8, 9, 7, 11, 6, 13, 9,
    3, 5, 8, 8, 6, 0, 8, 9, 0, 0, 4, 8, 8, 14, 13, 12, 2, 4,
    10, 2, 7, 5, 8, 13, 13, 13, 8, 10, 9, 7, 7, 7, 9, 0, 7, 10,
    10, 6, 2, 2, 7, 6)
pat <- rep(1:29, each = 2)
period <- rep(1:2, 29)
placebo.first <- c(2, 5, 8, 10, 12, 14, 15, 17, 20, 23, 26, 29)
group <- rep(1, 29)
group[placebo.first] <- 2
trt <- matrix(1:0, nrow = 29, ncol = 2, byrow = TRUE)
trt[placebo.first, 1] <- 0
trt[placebo.first, 2] <- 1

(d <- data.frame(id = pat, group = rep(group, each = 2), period = period,
    trt = c(t(trt)), dry = d0))

   id group period trt dry
1   1     1      1   1   8
2   1     1      2   0   5
3   2     2      1   0  12
4   2     2      2   1  11
5   3     1      1   1  14
6   3     1      2   0  10
7   4     1      1   1   8
8   4     1      2   0   0
9   5     2      1   0   6
10  5     2      2   1   8
11  6     1      1   1   9
12  6     1      2   0   7
13  7     1      1   1  11
14  7     1      2   0   6
15  8     2      1   0  13
16  8     2      2   1   9
17  9     1      1   1   3
18  9     1      2   0   5
```

```
19 10      2      1   0   8
20 10      2      2   1   8
21 11      1      1   1   6
22 11      1      2   0   0
23 12      2      1   0   8
24 12      2      2   1   9
25 13      1      1   1   0
26 13      1      2   0   0
27 14      2      1   0   4
28 14      2      2   1   8
29 15      2      1   0   8
30 15      2      2   1  14
31 16      1      1   1  13
32 16      1      2   0  12
33 17      2      1   0   2
34 17      2      2   1   4
35 18      1      1   1  10
36 18      1      2   0   2
37 19      1      1   1   7
38 19      1      2   0   5
39 20      2      1   0   8
40 20      2      2   1  13
41 21      1      1   1  13
42 21      1      2   0  13
43 22      1      1   1   8
44 22      1      2   0  10
45 23      2      1   0   9
46 23      2      2   1   7
47 24      1      1   1   7
48 24      1      2   0   7
49 25      1      1   1   9
50 25      1      2   0   0
51 26      2      1   0   7
52 26      2      2   1  10
53 27      1      1   1  10
54 27      1      2   0   6
55 28      1      1   1   2
56 28      1      2   0   2
57 29      2      1   0   7
58 29      2      2   1   6
```

```r
# Group 1: trt -> placebo
g1 <- subset(d, group == 1)
g2 <- subset(d, group == 2)

ms <- function(v) c(mean(v), sd(v)/sqrt(length(v)))

g1.diff <- matrix(g1$dry, ncol = 2, byrow = TRUE)
ms(g1.diff[, 1] - g1.diff[, 2])

[1] 2.824 0.841

g2.diff <- matrix(g2$dry, ncol = 2, byrow = TRUE)
ms(g2.diff[, 2] - g2.diff[, 1])

[1] 1.250 0.863
```

$$
\begin{aligned}
z_1 &= \frac{2.82 - 1.25}{\sqrt{0.8412^2 + 0.8627^2}} = 1.30 \\
z_2 &= \frac{2.82 + 1.25}{\sqrt{0.8412^2 + 0.8627^2}} = 3.38
\end{aligned}
$$

7.5 A two-period crossover design for the comparison of two active treatments and placebo

By GG Koch, IA Amara, BW Brown, T Colton, and DB Gillings (1989).

Consider sequences of treatments TT, TC, and CT.

1. The first period is parallel group design to address direct use in all patients

2. The second period for TT versus TC is a parallel group comparison design to address T versus C for patients who received T during the first period.

3. The second period for TT versus CT enables “delayed start” assessment of T relative to C if dropout during the first period is minimal and non-informative.

4. The second period for CT versus TC is for assessment of T relative C if carryover effects are small.

5. If T − C from 1, 2, 4 are similar (carryover effects of T to T, T to C, C to T are small), then an overall analysis of treatment effect differences have a very high power.
6. More patients are allocated to receive T within each period.

	P_1	P_2
$S_1 = CT$	β_0	$\beta_0 + \beta_1 + \beta_2$
$S_2 = TC$	$\beta_0 + \beta_1$	$\beta_0 + \beta_2 + \beta_3$
$S_3 = TT$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \tau$

$\beta_0 \cdots$ Treatment C effect
$\beta_1 \cdots$ Increment of treatment effect due to T.
$\beta_2 \cdots$ Carryover effect for C
$\beta_3 \cdots$ Increment of carryover effect for T

τ could represent additional treatment effects for longer duration.
Period 1 comparison between T and C is for primary treatment effects, and period 2 comparisons address effects of delayed start (CT vs. TT) and of long-duration effects.
Now consider TT, TC, CT, and CC.

1. This design can estimate all the parameters in the TT, TC, CT case.
2. CC vs. CT enables estimation of treatment effects with run-in period.
3. Relatively unethical to have many patients assigned to receive C.

	P_1	P_2
$S_0 = CC$	β_0	$\beta_0 + \beta_2$
$S_1 = CT$	β_0	$\beta_0 + \beta_1 + \beta_2$
$S_2 = TC$	$\beta_0 + \beta_1$	$\beta_0 + \beta_2 + \beta_3$
$S_3 = TT$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \tau$

$\beta_0 \cdots$ Treatment C effect
$\beta_1 \cdots$ Increment of treatment effect due to T.
$\beta_2 \cdots$ Carryover effect for C
$\beta_3 \cdots$ Carryover effect for T

τ could represent additional treatment effects for longer duration.
Example: Pincus T *et al.* (2004) "Patient preference for placebo, acetaminophen (paracetamol) or celecoxib efficacy studies (PACES): two randomised, double blind, placebo controlled, crossover clinical trials in patients with knee or hip osteoarthritis". *Ann Rheum Dis.* **63**: 931-939.

7.6 Latin squares

When there are k treatments and each patient is to receive all k treatments. Then there are $k!$ possible sequences. Three treatments yield 6 sequences, four treatments yield 24, and five yield 120.
$k = 3$: ABC, ACB, BAC, BCA, CAB, CBA
The idea is to use a reduced number of sequences (reduced sample size) but maintain a good "representation", i.e., every treatment is represented in every period with the same frequency.

	P_1	P_2	P_3
S_1	A	B	C
S_2	B	C	A
S_3	C	A	B

	P_1	P_2	P_3
S_1	A	C	B
S_2	B	A	C
S_3	C	B	A

There are $6!/(3!)(3!) = 20$ ways to choose 3 sequences from 6, but only 2 of those are Latin squares.

7.7 Optimal designs

There is an extensive literature on optimal choice of sequences for measuring treatment effects in the presence of carryover.

- More advanced theory $\cdots$
- Optimality depends on assumptions about carryover effects

Concerns about carryover can be reduced by using designs with more than two periods. (Laska E, Meisner M, Kushner HB. (1983) "Optimal crossover designs in the presence of carryover effects". *Biometrics*. **39**(4): 1087-1091.
Consider treatments A and B in two sequences: $AABB$ and $BBAA$. This design is not uniquely optimal, but it can be used to estimate treatment effects with more efficiency than using data from period 1.

	P_1	P_2	P_3	P_4
AABB	$\mu_{11} = \mu + \pi_1 + \tau_a$	$\mu_{12} = \mu + \pi_2 + \tau_a + \lambda_a$	$\mu_{13} = \mu + \pi_3 + \tau_b + \lambda_a$	$\mu_{14} = \mu + \pi_4 + \tau_b + \lambda_b$
BBAA	$\mu_{21} = \mu + \pi_1 + \tau_b$	$\mu_{22} = \mu + \pi_2 + \tau_b + \lambda_b$	$\mu_{23} = \mu + \pi_3 + \tau_a + \lambda_b$	$\mu_{24} = \mu + \pi_4 + \tau_a + \lambda_a$

Note μ is the overall mean, π is the period effect, τ is the treatment effect, and λ is the carryover effect.
To obtain an unadjusted (for carryover effect) treatment effect ($B - A$), use the following weights.

	P_1	P_2	P_3	P_4
AABB	$-1/4$	$-1/4$	$1/4$	$1/4$
BBAA	$1/4$	$1/4$	$-1/4$	$-1/4$

- Weights sum to 1 for B and -1 for A to form a contrast $B - A$.

- Weights sum to 0 over sequence and period.

$$\begin{aligned}&-\frac{1}{4}\mu_{11}-\frac{1}{4}\mu_{12}+\frac{1}{4}\mu_{13}+\frac{1}{4}\mu_{14}+\frac{1}{4}\mu_{21}+\frac{1}{4}\mu_{22}-\frac{1}{4}\mu_{23}-\frac{1}{4}\mu_{24}\\&=(\tau_b-\tau_a)+(\lambda_b-\lambda_a)/4\end{aligned}$$

When carryover effects are present, we can construct weights so that carryover effects will be eliminated.

	P_1	P_2	P_3	P_4
AABB	$-w_1$	$-w_2$	w_3	w_4
BBAA	w_1	w_2	$-w_3$	$-w_4$

Constraints on w's.

- $w_1+w_2+w_3+w_4=1$
- $w_2-w_3+w_4=0$

$$\begin{aligned}&-w_1\mu_{11}-w_2\mu_{12}+w_3\mu_{13}+w_4\mu_{14}+w_1\mu_{21}+w_2\mu_{22}-w_3\mu_{23}-w_4\mu_{24}\\&=-w_1\tau_a-w_2(\tau_a+\lambda_a)+w_3(\tau_b+\lambda_a)+w_4(\tau_b+\lambda_b)+w_1\tau_b+w_2(\tau_b+\lambda_b)-w_3(\tau_a+\lambda_b)-w_4(\tau_a+\lambda_a)\\&=(w_1+w_2+w_3+w_4)\tau_b-(w_1+w_2+w_3+w_4)\tau_a-(w_2-w_3+w_4)\lambda_a+(w_2-w_3+w_4)\lambda_b\\&=\tau_b-\tau_a\end{aligned}$$

Let σ^2 be the within-patient variance and n be the number of patients per sequence. The variance of the unadjusted estimator is

$$2\left\{\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^2\right\}\frac{\sigma^2}{n}=0.5\frac{\sigma^2}{n}.$$

And for adjusted estimator:

$$2\left\{w_1^2+w_2^2+w_3^2+w_4^2\right\}\frac{\sigma^2}{n}.$$

If we pick $w_1=4/10$, $w_2=2/10$, $w_3=3/10$, $w_4=1/10$, we have

$$2\left\{\left(\frac{4}{10}\right)^2+\left(\frac{2}{10}\right)^2+\left(\frac{3}{10}\right)^2+\left(\frac{1}{10}\right)^2\right\}\frac{\sigma^2}{n}=0.6\frac{\sigma^2}{n}.$$

If we only use data from the first period

$$2\left\{1^2+0^2+0^2+0^2\right\}\frac{\sigma^2}{n}=2\frac{\sigma^2}{n}.$$

The adjusted estimator has slightly higher variance, but it is unbiased with presence of carryover.

William's square

When an even number of treatments are considered in the same number of periods, William's square gives an optimal design. (Williams EJ (1949). "Experimental designs balanced for the estimation of residual effects of treatments". *Australian Journal of Scientific Research. Series A2*. 149-168.) It is a Latin square design in which every treatment precedes every other treatments exactly once.

	P_1	P_2	P_3	P_4
sequence 1	A	B	C	D
sequence 2	B	D	A	C
sequence 3	C	A	D	B
sequence 4	D	C	B	A

Latin square designs are a special type of incomplete block design.

Example:

An experiment was conducted to study the effects of different types of background music on the productivity (Y) of bank tellers. The treatments were defined as five combinations of temp and style of music:

A: slow, instrumental and vocal
B: medium, instrumental and vocal
C: fast, instrumental and vocal
D: medium, instrumental only
E: fast, instrumental only

There are 120 possible sequences of these treatments.

Chapter 8

Group sequential design

8.1 Introduction

Fully sequential method A test of significance is repeated after each observation.

Group sequential method A test of significance is repeated after a group of observations.

Some basic characteristics of a group sequential method

- The response variable needs to be observed immediately.
- Number of stages (or looks) can be 2 to 20.
- Looks are equally spaced. (This is not a critical requirement.)
- At each interim (and final) analysis, compute summary statistic based on the cumulative data.
- A group sequential method is a strategy to stop early as opposed to an “adaptive design”, which is often viewed as a strategy to extend the study if necessary.
- A set of critical values are computed so that the overall α is as specified.
  - Haybittle-Peto (1971)
    This is an *ad hoc* method in which a very conservative critical value (e.g., $Z > 3$) is used at every interim test. At the final analysis, no adjustment is used (i.e., $Z > -1.960$)
    It is highly unlikely to stop early.
  - Pocock (1977)
    A ”repeated test of significance” at a *constant* significance level to analyze accumulating data.
  - O’Brien-Fleming (1979)
    The significance levels increase as the study progress.

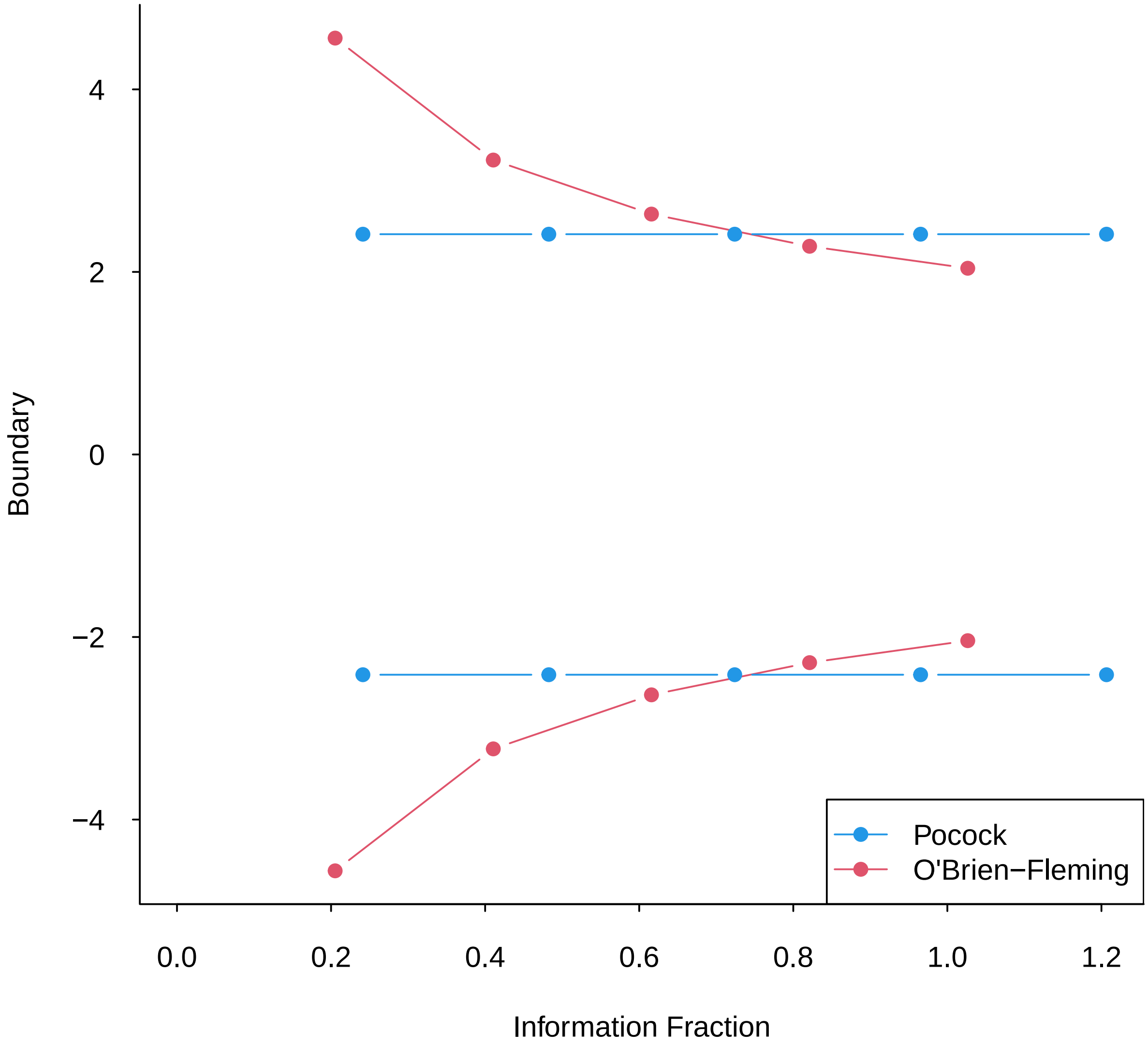
8.2 Example

For testing

$$\mu_t - \mu_c = 0$$
$$\mu_t - \mu_c > 0$$

With $\alpha = 0.025$ and power$= 0.90$ at $\delta_1 = 0.25$ $(\sigma^2 = 1)$, 5-stage group-sequential designs are:

```r
library(gsDesign)

x.of <- gsDesign(k = 5, test.type = 2, alpha = 0.025, beta = 0.1,
    delta0 = 0, delta1 = 0.25, n.fix = 1, sfu = "OF")
x.po <- gsDesign(k = 5, test.type = 2, alpha = 0.025, beta = 0.1,
    delta0 = 0, delta1 = 0.25, n.fix = 1, sfu = "Pocock")
```



Sample size is expressed in terms of ratios to the sample size of the conventional single-stage design.

```
## Pocock
x.po

Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
```

```
Spending computations assume trial stops
if a bound is crossed.

           Sample
            Size
  Analysis Ratio*  Z   Nominal p  Spend
         1  0.241 2.41    0.0079 0.0079
         2  0.483 2.41    0.0079 0.0059
         3  0.724 2.41    0.0079 0.0045
         4  0.965 2.41    0.0079 0.0037
         5  1.207 2.41    0.0079 0.0031
     Total                       0.0250

++ alpha spending:
 Pocock boundary.
* Sample size ratio compared to fixed design with no interim

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
  Theta      1      2      3      4      5 Total  E{N}
   0.00 0.0079 0.0059 0.0045 0.0037 0.0031 0.025 1.177
   3.24 0.2059 0.2603 0.2086 0.1402 0.0851 0.900 0.685

Lower boundary (futility or Type II Error)
          Analysis
  Theta      1      2      3      4      5 Total
   0.00 0.0079 0.0059 0.0045 0.0037 0.0031 0.025
   3.24 0.0000 0.0000 0.0000 0.0000 0.0000 0.000

## O'Brien-Fleming
x.of

Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.

           Sample
```

```
            Size
  Analysis Ratio*  Z   Nominal p  Spend
         1  0.205 4.56    0.0000 0.0000
         2  0.411 3.23    0.0006 0.0006
         3  0.616 2.63    0.0042 0.0038
         4  0.821 2.28    0.0113 0.0083
         5  1.026 2.04    0.0207 0.0122
     Total                       0.0250

++ alpha spending:
 O'Brien-Fleming boundary.
* Sample size ratio compared to fixed design with no interim

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
  Theta      1      2      3      4      5 Total E{N}
   0.00 0.000 0.0006 0.0038 0.0083 0.0122 0.025 1.02
   3.24 0.001 0.1244 0.3421 0.2840 0.1484 0.900 0.75

Lower boundary (futility or Type II Error)
          Analysis
  Theta 1      2      3      4      5 Total
   0.00 0 0.0006 0.0038 0.0083 0.0122 0.025
   3.24 0 0.0000 0.0000 0.0000 0.0000 0.000
```

- In the tables of the critical values, `Nominal p` is simply $P[Z > z]$, where $Z \sim Normal(0,1)$.

- `Spend` is the type I error probability that has been spent by the end of each stage, and it is based on *conditional* probability.
  For example, for the second stage of the Pocock design, it is 0.006. It can be computed as follows:

$$P[Z_2 > 2.413| - 2.413 \le Z_1 \le 2.413].$$

8.3 General applications

Let $k = 1, \cdots, K$ be denote the stages so that we have

$$\begin{aligned} \bar{X}_t^{(k)} - \bar{X}_c^{(k)} &= \frac{1}{n_{tk}} \sum_{i=1}^{n_{tk}} X_{ti} - \frac{1}{n_{ck}} \sum_{i=1}^{n_{ck}} X_{ci} \\ &\sim Normal\left(\mu_t - \mu_c, \frac{\sigma^2}{n_{tk}} + \frac{\sigma^2}{n_{ck}}\right), \end{aligned}$$

where n_{tk} and n_{ck} are the *cumulative* sample sizes for the treatment and control groups. Note that this is not a conditional distribution but a marginal distribution.

Define “information” as $I_k = (\sigma^2/n_{tk} + \sigma^2/n_{ck})^{-1}$. Roughly speaking, information is square of what appears in the denominator of the test statistic, Z. When $n_k = n_{tk} = n_{ck}$, $I_k = (2\sigma^2/n_k)^{-1}$.

The test statistic for stage k is

$$Z_k = \frac{\bar{X}_t^{(k)} - \bar{X}_c^{(k)}}{\sqrt{2\sigma^2/n_k}} = (\bar{X}_t^{(k)} - \bar{X}_c^{(k)})\sqrt{I_k}.$$

The vector, $(Z_1, \cdots, Z_k)$, has a multivariate normal distribution because each Z_k is a linear combination of the independent normal variates X_{ti} and X_{ci}. The marginal distribution of Z_k is

$$Z_k \sim Normal\left((\mu_t - \mu_c)\sqrt{I_k}, 1\right).$$

How about the covariance of Z_{k_1} and Z_{k_2} for $k_1 < k_2$?

$$\begin{aligned}
Cov(Z_{k_1}, Z_{k_2}) &= Cov\left(\{\bar{X}_t^{(k_1)} - \bar{X}_c^{(k_1)}\}\sqrt{I_{k_1}},\, \{\bar{X}_t^{(k_2)} - \bar{X}_c^{(k_2)}\}\sqrt{I_{k_2}}\right) \\
&= Cov\left(\{\bar{X}_t^{(k_1)} - \bar{X}_c^{(k_1)}\},\, \{\bar{X}_t^{(k_2)} - \bar{X}_c^{(k_2)}\}\right)\sqrt{I_{k_1}}\sqrt{I_{k_2}} \\
&= \left[Cov\left(\bar{X}_t^{(k_1)}, \bar{X}_t^{(k_2)}\right) + Cov\left(\bar{X}_c^{(k_1)}, \bar{X}_c^{(k_2)}\right)\right]\sqrt{I_{k_1}}\sqrt{I_{k_2}}
\end{aligned}$$

$$\begin{aligned}
Cov\left(\bar{X}_t^{(k_1)}, \bar{X}_t^{(k_2)}\right) &= Cov\left(\frac{1}{n_{k_1}}\sum_{i=1}^{n_{k_1}} X_i,\, \frac{1}{n_{k_2}}\sum_{i=1}^{n_{k_1}} X_i + \frac{1}{n_{k_2}}\sum X_i\right) \\
&= \frac{1}{n_{k_1}}\frac{1}{n_{k_2}} Var\left(\sum_{i=1}^{n_{k_1}} X_i\right) = \frac{1}{n_{k_2}}\sigma^2 \\
Cov(Z_{k_1}, Z_{k_2}) &= \sigma^2\left(\frac{1}{n_{k_2}} + \frac{1}{n_{k_2}}\right)\sqrt{I_{k_1}}\sqrt{I_{k_2}} \\
&= \sqrt{I_{k_1}/I_{k_2}}.
\end{aligned}$$

Therefore,

- $(Z_1, \cdots, Z_K)$ is multivariate normal.
- $E[Z_k] = (\mu_t - \mu_c)\sqrt{I_k}, \qquad k = 1, \cdots, K$, and
- $Cov(Z_{k_1}, Z_{k_2}) = \sqrt{I_{k_1}/I_{k_2}}, \qquad 1 \le k_1 \le k_2 \le K.$

General decision rule for a group sequential design is
After group $k = 1, \cdots, K-1$
if $|Z_k| \ge c_k$ stop and reject H_0.
otherwise continue to group $k+1$.

After group K
if $|Z_k| \ge c_K$ stop and reject H_0.
otherwise stop for futility.
The test's type I error rate can be expressed as

$$P\left\{|Z_k| \ge c_k \text{ for some } k = 1, \cdots, K\right\}.$$

The critical values, c_k, are chosen so that the above probability is equal to α. And the power of the study at δ_1 is

$$P\left\{\bigcup_{k=1}^{K}\left(|Z_j| < c_j, \text{for } j = 1, \cdots, k-1 \text{ and } |Z_k| \ge c_k\right)\right\}.$$

Evaluation of this probability requires knowing the distribution of $(Z_1, \cdots, Z_K)$. Refer to tables of c_K values or a computer software.

- For a Pocock method, the critical values are constant, so $c_k = C_P(K, \alpha)$. That is, specifying α and K uniquely determines the critical values.

- For the previous example, $C_P(5, 0.025) = 2.413$.

- For an O'Brien-Fleming method, the critical values have the form, $c_k = C_B(K, \alpha)\sqrt{K/k}$

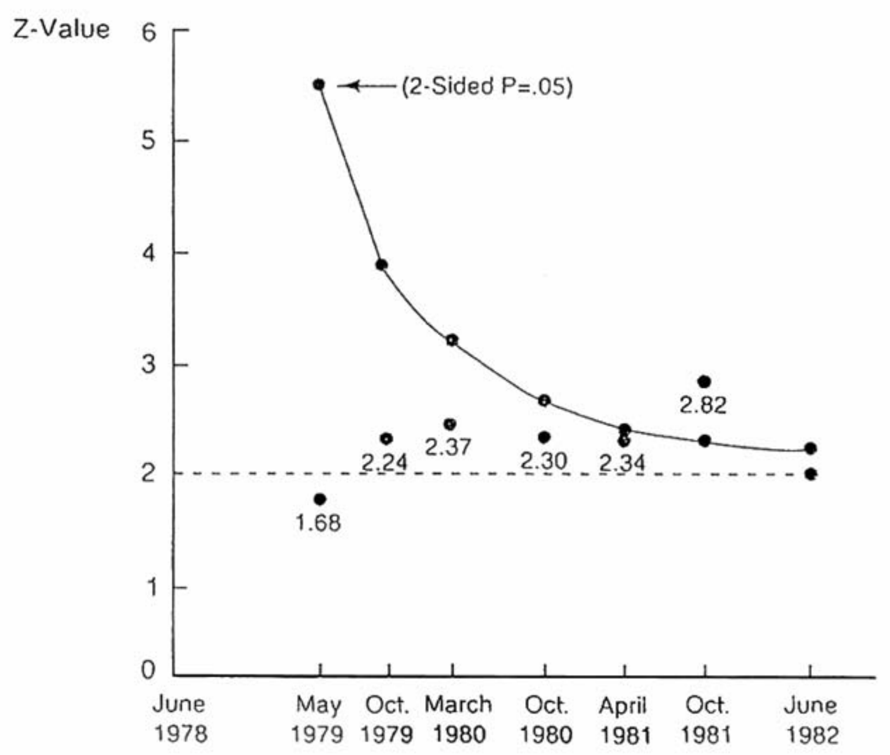
- For the same example, $C_B(5, 0.025) = 2.040$. And the other critical values are: $2.040\sqrt{5/4}$, $2.040\sqrt{5/3}$, and so on.

```
(K.of)

[1] 2.04

K.of * sqrt(5/(5:1))

[1] 2.04 2.28 2.63 3.23 4.56
```

- More generally, if stage sample sizes are different, use I_k, that is, $c_k = C_B(K, \alpha)\sqrt{I_K/I_k}$.

8.3.1 Beta blocker heart attack trial

Seven analyses (including the final one) were planned (corresponding to the timing of the Data Monitoring Committee meetings) using O'Brien-Fleming bounds with two-sided type I error rate of 5%. The primary outcome was survival, and log-rank test was used.



If Pocock boundary had been used, $N = 7$ and $\alpha = 0.05$ give $Z = 2.485$. Therefore, the trial would have been stopped at the same point.

8.3.2 non-Hodgkin's lymphoma

Pocock 1983 *Clinical Trials: A Practical Approach*. A trial was conducted in patients with non-Hodgkin's lymphoma for two drug combinations (cytoxanprednisone -CP- and cytoxan-vincristine-prednisone -CVP-). The primary endpoint was tumor shrinkage (Yes/No).
Statistical analyses were planned after approximately 25 patients. With 5 looks and one-sided $\alpha = 0.05$. The Pocock procedure requires a significance level of 0.017 at each analysis. χ^2 tests without the continuity correction were performed at each of the 5 scheduled analyses.

```
gsDesign(k = 5, test.type = 1, alpha = 0.05, n.fix = 1, sfu = "Pocock")

One-sided group sequential design with
90 % power and 5 % Type I Error.
           Sample
            Size
  Analysis Ratio*  Z   Nominal p  Spend
         1  0.246 2.12    0.0169 0.0169
         2  0.491 2.12    0.0169 0.0117
         3  0.737 2.12    0.0169 0.0087
         4  0.982 2.12    0.0169 0.0069
         5  1.228 2.12    0.0169 0.0057
     Total                       0.0500

++ alpha spending:
 Pocock boundary.
* Sample size ratio compared to fixed design with no interim

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
  Theta      1      2      3      4      5 Total  E{N}
   0.00 0.0169 0.0117 0.0087 0.0069 0.0057  0.05 1.197
   2.93 0.2510 0.2574 0.1900 0.1249 0.0767  0.90 0.668
```

	Tumor shrinkage		
	CP	CVP	p-value
Analysis 1	$3/14$	$5/11$	0.40
Analysis 2	$11/27$	$13/24$	0.50
Analysis 3	$18/40$	$17/36$	1.00
Analysis 4	$18/54$	$24/48$	0.13
Analysis 5	$23/67$	$31/59$	0.06

The CVP appeared better than the CP, but difference was not statistically significant. Further analyses of secondary endpoints convinced the researchers that the CVP was better than the CP.

8.4 Alpha-spending

“Classical” group sequential designs have equal information (sample size) at every stage, but we may want to be a little more flexible. And when I_k is not a constant we might want to change α spent accordingly.

Decompose the rejection region.

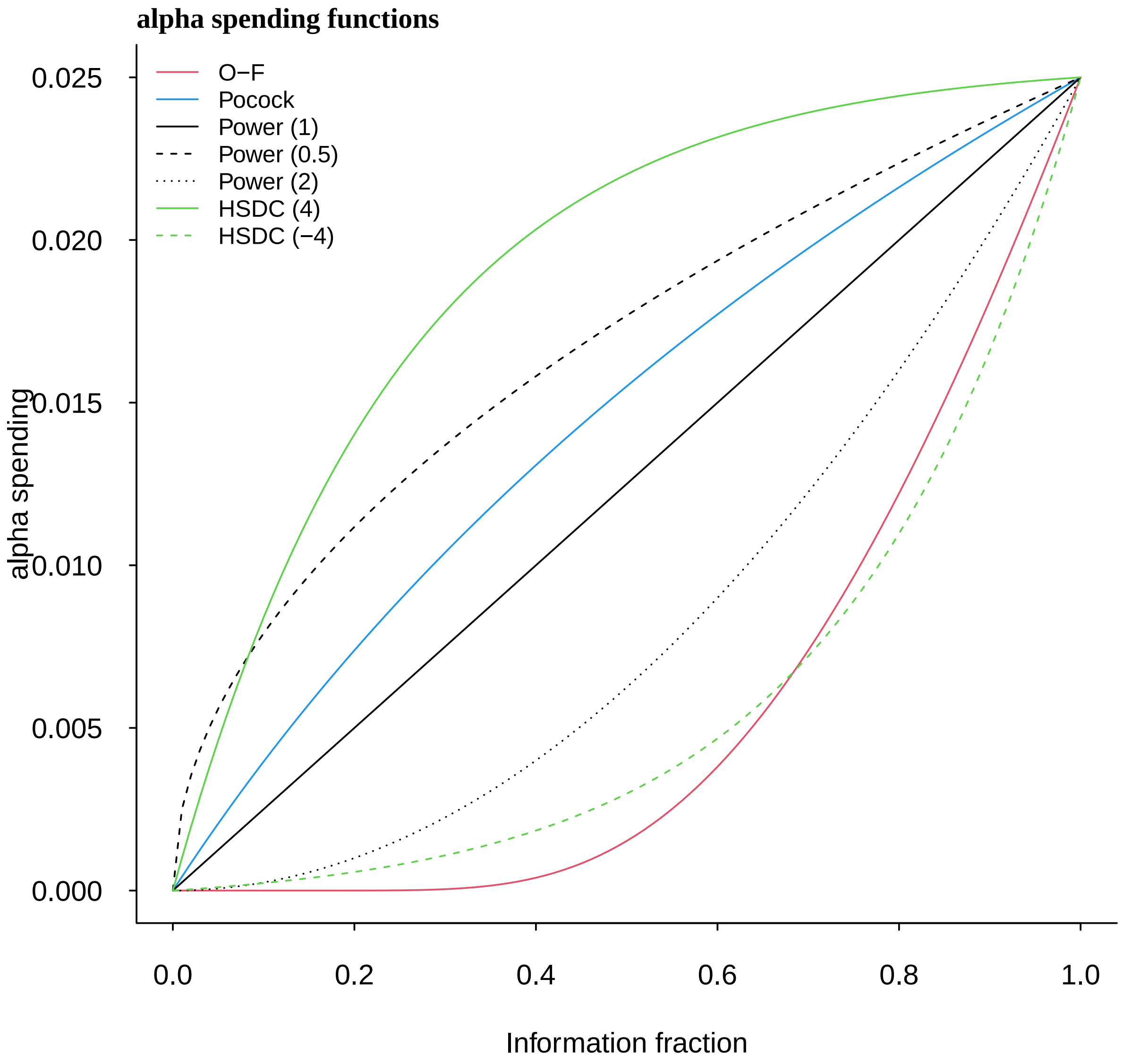
$$\begin{aligned}
R &= P\{|Z_k| \geq c_k \text{ for some } k = 1, \cdots, K\} \\
&= P\{(|Z_1| \geq c_1) \text{ or } (|Z_1| < c_1 \text{ and } |Z_2| \geq c_2) \text{ or } \cdots\} \\
&= P\{|Z_1| \geq c_1\} + P\{|Z_1| < c_1 \text{ and } |Z_2| \geq c_2\} + P\{|Z_1| < c_1 \text{ and } |Z_2| < c_2 \text{ and } |Z_3| \geq c_3\} + \cdots \\
&= \alpha(I_1) + (\alpha(I_2) - \alpha(I_1)) + (\alpha(I_3) - \alpha(I_2) - \alpha(I_1)) + \cdots
\end{aligned}$$

The biggest advantage of alpha-spending approach is its flexibility; neither the number nor timing of the interim analyses need to be specified in advance. The monitoring plan can be changed during the trial and still type I error rate is preserved. The power depends relatively little on the number and timing of the interim looks[1].

Alpha-spending functions

O'Brien-Fleming	$\alpha(t) = 2\left[1 - \Phi\left(z_{\alpha/2}/\sqrt{t}\right)\right]$
Pocock	$\alpha(t) = \alpha \log(1 + (e-1)t)$
Kim-DeMets (Power)	$\alpha(t, \theta) = \alpha t^{\theta}$ (for $\theta > 0$)
Hwang-Shih-DeCani	$\alpha(t, \phi) = \alpha \frac{1-e^{-\phi t}}{1-e^{-\phi}}$ (for $\phi \neq 0$)

[1] "Fundamentals of clinical trials (4th ed)" by Friedman LM, Furberg CD, DeMets DL



8.5 One-sided test

If “stop for futility” is not an option, the same boundary can be used. If a futility stop is an option, then

After group $k = 1, \cdots, K-1$
if $Z_k \geq b_k$ stop and reject H_0.
if $Z_k \leq a_k$ stop for futility (accept H_0).
After group K
if $Z_k \geq b_K$ stop and reject H_0.
if $Z_k < a_K$ stop for futility.
Note that $a_K = b_K$ ensures that the test terminates at analysis K.

8.6 Repeated confidence intervals

If we compute unadjusted confidence intervals $\bar{X}_{\text{so far}} \pm 1.96\sigma/\sqrt{n_{\text{so far}}}$ at the end of each stage, we get low coverage probabilities. Armitage, McPherson, Rowe ("Repeated significance tests on accumulating data". *JRSS-A* 1969) computed the actual coverage probabilities (Table 2).

Number of looks	Overall probability that all intervals contain θ
1	0.95
2	0.92
3	0.89
4	0.87
5	0.86
10	0.81
20	0.75
50	0.68
∞	0

The idea of repeated confidence intervals (RCIs) is to use an adjusted value, $c_k(\alpha, K)$, instead of 1.96 so that the overall coverage probability is $1 - \alpha/2$. The value of $c_k(\alpha, K)$ is the critical value (border) for each stage and depends on α and K if Pocock boundary is used, and additionally k if O'Brien-Fleming boundary is used.

Example: Suppose we use a 6-stage group sequential design of O'Brien-Fleming type with a two-sided $\alpha = 5\%$. The critical values are:

```
gsDesign(k = 6, test.type = 2, alpha = 0.025, sfu = "OF")

Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.

           Sample
            Size
```

```
  Analysis Ratio*  Z   Nominal p  Spend
         1  0.172 5.03    0.0000 0.0000
         2  0.343 3.56    0.0002 0.0002
         3  0.515 2.90    0.0018 0.0017
         4  0.686 2.51    0.0060 0.0047
         5  0.858 2.25    0.0123 0.0079
         6  1.030 2.05    0.0200 0.0105
     Total                       0.0250

++ alpha spending:
 O'Brien-Fleming boundary.
* Sample size ratio compared to fixed design with no interim

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
  Theta      1      2      3      4      5      6 Total  E{N}
   0.00 0.0000 0.0002 0.0017 0.0047 0.0079 0.0105 0.025 1.022
   3.24 0.0001 0.0487 0.2350 0.2915 0.2088 0.1159 0.900 0.739

Lower boundary (futility or Type II Error)
          Analysis
  Theta 1      2      3      4      5      6 Total
   0.00 0 0.0002 0.0017 0.0047 0.0079 0.0105 0.025
   3.24 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.000
```

The critical values are

```
[1] 5.03 3.56 2.90 2.51 2.25 2.05
```

First, let's confirm that the critical values have the form $c_k = C_{OB}(K,\alpha)\sqrt{I_K/I_k}$. The final critical value $C_{OB}(6,\alpha) = 2.053$, and assuming the looks are equi-distant (same group sample size), we have:

$$
\begin{array}{ll}
c_1 = 2.053\sqrt{6/1} = 5.028 & c_2 = 2.053\sqrt{6/2} = 3.556 \\
c_3 = 2.053\sqrt{6/3} = 2.903 & c_4 = 2.053\sqrt{6/4} = 2.514 \\
c_5 = 2.053\sqrt{6/5} = 2.249 & c_6 = 2.053\sqrt{6/6} = 2.053
\end{array}
$$

Then after stage 1, we would use 2.053 in place of the regular 1.96 when computing a 95% confidence interval. In general after stage k $(k = 1, \cdots, 6)$,

$$(\bar{x}_{kt} - \bar{x}_{kc}) \pm c_k \frac{\sqrt{2\sigma^2}}{\sqrt{mk}},$$

where m is per-group sample size for each stage.
This method (RCI) is consistent with the corresponding hypothesis testing. Only when is H_0 rejected in stage k, the confidence interval for that stage will exclude the null value. Thus, we can use the idea of “inverting hypothesis test” to get the same confidence interval. (more later)

8.7 p-values

Recall how we construct a proper p-value for a Simon’s two-stage design in phase II methodology. We needed to define “more or as extreme as the observed data”. To be able to do this, we need to have an ordering of all the sample paths. In a simple single-stage design, the ordering is usually based on z-values (or absolute value of z-values if two-sided test), i.e., the bigger the observed z, the stronger the evidence against H_0. Then a one-sided p-value is computed by

$$p = P_0[Z \geq z].$$

With a group sequential design, or more generally, with a multi-stage design with pre-specified group-wise sample sizes, the following orderings have been proposed. Notation: $(k', z') \succ (k, z)$ to denote (k', z') is above (k, z).

- Stage-wise ordering.
  $(k', z') \succ (k, z)$ if any of the following is true:

  1. $k' = k$ and $z' \geq z$.
  2. $k' < k$ and $z' \geq b_{k'}$ (upper critical value).
  3. $k' > k$ and $z \leq a_k$ (lower critical value).

- MLE ordering.
  $(k', z') \succ (k, z)$ if $z'/\sqrt{I_{k'}} > z/\sqrt{I_k}$. Originally proposed in connection with a test for a binomial proportion The bigger value of the MLE gets a higher order. Sometimes called “sample mean ordering” because this is equivalent to ordering based on the sample mean (one-sample) or the difference of sample means (two-samples).

- Likelihood ratio ordering.
  $(k', z') \succ (k, z)$ if $z' > z$. (Stages do not matter.)

- Score test ordering.
  $(k',z') \succ (k,z)$ if $z\sqrt{I_{k'}} > z\sqrt{I_k}$.

Whichever ordering is used, we can compute a one-sided p-value is

$$P_0[(T,Z_T) \succ (k^*,z^*)]$$

For example, if we use a stage-wise ordering and test terminates in the $K-1$ stage with $Z_{K-1} > b_{K-1}$ (reject H_0).

$$p = \int_{b_1}^{\infty} g_1(z;0)dz + \cdots + \int_{z^*}^{\infty} g_{K-1}(z;0)dz.$$

In the above expression, $g_k(z;\theta)$ is a density function of z in stage k. Conceptually, the density function of z in k stage depends on all the data in the previous stages, $1 \cdots k-1$, requiring multivariate integration.
Armitage, McPherson, Rowe (1969) derived a recursive formula so that the computation is much simplified, requiring only a succession of univariate integrations. For $k = 2, \cdots, K$,

$$g_k(z;\theta) = \int_{C_{k_1}} g_{k-1}(\mu;\theta)\frac{\sqrt{I_k}}{\sqrt{\Delta_K}}\phi\left(\frac{z\sqrt{I_k} - \mu\sqrt{I_{k-1}} - \Delta_k\theta}{\sqrt{\Delta_k}}\right)d\mu,$$

where C_{k_1} is the continuation region of the stage k_1, and Δ_k is the increment information, $I_k - I_{k-1}$.
If stage-wise ordering is used, it automatically ensures that item the p-value is less than the significance level α if and only if H_0 is rejected.
Once we define the ordering to use with the group sequential test then we can compute a p-value for testing $H_0 : \theta = 0$ by “inverting hypothesis test”. A $(1-\alpha/2)$ confidence interval is a collection of θ_0' such that $H_0' : \theta = \theta_0'$ would be accepted with the observed sample path. (More details with general adaptive designs.)

Chapter 9

General Adaptive Designs

Adaptive clinical trials

- are prospectively planned.
  - All possible adaptations are defined in protocol.
- require much more forethought than a single stage design.
- require same level of evidence as usual.
- are accepted by regulators (e.g., FDA)
  - FDA guidance available
    Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics
- are nicer with the Bayesian method.

9.1 Introduction

Much of discussion in the literature for flexible designs in phase III clinical trial methodologies revolves around 2 stage designs. Practically speaking, implementing flexible Frequentist clinical trials beyond two stages is difficult, and perhaps these multi-stage flexible designs add only little to the designs with just two stages. Moreover, phase III clinical trials are for confirmatory purposes, and adaptively changing the design more than once in the middle of a confirmatory trial is not seen favorable by the regulatory figure.
So we will only consider two-stage adaptive designs. Two-stage group sequential designs are examples of such designs.

9.2 Background

We will look at unblinded two stage designs in which all the information from stage 1 is available. Design of the second stage (sample size and critical value) may be specified as functions of stage 1 data. If both sample size and critical value are constants in stage 1 data, then it reduces to a two-stage group sequential design.
Adaptive designs can be categorized into the following two types:

- **prespecified designs**
  Design of the second stage (e.g., sample size and critical value) is specified before the first stage. There is nothing to decide at the end of stage 1. The design of the second stage is defined flexibility as functions of stage 1 data. Group sequential designs fall into this category.
- **unspecified designs**
  Design of the second stage is not specified in advance and determined after stage 1 data are observed. These designs are generally not accepted by the regulatory body.

Characteristics of these types of designs:

prespecified designs

- Type I error can be controlled.
- Type II error can be controlled; the power can be specified.
- Design characteristics (e.g., expected and maximum sample sizes) can be computed prior to the initiation of the study.

unspecified designs

- Type I error can be controlled.
- These designs give much flexibility to handle unexpected situations (e.g., variance much bigger than anticipated).
- Probably not accepted.

What they say about adaptive designs...

PhRMA (2006) "A clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial."
"... changes are made by design, and not on an ad hoc basis; therefore, adaptation is a design feature aimed to enhance the trial, not a remedy for inadequate planning."

EMA (2006) "A study design is called 'adaptive' if statistical methodology allows the modification of a design element (e.g. sample-size, randomisation ratio, number of treatment arms) at an interim analysis with full control of type I error rate."
"adaptive designs should not be seen as a means to alleviate the burden of rigorous planning of clinical trials."

FDA (2010) "... adaptive design clinical study is defined as a study that includes a prospectively planned opportunity for modification of one or more aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study."

9.3 Conditional power

- Power is the probability of rejecting H_0.
- Type I error rate is the power computed under H_0.
- Conditional power is the probability of rejecting H_0 at the end of the second stage conditioned on the first stage data.

Example: Suppose in stage 1, we will reject H_0 if $Z_1 > 2.3$ otherwise continue to stage 2. At the end of stage 2, we will reject H_0 if $Z_2 > 2.1$, where Z_1 is from the stage 1 data, and Z_2 is from the combined data from both stages. Then

- Stage 1 power is P[Reject H_0 at the end of stage 1] $= P[Z_1 > 2.3]$.
- Stage 2 conditional power is P[Reject H_0 at the end of stage 2 | Stage 1 data] $= P[Z_2 > 2.1 | Z_1 = z_1]$.
  - If there is no stage 2, conditional power does not exist, but...
  - we can let conditional power $= 1$ if H_0 is rejected in stage 1 and $= 0$ if the trial stops for futility in stage 1.
- Conditional type I error rate is the probability of rejecting H_0 in stage 2 under H_0 conditioned on the stage 1 data.

Suppose

$$Z_2 = \frac{Z_1 + Z_{2-1}}{\sqrt{2}}$$

where Z_2 are from the cumulative data, and Z_{2-1} are from the stage 2 data only. Under H_0 $Z_2 \sim Normal(0,1)$. Conditioned on $Z_1 = z_1$, we have

$$Z_2 = \frac{z_1 + Z_{2-1}}{\sqrt{2}}.$$

Therefore, $Z_2 > c$ is equivalent to

$$Z_{2-1} > \sqrt{2}c - z_1.$$

```
# Stage 1 type I error rate
1 - pnorm(2.3)

[1] 0.0107

# Conditional type I error rate:
condTypeI <- function(z1) {
    # Reject H0 if Z_2 > 2.1.
    1 - pnorm(2.1 * sqrt(2) - z1)
}
# Conditional type I error rate if Z_1 = 0.
condTypeI(0)

[1] 0.00149

# Conditional type I error rate if Z_1 = 1.
condTypeI(1)

[1] 0.0244

# Conditional type I error rate if Z_1 = 2.
condTypeI(2)

[1] 0.166
```

When planning a two-stage adaptive design, we need to make sure that the (unconditional) type I error rate is α, which is usually 2.5% To compute the unconditional type I error rate, or more generally, the unconditional power, we need to integrate the unconditional type I error rate (power) with respect to the distribution of Z_1.

$$\text{P[Unconditional power]} = \int \text{P[Conditional power} \mid Z_1 = z_1] f(z_1) dz_1$$

```
Integrand <- function(z1) {
    condTypeI(z1) * dnorm(z1)
}

## Stage 2 (unconditional) type I error rate.
integrate(Integrand, lower = -Inf, upper = 2.3)$value

[1] 0.0138
```

Sometimes, the term, “predictive power” is confused with the conditional power. The predictive power is a Bayesian concept, which is a weighted average of the power computed for different values of θ with the weight is defined by the distribution of θ.
In the above example, we used a simple test statistic

$$Z_2 = \frac{Z_1 + Z_{2-1}}{\sqrt{2}}$$

to combine the data from both stages; however, this is rarely used in practice (despite its simplicity) because it puts the same weight on the stage 1 and 2 z-values regardless of the sample sizes.

9.4 Two-stage design -without all the math-

We will consider a simple situation, where we aim to test $H_0 : \delta = 0$, where $\delta = \mu_t - \mu_c$. Random samples are taken from the treatment and control populations

$$X_t \sim Normal(\mu_t, \sigma_t^2) \qquad X_c \sim Normal(\mu_c, \sigma_c^2).$$

Assume the true variances are equal and known: $\sigma_t^2 = \sigma_c^2 = \sigma^2$. Also assume the stage 1 sample sizes are equal in the control and treatment groups: $n_{1t} = n_{1c} = n_1$. Then

$$\bar{X}_{1t} \sim Normal(\mu_t, \sigma^2/n_1) \qquad \bar{X}_{1c} \sim Normal(\mu_c, \sigma^2/n_1)$$

and

$$Z_1 = \frac{\sqrt{n_1}(\bar{X}_{1t} - \bar{X}_{1c})}{\sqrt{2}\sigma}.$$

The distribution of Z_1 under the null is $Norlam(0,1)$.
In stage 1, we observe Z_1 and use the following decision rule:

- If $Z_1 < k_1$, stop for futility.
- If $Z_1 > k_2$, stop and reject H_0.
- If $k_1 < Z_1 < k_2$ then continue to stage 2.

In stage 2, we take a sample of size n_2 from each arm, and define

$$Z_{2-1} = \frac{\sqrt{n_2}(\bar{X}_{2t} - \bar{X}_{2c})}{\sqrt{2}\sigma}.$$

Conditioned on $Z_1 = z_1 \in (k_1, k_2)$,

$$Z_{2-1} \sim Normal(\sqrt{n_2}\delta/(\sqrt{2}\sigma), 1),$$

which is $Normal(0,1)$ under H_0. Note that n_2 may be a function of the stage 1 data.
Instead of combining the stage 1 and 2 data with $(Z_1 + Z_{2-1})/\sqrt{2}$, we will use the following test statistic so that each X is weighted equally.

$$Z_2 = \frac{\sqrt{n_1}}{\sqrt{n_1+n_2}} Z_1 + \frac{\sqrt{n_2}}{\sqrt{n_1+n_2}} Z_{2-1}.$$

We can write down the conditional distribution of this test statistic given $Z_1 = z_1$ (omitted).
The important ideas here are:

- n_2 is not a constant, but it is a random variable. $n_2(z_1)$.
- When conditioned on $Z_1 = z_1$, Z_1 is no longer a random variable but a constant.
- There are many decent ways to combine the stage 1 and 2 data.
- The critical value to use at the end of stage 2 can be for Z_{2-1}, and it may be a function of Z_1.

In the second stage, a sample of size $n_2(z_1)$ is taken, and the decision rule at the end of stage 2 is:

- If $Z_{2-1} \leq c(z_1)$, stop and conclude futility.
- If $Z_{2-1} > c(z_1)$, stop and conclude efficacy.

9.5 Conditional power functions

Like with the group sequential designs, we need to have the conditional type I error rate that satisfies

$$\alpha = \alpha_1 + \int_{k_1}^{k_2} P_0[\text{Reject } H_0 | Z_1 = z_1] g(z_1) dz_1.$$

$P_0[\text{Reject } H_0 | Z_1 = z_1]$ is the conditional power, and we can use any form of conditional power as long as the above criterion is satisfied. Here let $CP(Z_1, \delta)$ to denote the conditional power function, so that

$$\begin{aligned} CP(Z_1, \delta = \delta_0) &\text{ is the conditional type I error rate.} \\ CP(Z_1, \delta = \delta_1) &\text{ is the conditional power at the original alternative.} \\ CP(Z_1, \delta = \hat{\delta}) &\text{ is the conditional power at the stage 1 trend.} \end{aligned}$$

9.6 Example

Design a two-stage procedure to test $H_0 : \mu_t - \mu_c = 0$ and $H_1 : \mu_t - \mu_c > 0$. Assume σ is known to be 4. We want one sided α to be 0.025 and power to be 0.90 at $\mu_t - \mu_c = 1$. For a single stage design, the sample size is

$$N = 4^2(2)\left(z_{0.025} + z_{0.10}\right)^2$$

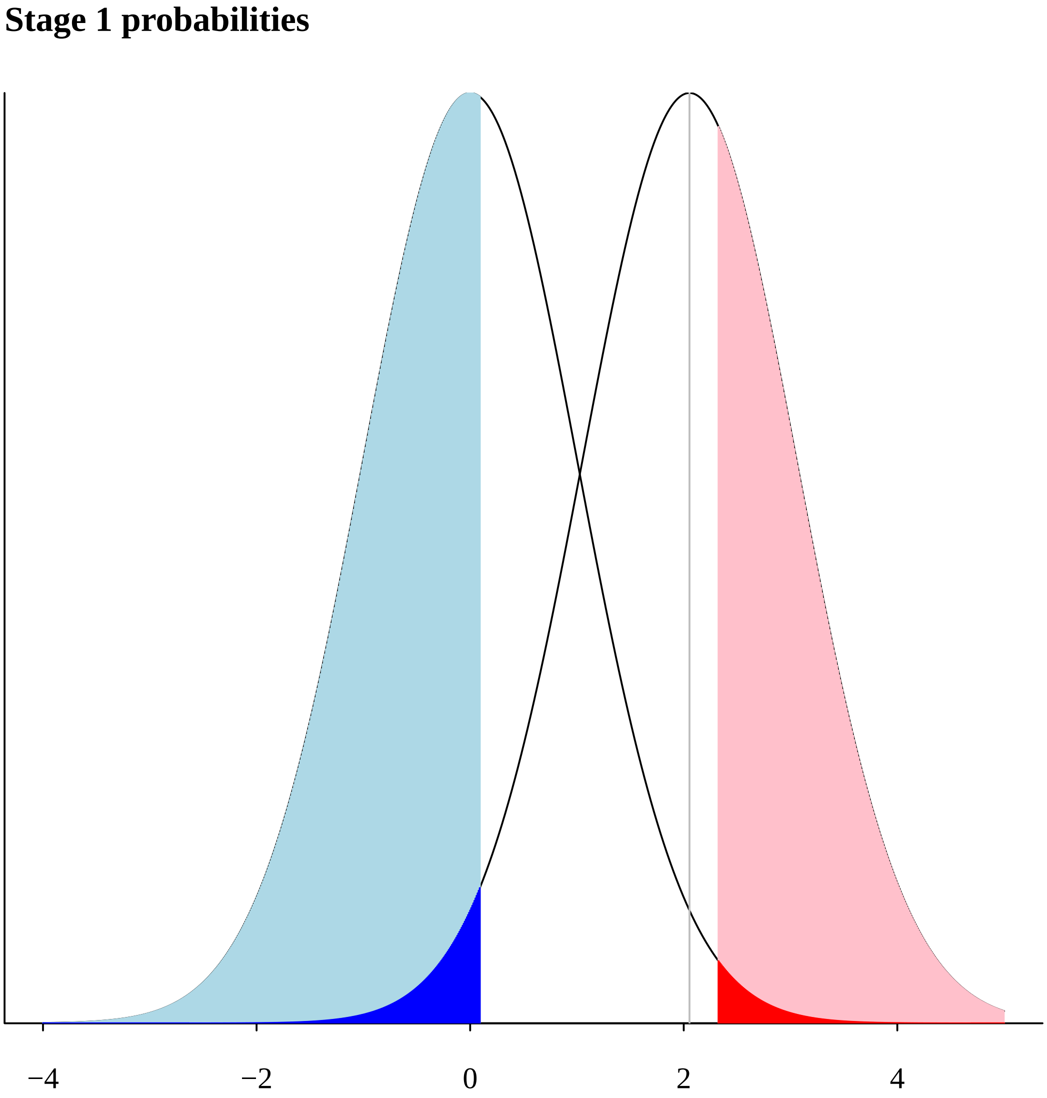
Stage 1 design
Let's decide to look at the data when $n_1 = 135$ observations are available from each group. (approximately 40% of N) We also need to decide how much of α and β we want to "spend" in stage 1. Let's choose $\alpha_1 = 0.01$ and $\beta_1 = 0.025$.

$$\begin{aligned} \alpha_1 &= P_0[\text{Reject } H_0 \text{ in stage 1}] \\ &= P_0[Z_1 > k_2] \qquad \text{where } Z_1 \sim Normal(0,1) \end{aligned}$$

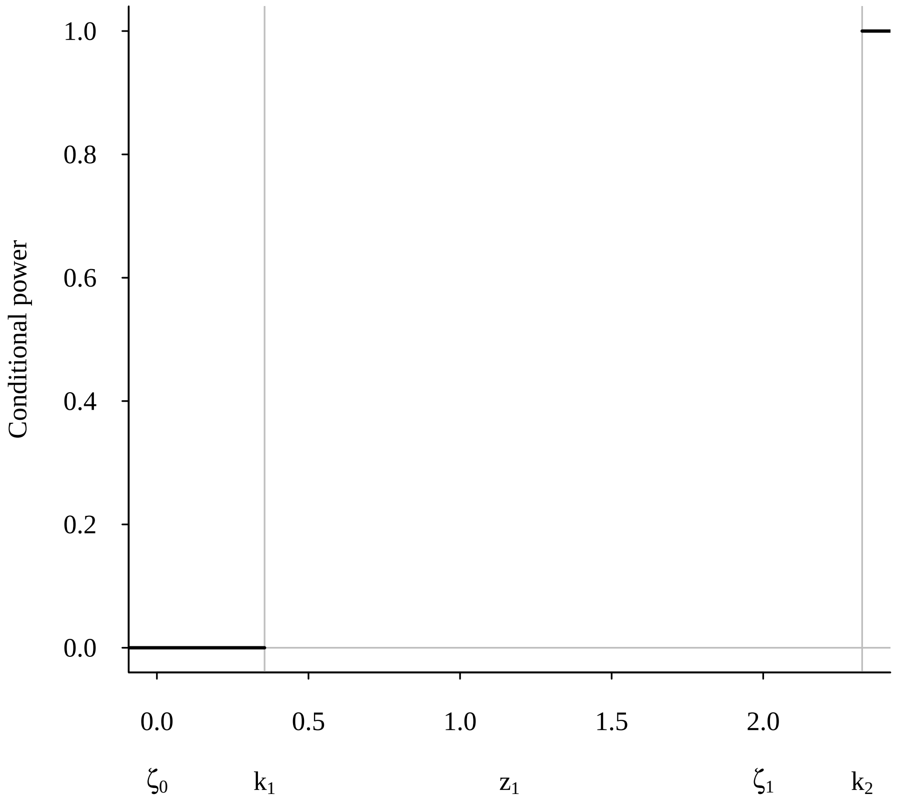
Then $k_2 = 2.326$.

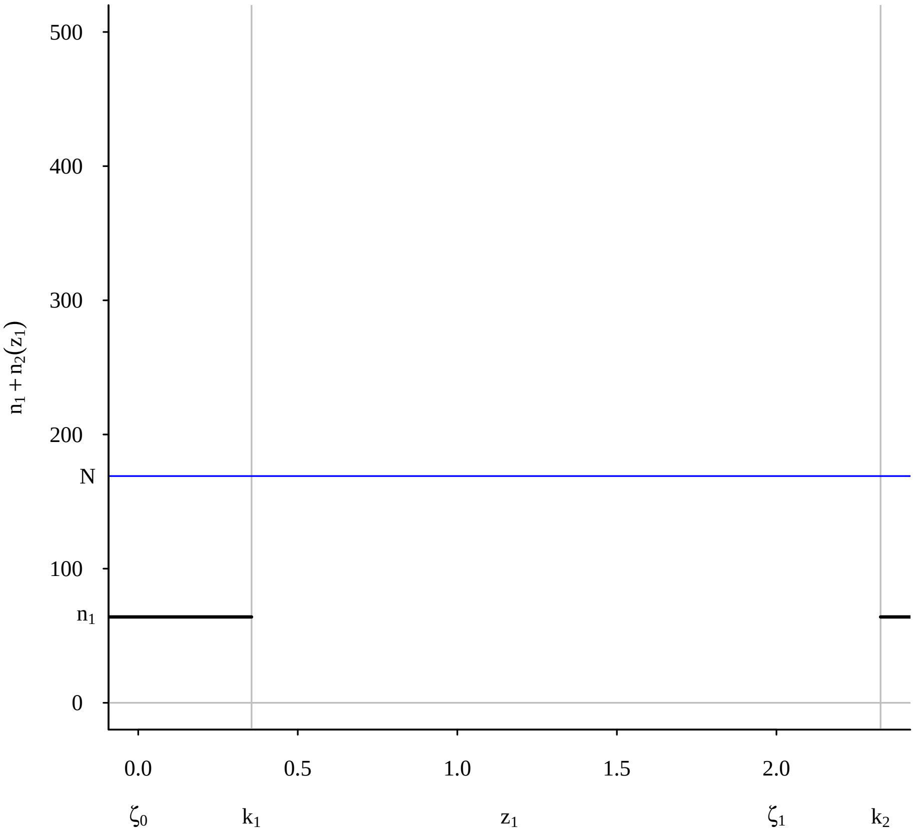
$$\begin{aligned} \beta_1 &= P_1[\text{Accept } H_0 \text{ in stage 1}] \\ &= P_0[Z_1 < k_1] \qquad \text{where } Z_1 \sim Normal(\sqrt{n_1}\xi_1, 1) = Normal(2.05, 1) \end{aligned}$$

Then $k_1 = 0.094$.



Also let's set the maximum sample size to be 500 (approximately 50% increase from N).


Conditional power
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
ζ_0
k_1
z_1
ζ_1
k_2



First, let's look at some stage 1 design characteristics:

Some stage 1 characteristics

				Stage 1	
μ	ξ	ζ	Accept	Continue	Reject
0	0	0	0.538	0.453	0.010
0.5	0.0884	1.027	0.175	0.728	0.097
1	0.1768	2.054	0.025	0.582	0.393

Stage 2 design:
For stage 2, we will choose a conditional type I error function so that the type I error rate is controlled at 2.5%. Also, we will choose a conditional power function so that the overall (unconditional) power is 90%.

– Some math –

We choose to use the following conditional type I error function:

$$CP(Z_1, \delta = \delta_0) = 0.002 + 5.255(z_1 - k_1).$$

And the conditional power function:

$$CP(Z_1, \delta = \delta_1) = 0.75 + 1.004(z_1 - k_1).$$

These two functions will satisfy the type I error rate condition (2.5%) and power condition (90%).

Now consider A-functions of the form $A(z_1, \xi_0) = a_0 + a_1(z_1 - k_1)^2$ and $A(z_1, \xi_1) = b_0 + b_1(z_1 - k_1)$. We can use any A functions as long as they satisfy α and power conditions.

First we pick a_0 (the value of $A(z_1, \xi_0)$ at $z_1 = k_1$ to be 0.002 and solve for a_1 so that

$$\alpha_2 = 0.015 = \int_{k_1}^{k_2} A(z_1, \xi_0) g_1(z_1, \xi_0)\, dz_1.$$

Numerical integration finds $a_1 = 5.2547$

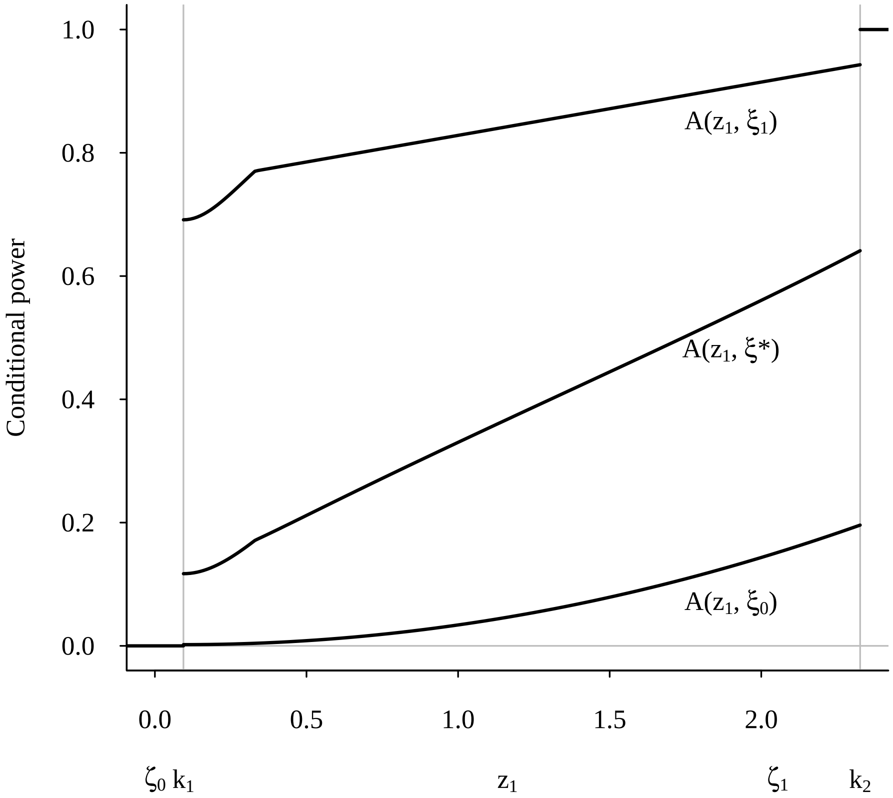
$$A(z_1, \xi_0) = 0.002 + 5.2547(z_1 - k_1)^2.$$

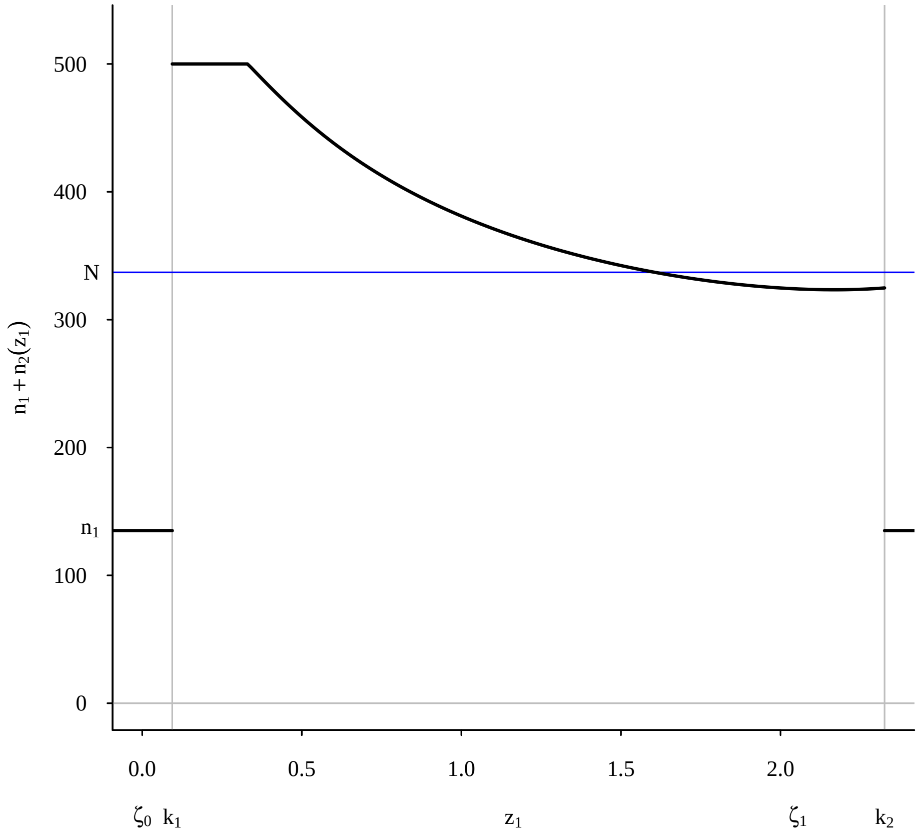
Similarly for $A(z_1, \xi_1)$, by specifying $b_0 = 0.75$, we find $b_1 = 1.004$ so that

$$A(z_1, \xi_1) = 0.75 + 1.004(z_1 - k_1).$$

Using these two A functions, we can compute $n_2(z_1)$, and it turns out $\max\{n_1 + n_2(z_1)\} > 500$, and we need to modify the design a little. It is relatively simple to make small modifications to the design because we understand how the design elements $A(z_1, \xi_0)$, $A(z_1, \xi_1)$, $n_2(z_1)$, and $c(z_1)$, are interrelated.

First while fixing $A(z_1, \xi_0)$, we "tap" $n_2(z_1)$ so that $\max\{n_1 + n_2(z_1)\} = 500$. This action changes $A(z_1, \xi_1)$ slightly resulting a smaller power than 0.90. To make the power 0.90 again, we add a constant to the new $A(z_1, \xi_1)$ but capping the resulting $n_1 + n_2(z_1)$ at 500. The final design is shown below graphically.


1.0
0.8
0.6
0.4
0.2
0.0
Conditional power
$A(z_1, \xi_1)$
$A(z_1, \xi^*)$
$A(z_1, \xi_0)$
0.0
0.5
1.0
1.5
2.0
ζ_0 k_1
z_1
ζ_1
k_2



Design with flat A_0 and flat A_1

				Stage 1		Stage 2		This design		Single stage	
μ	ξ	n_1	Accept	Continue	Reject	Reject	Power	$E[N]$	Max N	Power	N
0	0	135	0.538	0.452	0.010	0.015	0.025	264.2	500	0.025	337
0.25	0.044	135	0.337	0.628	0.035	0.085	0.120	303.8	500	0.125	337
0.50	0.088	135	0.175	0.728	0.097	0.264	0.360	318.4	500	0.367	337
0.75	0.133	135	0.074	0.710	0.216	0.464	0.680	302.7	500	0.681	337
1	0.177	135	0.025	0.582	0.393	0.507	0.900	264.7	500	0.900	337

In the literature, many specific A functions have been proposed. A few examples include:

- Proschan & Hunsberger (1995)

$$A_{\text{PH}}(z_1,\xi) = 1 - \Phi\left[\sqrt{n_1}\sqrt{(k_2 - \xi\sqrt{n_1})^2 - (z_1 - \xi\sqrt{n_1})^2}\right]$$

- Chen, DeMets & Lan (2004)

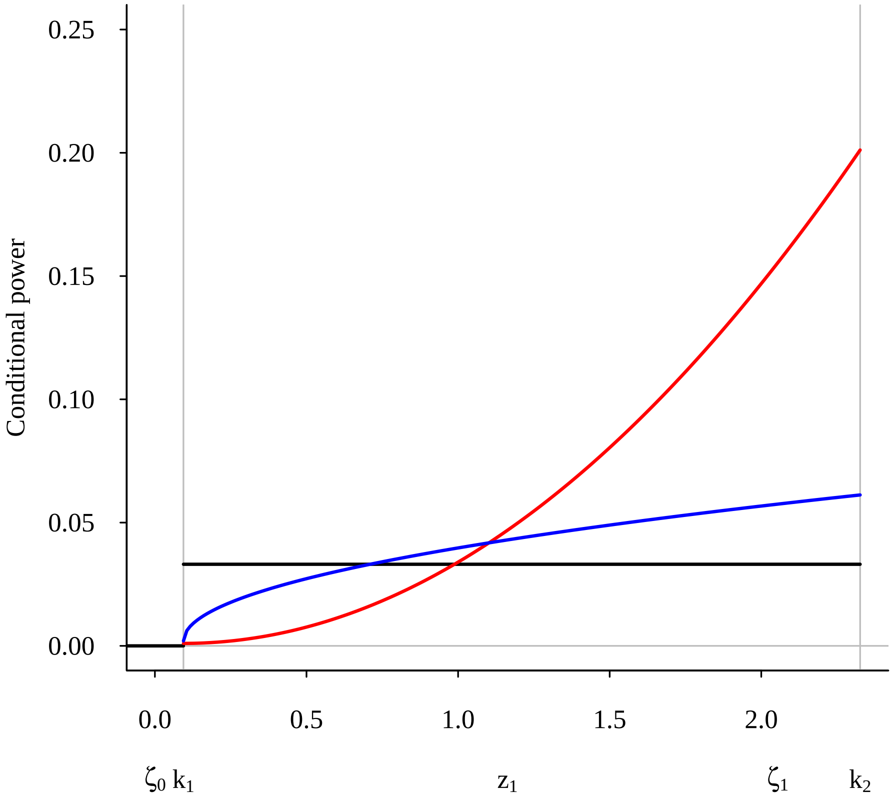
$$A_{\text{CDL}}(z_1, \xi) = 1 - \Phi\left[\sqrt{2} z_\alpha - z_1 - \xi\sqrt{n_1}\right]$$

$A(z_1, z_1)$ "Conditional power under the current trend"

9.7 Unspecified designs

The minimum requirement to control type I error rate is to pre-specify $A(z_1, \xi_0)$ function that satisfies α condition. Then after the first stage, when the actual z_1 from the data are available, pick $n_2(z_1)$ so that conditional powers at any a value of ξ (other than ξ_0) can be set.

If we allow even $A(z_1, \xi_0)$ to be specified after the fist stage, type I error rate cannot be controlled. There exist many (in fact infinite number of) $A(z_1, \xi_0)$ functions that give the desired value of α_1 (0.01 in our example). Depending on z_1 the required sample size to guarantee a certain conditional power differs. We cannot choose an $A(z_1, \xi_0)$ function that gives the minimum sample size for the observed z_1. Roughly speaking, when the conditional type I error rate at the observed z_1 is large, the required sample size is small for the same conditional power. All three conditional type I error rates in the following plot give $\alpha = 0.025$.


Conditional power
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.5
1.0
1.5
2.0
ζ_0 k_1
z_1
ζ_1
k_2

–END–

9.8 Ordering of sample space

To compute p-values and confidence interval (through inverting hypothesis tests), we need to define an ordering of sample space; however, this task is difficult because of sample size difference for potential values of z_1.

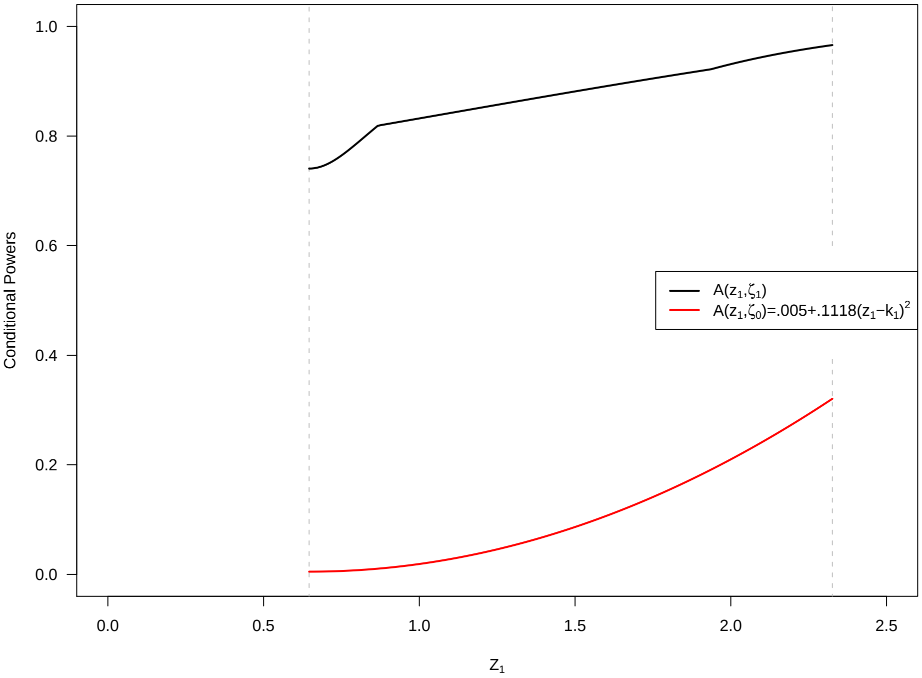
One useful fact (not too difficult to show) is that the decision rule, "reject if $Z_2 > c(z_1)$" is equivalent to the rule "reject if stage 2 (conditional) p-value is less than $A(z_1, \xi_0)$ evaluated at the observed z_1." So we can compute a (conditional) p-value just using the stage 2 data, $P_0[Z_2 > z_2]$, and compare it to the conditional type I error rate computed at the observed z_1.

Suppose the red line in the previous plot is used as $A(z_1, \xi_0)$, then

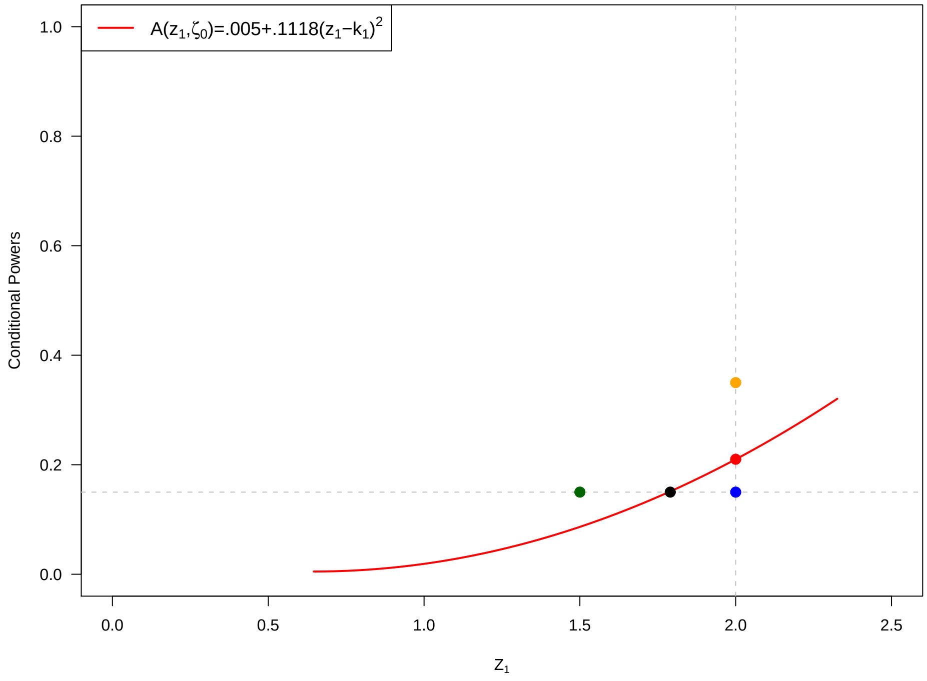
- if $z_1 = 2.0$ and stage 2 conditional p-value is 0.10 then H_0 will be rejected because this p-value is less than $A(z_1, \xi_0)$ i.e., below the red line.
- if $z_1 = 1.0$ and stage 2 conditional p-value is 0.10 then H_0 will not be rejected.

Therefore, we need an ordering of the sample space that takes into account not only the sample size of stage 2, $n_2(z_1)$, but also the conditional type I error rate for the stage 2, $A(z_1, \xi_0)$.

Suppose we choose to use $A(z_1, \xi_0)$ and $A(z_1, \xi_1)$ shown in the plot below.



And let's consider the following 5 sample paths indicated by the conditional p values. Can we order the strength of evidence against H_0 for these data?



- When z_1 is the same, the second stage sample size is the same, so it should be simple to order the sampling paths. The smaller p value, the stronger evidence against H_0.
  Blue $\succ$ Red $\succ$ Yellow.

- When the conditional p values are the same, then we can order them by the strength of evidence in the first stage.
  Blue $\succ$ Black $\succ$ Green.

- The black and red dots should indicate equal strength of evidence because they both result in "just" rejecting H_0.

So in the above picture, the only unclear ordering is between Green and Yellow.
The third rule gives a hint as to how to proceed; the data leading to the black and red dots indicate that those data have just enough evidence to reject H_0. The blue dot is for a sampling path that gives stronger evidence against H_0; we could reject H_0' that is more extreme.
We can find a value of ξ_0^* (equivalently, μ_0^*) so that $A(z_1,\xi_0^*)$ goes through the blue dot, and say we could have rejected $H_0^* : \mu = \mu_0^*$.

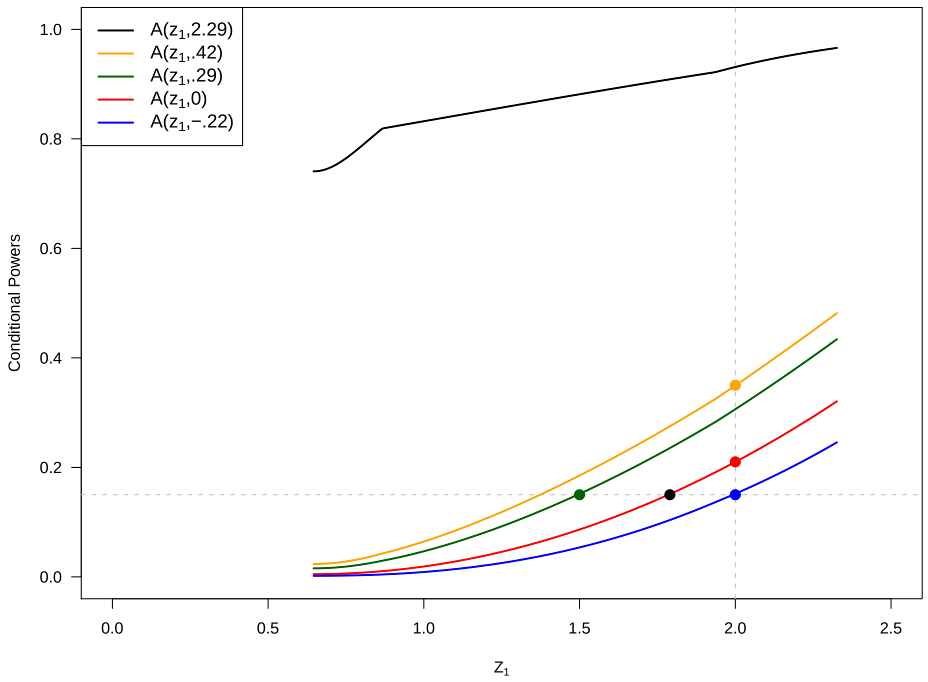
Technically speaking, we can find ξ^* by solving the following:

$$p = A(z_1, \xi_0^*) = 1 - \Phi\left[c(z_1) - \sqrt{n_2(z_1)}\,\xi_0^*\right]$$

Note that $p = P_0[Z_2 > z_2]$, and once a value of z_1 is observed, we can evaluate $c(z_1)$ and $n_2(z_1)$, so the only unknown quantity in the above expression is ξ_0^*.

From the above picture, we know the ordering is: Blue $\succ$ Red $=$ Black $\succ$ Green $\succ$ Yellow. And “some as or more extreme” than the observed is anything on and below the line, and we can compute the p-value by computing

$$\int_{k_2}^{\infty} g_1(z_1, \xi_0)\,dz_1 + \int_{k_1}^{k_2} A(z_1, \xi_0^*) g_1(z_1, \xi_0)\,dz_1$$



This method (ordering) guarantees that the p-value and the corresponding hypothesis testing are consistent (p-value $< \alpha$ iff H_0 is rejected).

And it can be shown that when $n_2(z_1)$ and $c_u(z_1)$ {critical value for the combined statistic} are constants, this ordering reduces to the stage-wise ordering.

9.9 Predictive power

With an unspecified design, some people are reluctant to use the conditional power to determine the design of the second stage. One issue is that where to compute the conditional power is not always clear.

The original alternative is usually a reasonable choice ($A(z_1, \xi_1)$). However, when the observed z_1 is much different (smaller) from ξ_1 we may not be interested in the conditional power at ξ_1 but at some smaller value that is still clinically meaningful. (Minimum clinically relevant alternative $= \xi_1^\dagger$)

Another popular choice is $\hat{\xi} \equiv z_1/\sqrt{n_1}$ ("alternative under the current trend").

Or maybe we should compute the conditional power at somewhere in between ξ_1 and $\xi_1^\dagger$. Average?

Now we are talking like a Bayesian because we are talking about an average of ξs which are, for a frequentist, parameters. Maybe we have a prior distribution of ξ (or equivalently μ). And a posterior distribution of ξ after the first stage, $\pi(\xi|z_1)$, and we can compute a weighted average of the conditional power with respect to the posterior distribution. Something like

$$\int_{-\infty}^{\infty} A(z_1, \xi)\pi(\xi|z_1)\, d\xi,$$

and this is often called a predictive power given the stage one data.

The conditional power is a frequentist concept, and it is computed at one value of ξ. The predictive power is a Bayesian concept, and it is a weighted average of the conditional power with respect to a posterior distribution of ξ.