

# ST 520: Statistical Principles of Clinical Trials and Epidemiology

Daowen Zhang

*NC State University*  
DEPARTMENT OF STATISTICS



zhang@stat.ncsu.edu

<http://www4.stat.ncsu.edu/~dzhang2>

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Brief Introduction to Epidemiology . . . . .	7
1.2	Brief Introduction and History of Clinical Trials . . . . .	29
<b>2</b>	<b>Phase I and Phase II Clinical Trials</b>	<b>42</b>
2.1	Phase I clinical trials (from Dr. Marie Davidian) . . . . .	44
2.2	Phase II Clinical Trials . . . . .	99
<b>3</b>	<b>Phase III Clinical Trials</b>	<b>133</b>
<b>4</b>	<b>Randomization</b>	<b>159</b>
<b>5</b>	<b>Some Additional Issues in Phase III Clinical Trials</b>	<b>201</b>
<b>6</b>	<b>Sample Size Calculations</b>	<b>209</b>
<b>7</b>	<b>Comparing More Than Two Treatments</b>	<b>234</b>

---

<b>8</b>	<b>Causality, Non-compliance and Intent-to-treat</b>	<b>275</b>
8.1	Causality and Counterfactual Random Variables . . . . .	275
8.2	Noncompliance and Intent-to-treat analysis . . . . .	282
8.3	A Causal Model with Noncompliance . . . . .	288
<b>9</b>	<b>Survival Analysis in Phase III Clinical Trials</b>	<b>300</b>
9.1	Describing the Distribution of Time to Event . . . . .	301
9.2	Censoring and Life-Table Methods . . . . .	310
9.3	Kaplan-Meier or Product-Limit Estimator . . . . .	320
9.4	Two-sample Log-rank Tests . . . . .	326
9.5	Power and Sample Size Based on the Log-rank Test . . . . .	336
9.6	K-Sample Log-rank Tests . . . . .	352
9.7	Sample-size Considerations for the K-sample Log-rank Test	355
9.8	Analyzing Data Using $K$ -sample Log-rank Test . . . . .	357
<b>10</b>	<b>Early Stopping of Clinical Trials</b>	<b>360</b>
10.1	General Issues in Monitoring Clinical Trials . . . . .	360

---

10.2 Information Based Design and Monitoring . . . . . 366

10.3 Choice of Boundaries . . . . . 382

10.4 Power and Sample Size in Terms of Information . . . . . 388

# 1 Introduction

Two areas of studies on human beings: **EPIDEMIOLOGY** and **CLINICAL TRIALS**

**EPIDEMIOLOGY**: Systematic study of disease etiology (causes and origins of disease) using observational data (i.e. data collected from a population not under a controlled experimental setting).

- Second hand smoking and lung cancer
- Air pollution and respiratory illness
- Diet and Heart disease
- Water contamination and childhood leukemia
- Finding the prevalence and incidence of HIV infection and AIDS

**CLINICAL TRIALS:** The evaluation of intervention (treatment) on disease in a controlled experimental setting.

- The comparison of AZT versus no treatment on the length of survival in patients with AIDS
- Evaluating the effectiveness of a new anti-fungal medication on Athlete's foot
- Evaluating hormonal therapy on the reduction of breast cancer (Womens Health Initiative)

## 1.1 Brief Introduction to Epidemiology

**I. Cross-sectional study:** data are obtained from a random sample at one point in time. This gives a snapshot of a population.

- **Example:** Based on a survey or a random sample thereof, we determine the proportion of individuals with heart disease at one time point. This is referred to as the prevalence of disease. The prevalence can be broken down by age, race, sex, socio-economic status, geographic, etc.
- Important public health information can be obtained. Useful in determining how to allocate health care resources.
- However such data are generally not very useful in determining causation.

- If exposure ( $E$ ) and disease ( $D$ ) are binary (yes/no), data from a cross-sectional study can be represented as

	$D$	$\bar{D}$	
$E$	$n_{11}$	$n_{12}$	$n_{1+}$
$\bar{E}$	$n_{21}$	$n_{22}$	$n_{2+}$
	$n_{+1}$	$n_{+2}$	$n_{++}$

where  $E$  = exposed (to risk factor),  $\bar{E}$  = unexposed;  $D$  = disease,  $\bar{D}$  = no disease.

Here all 4 counts  $n_{11}, n_{12}, n_{21}, n_{22}$  are random variables.

$$(n_{11}, n_{12}, n_{21}, n_{22})$$

$$\sim \text{multinomial}(n_{++}, P[DE], P[\bar{D}E], P[D\bar{E}], P[\bar{D}\bar{E}]).$$

- With this study, we can obtain estimates of the following parameters of interest

prevalence of disease  $P[D]$  (by  $\frac{n_{+1}}{n_{++}}$ )

exposure probability  $P[E]$  (by  $\frac{n_{1+}}{n_{++}}$ )

prevalence of disease among exposed  $P[D|E]$  (by  $\frac{n_{11}}{n_{1+}}$ )

prevalence of disease among unexposed  $P[D|\bar{E}]$  (by  $\frac{n_{21}}{n_{2+}}$ )

...

- **relative risk**  $\psi$  of getting disease between exposed and un-exposed:

$$\psi = \frac{P[D|E]}{P[D|\bar{E}]}$$

- ★  $\psi > 1 \Rightarrow$  positive association
  - ★  $\psi = 1 \Rightarrow$  no association
  - ★  $\psi < 1 \Rightarrow$  negative association
- Estimate of  $\psi$  from a cross-sectional study:

$$\hat{\psi} = \frac{\hat{P}[D|E]}{\hat{P}[D|\bar{E}]} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

- odds ratio  $\theta$  of getting disease between exposed and un-exposed:

$$\theta = \frac{P[D|E]/(1 - P[D|E])}{P[D|\bar{E}]/(1 - P[D|\bar{E}])}.$$

$$\star \psi > 1 \iff \theta > 1$$

$$\star \psi = 1 \iff \theta = 1$$

$$\star \psi < 1 \iff \theta < 1$$

- Estimate of  $\theta$  from a cross-sectional study:

$$\begin{aligned} \hat{\theta} &= \frac{\hat{P}[D|E]/(1 - \hat{P}[D|E])}{\hat{P}[D|\bar{E}]/(1 - \hat{P}[D|\bar{E}])} \\ &= \frac{n_{11}/n_{1+}/(1 - n_{11}/n_{1+})}{n_{21}/n_{2+}/(1 - n_{21}/n_{2+})} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \end{aligned}$$

- Variance of  $\log(\hat{\theta})$  (log is the **natural log**):

$$\widehat{\text{var}}(\log(\hat{\theta})) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}.$$

$n_{++}$  and  $n_{ij}$  have to be large.

- A  $(1 - \alpha)$  CI for  $\log(\theta)$ :

$$\log(\hat{\theta}) \pm z_{\alpha/2} [\widehat{\text{Var}}(\log(\hat{\theta}))]^{1/2}.$$

Exponentiating both limits  $\implies$  a  $(1 - \alpha)$  CI for  $\theta$ .

- Alternatively,

$$\widehat{\text{Var}}(\widehat{\theta}) = \widehat{\theta}^2 \left[ \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right]$$

- A  $(1 - \alpha)$  CI for  $\theta$ :

$$\widehat{\theta} \pm z_{\alpha/2} [\widehat{\text{Var}}(\widehat{\theta})]^{1/2}.$$

- When  $\alpha = 0.05$ ,  $z_{\alpha/2} = 1.96$ .

- If the disease under study is a rare one, then

$$P[D|E] \approx 0, \quad P[D|\bar{E}] \approx 0.$$

In this case, we have

$$\theta \approx \psi.$$

**II. Prospective study:** a cohort of individuals are identified who are free of a particular disease under study and data are collected on certain risk factors; i.e. smoking status, drinking status, exposure to contaminants, age, sex, race, etc. These individuals are then followed over some specified period of time to determine whether they get disease or not. The relationships between the probability of getting disease during a certain time period (called incidence of the disease) and the risk factors are then examined.

- Data from a prospective study may be summarized as

	$D$	$\bar{D}$	
$E$	$n_{11}$	$n_{12}$	$n_{1+}$
$\bar{E}$	$n_{21}$	$n_{22}$	$n_{2+}$

$n_{1+}$  and  $n_{2+}$  are fixed sample sizes for each group.

- $n_{11}$  and  $n_{21}$  have the following distributions:

$$n_{11} \sim \text{Bin}(n_{1+}, P[D|E]), \quad n_{21} \sim \text{Bin}(n_{2+}, P[D|\bar{E}]).$$

- Relative risk  $\psi$  and odds-ratio  $\theta$  can be estimated in the same way!

- One problem in a prospective study is drop-out before the event is observed.
- **Example:** 40,000 British doctors were followed for 10 years. The following data were collected:

Table 1: *Death Rate from Lung Cancer per 1000 person years.*

# cigarettes smoked per day	death rate
0	.07
1-14	.57
15-24	1.39
25+	2.27

For presentation purpose, the estimated rates are multiplied by 1000.

- Denote  $T =$  time to death due to lung cancer; the death rate at time  $t$  is defined by

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{P[t \leq T < t + h | T \geq t]}{h}.$$

Assume the death rate  $\lambda(t)$  is a constant  $\lambda$ , then it can be estimated by

$$\hat{\lambda} = \frac{\text{total number of deaths from lunge cancer}}{\text{total person years of smoking during the 10 year period}}.$$

- The probability of dying from lung cancer in one year:

$$P[D] = P[t \leq T < t + 1 | T \geq t] = 1 - e^{-\lambda} \approx \lambda, \quad \text{if } \lambda \text{ is very small.}$$

- $\hat{P}[D]$  for non-smoking British doctors:

$$\hat{P}[D] = 0.07/1000 = 0.00007$$

- $\hat{P}[D]$  for the heaviest smokers:

$$\hat{P}[D] = 2.27/1000 = 0.00227$$

- Relative risk of dying from lung cancer in one year between heavy smokers and non-smokers:

$$\hat{\psi} = 0.00227/0.00007 = 2.27/0.07 = 32.43.$$

- Odds-ratio of dying from lung cancer in one year between heavy smokers and non-smokers:

$$\hat{\theta} = \frac{.00227/(1 - .00227)}{.00007/(1 - .00007)} = 32.50.$$

**III. Retrospective or Case-Control study:** Individuals with disease (called cases) and individuals without disease (called controls) are identified. Using records or questionnaires the investigators go back in time and ascertain exposure status and risk factors from their past. Such data are used to estimate the relative risk of developing disease between exposed and un-exposed.

- Example: A sample of 1357 male patients with lung cancer (cases) and a sample of 1357 males without lung cancer (controls) were surveyed about their past smoking history. This resulted in the following table:

smoke	cases	controls
yes	1,289	921
no	68	436
	1357	1357

We would like to estimate the relative risk  $\psi$  or the odds-ratio  $\theta$  of getting lung cancer between smokers and non-smokers.

- In general, data from a case-control study can be represented by the following  $2 \times 2$  table:

	$D$	$\bar{D}$
$E$	$n_{11}$	$n_{12}$
$\bar{E}$	$n_{21}$	$n_{22}$
	$n_{+1}$	$n_{+2}$

- By the study design, the margins  $n_{+1}$  and  $n_{+2}$  are fixed numbers, and  $n_{11}$  and  $n_{12}$  are random variables having the following distributions:

$$n_{11} \sim \text{Bin}(n_{+1}, P[E|D]), \quad n_{12} \sim \text{Bin}(n_{+2}, P[E|\bar{D}]).$$

- We hope to estimate the relative risk  $\psi$  in a case-control study

$$\psi = \frac{P[D|E]}{P[D|\bar{E}]}$$

But we can only estimate  $P[E|D]$  and  $P[E|\bar{D}]$ .

- What if we treat the case-control study as a prospective or cross-sectional study and use the incorrect formulas to estimate  $\psi$ ?

$$\hat{P}[D|E] = \frac{n_{11}}{n_{1+}} = \frac{n_{11}}{n_{11} + n_{12}} \text{ (incorrect!)}$$

$$\hat{P}[D|\bar{E}] = \frac{n_{21}}{n_{2+}} = \frac{n_{21}}{n_{21} + n_{22}} \text{ (incorrect!)}$$

- We can make  $\hat{P}[D|E]$  and  $\hat{P}[D|\bar{E}]$  go to one! (Not sensible!)

- For our example, if incorrect formulas are used  $\implies$

$$\hat{P}[D|E] = \frac{1289}{1289 + 921} = 0.583 \text{ (incorrect!)}$$

$$\hat{P}[D|\bar{E}] = \frac{68}{68 + 436} = 0.135 \text{ (incorrect!)}$$

$$\hat{\psi} = \frac{\hat{P}[D|E]}{\hat{P}[D|\bar{E}]} = \frac{0.583}{0.135} = 4.32 \text{ (incorrect!).}$$

- Let us try to estimate the **odds ratio**:

$$\begin{aligned}
 \theta &= \frac{P[D|E]/(1 - P[D|E])}{P[D|\bar{E}]/(1 - P[D|\bar{E}])} \\
 &= \frac{P[D|E]/P[\bar{D}|E]}{P[D|\bar{E}]/P[\bar{D}|\bar{E}]} \\
 &= \frac{P[D|E]/P[D|\bar{E}]}{P[\bar{D}|E]/P[\bar{D}|\bar{E}]} \\
 &= \frac{P[E|D]/P[\bar{E}|D]}{P[E|\bar{D}]/P[\bar{E}|\bar{D}]} \\
 &= \frac{P[E|D]/(1 - P[E|D])}{P[E|\bar{D}]/(1 - P[E|\bar{D}])}.
 \end{aligned}$$

- right hand side = odds-ratio of being exposed between cases and controls and can be estimated.

- From the distributions:

$$n_{11} \sim \text{Bin}(n_{+1}, P[E|D]), \quad n_{12} \sim \text{Bin}(n_{+2}, P[E|\bar{D}]).$$

$\implies$

$$\hat{P}[E|D] = \frac{n_{11}}{n_{+1}}, \quad \hat{P}[E|\bar{D}] = \frac{n_{12}}{n_{+2}},$$

- $\theta$  can be estimated by

$$\begin{aligned} \hat{\theta} &= \frac{\hat{P}[E|D]/(1 - \hat{P}[E|D])}{\hat{P}[E|\bar{D}]/(1 - \hat{P}[E|\bar{D}])} \\ &= \frac{n_{11}/n_{+1}/(1 - n_{11}/n_{+1})}{n_{12}/n_{+2}/(1 - n_{12}/n_{+2})} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \end{aligned}$$

- The same formula!  $\text{var}(\log(\hat{\theta}))$  and  $\text{var}(\hat{\theta})$  have the same formulas too, inference similar.

- Lung cancer example:

$$\hat{\theta} = \frac{1289 \times 436}{921 \times 68} = 8.97.$$

- If we can view lung cancer as a rare event, then the relative risk

$$\hat{\psi} \approx \hat{\theta} = 8.97.$$

- Smokers are 9 times as likely as non-smokers to develop lung cancer.
- The **incorrect** estimate of the relative risk 4.32 is too low.

## Pros and Cons of a case-control study

- Pros

- ★ Can be done more quickly. You don't have to wait for the disease to appear over time.
- ★ If the disease is rare, a case-control design is more efficient:

$$\widehat{\text{var}}(\log(\hat{\theta})) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}.$$

- Cons

- ★ Bias in getting exposure information! This can be a severe drawback.

## 1.2 Brief Introduction and History of Clinical Trials

- Definition of a clinical trial:
  - ★ A clinical trial is a study in human subjects in which treatment (intervention) is initiated specifically for therapy evaluation.
  - ★ A prospective study comparing the effect and value of intervention against a control in human beings.
  - ★ A clinical trial is an experiment which involves patients and is designed to elucidate the most appropriate treatment of future patients.
  - ★ A clinical trial is an experiment testing medical treatments in human subjects.

- **Historical perspective**

- ★ Historically, the quantum unit of clinical reasoning has been the case history and the primary focus of clinical inference has been the individual patient. Inference from the individual to the population was informal. The advent of formal experimental methods and statistical reasoning made this process rigorous.

By statistical reasoning or inference we mean the use of results on a limited sample of patients to infer how treatment should be administered in the general population who will require treatment in the future.

★ **Early History**

**1600 East India Company** (A British company founded in 1600)

In the first voyage of four ships– only one ship was provided with lemon juice. This was the only ship relatively free of scurvy.

**Note:** This is observational data and a simple example of an epidemiological study.

★ **1753 James Lind** (British doctor, Father of Nautical Medicine)

“I took 12 patients in the scurvy aboard the Salisbury at sea. The cases were as similar as I could have them... they lay together in one place... and had one common diet to them all...

To two of them was given a quart of cider a day, to two an elixir of vitriol, to two vinegar, to two oranges and lemons, to two a course of sea water, and to the remaining two the bigness of a nutmeg. The most sudden and visible good effects were perceived from the use of oranges and lemons, one of those who had taken them being at the end of six days fit for duty... and the other appointed nurse to the sick...

**Note:** This is an example of a controlled clinical trial.

★ **1794 Rush** (American doctor) *Treatment of yellow fever by bleeding*

“I began by drawing a small quantity at a time. The appearance of the blood and its effects upon the system satisfied me of its safety and efficacy. Never before did I experience such sublime joy as I now felt in contemplating the success of my remedies... The reader will not wonder when I add a short extract from my notebook, dated 10th September. “Thank God”, of the one hundred patients, whom I visited, or prescribed for, this day, I have lost none.”

- ★ **Louis** (French physician): Lays a clear foundation for the use of the *numerical method* in assessing therapies.

Louis (1835) studied the value of bleeding as a treatment of pneumonia, erysipelas and throat inflammation and found no demonstrable difference in patients bled and not bled. This finding contradicted current clinical practice in France and instigated the eventual decline in bleeding as a standard treatment. Louis had an immense influence on clinical practice in France, Britain and America and can be considered the founding figure who established clinical trials and epidemiology on a scientific footing.

Table 2: *Pneumonia: Effects of Blood Letting*

Days bled after onset	Died	Lived	proportion surviving
1-3	12	12	50%
4-6	12	22	65%
7-9	3	16	84%

In 1827: 33,000,000 leeches were imported to Paris.

In 1837: 7,000 leeches were imported to Paris.

- **Modern clinical trials:**

- ★ The **first** clinical trial with a properly randomized control group was set up to study streptomycin in the treatment of pulmonary tuberculosis, sponsored by the Medical Research Council, 1948 (UK). This was a multi-center clinical trial where patients were randomly allocated to streptomycin + bed rest versus bed rest alone.

The evaluation of patient x-ray films was made independently by two radiologists and a clinician, each of whom did not know the others evaluations or which treatment the patient was given.

Both patient survival and radiological improvement were significantly better on streptomycin.

★ **The field trial of the Salk Polio Vaccine:**

In 1954, 1.8 million first to third graders participated in the trial to assess the effectiveness of the Salk vaccine in preventing paralysis or death from poliomyelitis.

Incidence is low (1 in 2000).

**Randomized component:** 0.8 million children were randomized in a **double-blind placebo-controlled** trial.

**Result:** Incidence of polio in treated group is less than half of that in the control group.

**Non-randomized component:** Second graders were offered vaccine and first and third graders were formed control group.

**Result:** similar.

**However**, it turned out that the incidence of polio among children (second graders) offered vaccine and not taking it (non-compliers) was different than those in the control group (first and third graders).

**Question:** were treated children (second graders) and the control (first and third graders) similar?

- **Government sponsors clinical trials:** NIH (National Institutes of Health)
  - ★ NHLBI- (National Heart Lung and Blood Institute) funds individual and often very large studies in heart disease.
  - ★ NIAID- (National Institute of Allergic and Infectious Diseases) Much of their funding now goes to clinical trials research for patients with HIV and AIDS.
  - ★ NIDDK- (National Institute of Diabetes and Digestive and Kidney Diseases). Funds large scale clinical trials in diabetes research.

- **Pharmaceutical Industry:**

- ★ Before World War II no formal requirements were made for conducting clinical trials before a drug could be freely marketed.
- ★ In 1938, animal research was necessary to document toxicity, otherwise human data could be mostly anecdotal.
- ★ In 1962, it was required that an “adequate and well controlled trial” be conducted.
- ★ In 1969, it became mandatory that evidence from a randomized clinical trial was necessary to get marketing approval from the Food and Drug Administration (FDA).
- ★ More recently there is effort in standardizing the process of drug approval worldwide. This has been through efforts of the International Conference on Harmonization (ICH).  
website:  
<http://www.pharmweb.net/pwmirror/pw9/ifpma/ich1.html>
- ★ There are more clinical trials currently taking place than ever

before. The great majority of the clinical trial effort is supported by the Pharmaceutical Industry for the evaluation and marketing of new drug treatments.

## 2 Phase I and Phase II Clinical Trials

### Phases of Clinical Trials:

- Preclinical (drug discovery): experimentation before a drug is given to human subjects
  - ★ lab testing for biologic activity (in vitro)
  - ★ testing on animals (in vivo)

- Clinical:
  - ★ **Phase I:** To explore possible toxic effects of drugs and determine a tolerated dose for further experimentation. Also explore pharmacology of the drug and investigate its interaction with other drugs, food and alcohol (these can be parallel to phase II-phase III trials). First-in-human trials.
  - ★ **Phase II:** Screening and feasibility by initial assessment for therapeutic effects; dose finding and further assessment of toxicities (safety and tolerability).
  - ★ **Phase III:** Comparison of new intervention (drug or therapy) to the current standard of treatment; both with respect to efficacy and toxicity.
  - ★ **Phase IV:** (post marketing) Observational study of morbidity/adverse effects.

## 2.1 Phase I clinical trials (from Dr. Marie Davidian)

**Broad definition:** Phase I trials are the first studies in which a new drug is administered to human subjects.

- Previous studies in the laboratory (in vitro)
- Previous studies in animals, e.g. rats, dogs (in vivo)

**Objectives:** Before efficacy (activity of a drug on disease) of the drug may be established, first must

- Determine a “safe,” “tolerable” dose (through dose-escalation)
- Develop an appropriate schedule of administration
- Gain understanding of the *pharmacology* of the drug
- Need to examine interaction effects (drug-drug, drug-food, drug-alcohol) for safety profile and proper labeling.

**In addition:** Do this in a timely manner, using a small number of subjects.

**Features:**

- Most are not comparative but rather are “informational”
- “Interaction studies” are comparative, but not aimed for efficacy, still “informational” (for safety).

## Types of studies:

- “Dose-finding” studies – determine the maximum dosage that can be given without serious “problems” – these studies are often themselves called “phase I studies” (especially for cancer studies)
- Clinical pharmacology studies – determine the *pharmacokinetics* of the drug to aid in setting dosage schedules, understanding how “problems” are related to amount of drug present
- Drug-drug interaction studies – determine how other commonly used drugs affect important *PK* parameters, using a cross-over or parallel design
- Drug-food interaction studies – determine how food affects important *PK* parameters, using a cross-over or parallel design
- Alcohol/benzo interaction studies – determine how alcohol/benzo affects *PD* parameters, using a cross-over or parallel design

## Pre-clinical Studies

### Before administration to humans:

- *In vitro* studies – laboratory investigations using biological material but not actual organisms; look for biologic activity of the drug
- Dose-finding studies in rodents, large animal species (e.g. dogs)
- Pharmacology studies in rodents, dogs, etc
- Goal – “scale up” previous results to provide first idea of behavior in humans
- Advantage – Ethical issues involved in human experimentation circumvented (coming up)

## Phase I Dose-finding Studies

**Paradox:** Although results of “dose-finding” will be carried forward to be used in studies of efficacy (phase II) and later to comparative trials (phase III) and eventually to routine patient care

- Standard approaches to design and analysis have little statistical justification
- Many texts on clinical trials devote little or no discussion to dose-finding (or pharmacology) studies, e.g. Freedman et al. mention only on p. 3–4!

**Toxicity:** “Problems” that may arise in direct response to administration of the drug – side effects

- Nature depends on the drug
- E.g. change in organ function – a drug to treat cancer may induce irregular heartbeat
- May be life-threatening and irreversible
- May be life-threatening and reversible
- May be non-life-threatening

**Characterization:** Often done on a standard, “graded” scale (especially for cancer research)

- Ordered categories increasing in severity – Grades I – IV or V
- Examples:

	I	II	III	IV
Abdominal Pain	Mild	Moderate (No trt)	Moderate (Trt)	Severe (Hospital)
Creatinine Clearance (cc/min/1.73 m <sup>2</sup> )	60–75	50–59	35–49	< 35

**Defining toxicity:**

- Which toxicities are relevant depends on intervention, nature of likely subjects, clinical judgment – must be defined by the investigators
- What degree of toxicity is “acceptable” must be established

**Terminology: *Dose Limiting Toxicity (DLT)***

- Serious or life-threatening but reversible
- Often used as the definition for dose-finding in cancer research

**Assumption:** Maximum benefit occurs at maximum doses

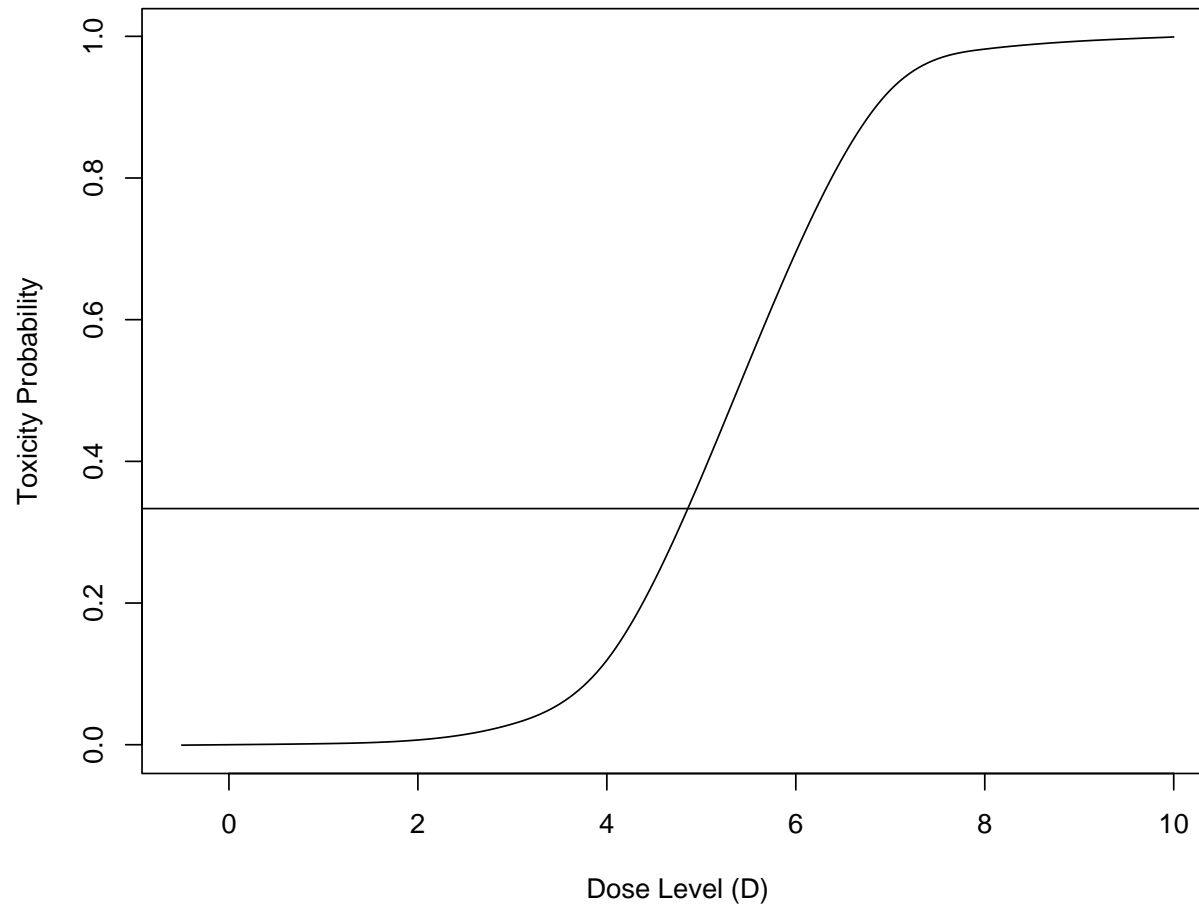
**Main objective:** Determine the *Maximum Tolerated Dose* (MTD)

- The “highest possible while still tolerable” dose
- Involves some level of toxicity we are willing to tolerate because of the drug’s potential benefit
- E.g. “the dose that produces toxicity of grade III or worse in not more than one out of three patients” (often used for anticancer agents)

**Statistically speaking:** Determine the dose at which the proportion of subjects in the population who would develop toxicity if given this dose is  $1/3$

- I.e. the 33rd percentile of the distribution of toxicity in the population

## Illustration of MTD: a hypothetical example



**Subjects:** Nature of subjects used depends on drug

- Healthy volunteers – used where toxicities unlikely to be severe, e.g. topical agents
- Patients with advanced disease – used where subjects have failed other therapies, toxicities likely. e.g. chemotherapy

**Goal:** Establish appropriate dose quickly while exposing as few subjects as possible to suboptimal doses that are likely not to be efficacious

**Ideal design:**

- Select doses  $D_1, \dots, D_k$  for study such that one of these is (close to) the MTD
- Randomize subjects to each dose,  $n_i$  subjects at dose  $i$
- Observe number  $r_i$  exhibiting DLT at each dose, and calculate proportions exhibiting DLT  $p_i = r_i/n_i$
- Model dose-response (probability of toxicity) and fit to observed proportions at each dose
- Estimate MTD from fitted model

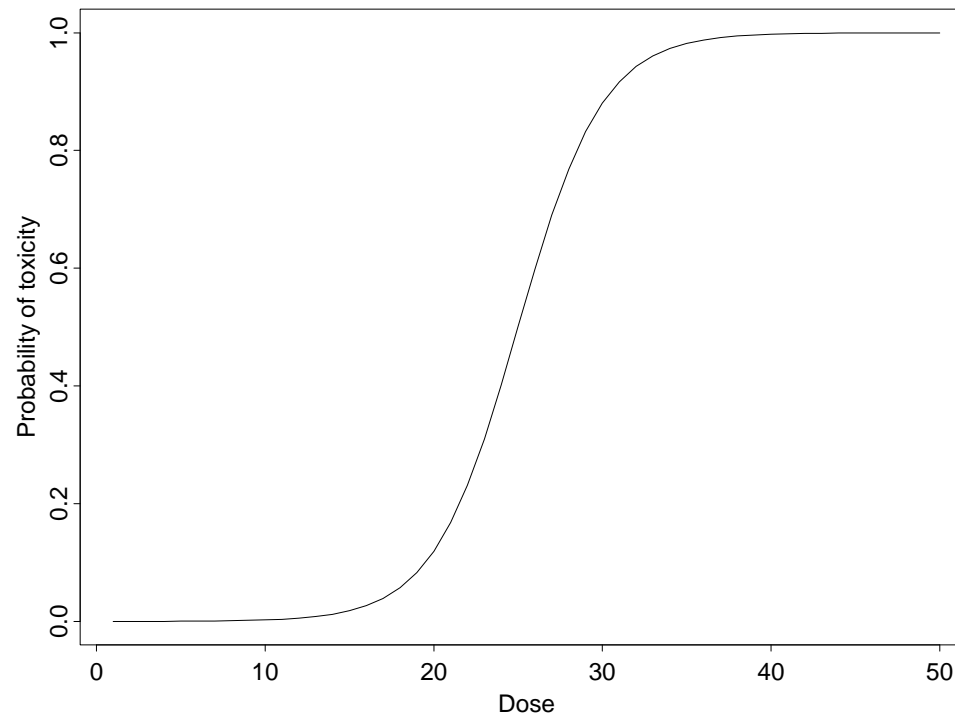
## Models for dose-response:

- $Y = 1$  if toxicity observed for a subject at dose  $d$
- $P(Y = 1|d) = f(d, \beta)$ ,  $f$  monotone in dose
- Dose  $d_0$  associated with specified probability  $p_0 = P(Y = 1|d_0)$  is  $d_0 = f^{-1}(p_0, \beta)$
- E.g. If MTD is defined as dose where toxicity is 33%,  $MTD = f^{-1}(0.33, \beta)$
- Estimate  $\beta$  from data and use to estimate MTD, i.e.  $\widehat{MTD} = f^{-1}(0.33, \hat{\beta})$
- More generally
  - ★ May estimate dose associated with any probability
  - ★ May estimate probability associated with any fixed dose
- Standard approach for animal experiments

---

**Logistic model:**

$$P(Y = 1|d) = \frac{\exp(\beta_0 + \beta_1 d)}{1 + \exp(\beta_0 + \beta_1 d)}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1) = (-10, 0.4)$$



**Problem:** This approach not feasible in human subjects due to *ethical considerations*

- Because drug not previously used in humans, must test at lower doses first before feeling confident enough to move to higher doses
- Do not want to treat many subjects at dose that is too low to do good or too high (toxic)
- Result: cannot simply randomize subjects to different dose levels

**Approach:** “Adaptive” designs – *dose escalation*

- Try a dose in several subjects
- If no toxicities, try a higher dose in several (new) subjects
- Continue until dose is found that yields toxicity
- Sometimes, may “de-escalate” from a dose that is not tolerated
- Many variations on this idea

**Result:** Sample size is not specified in advance; rather, it is an outcome of the study

- Usually, sample size is small ( $\sim 20$ )

**Standard design** (to find MTD): Select a sequence of increasing doses. Start at lowest dose level and administer drug to 3 subjects

1. If no toxicity observed in any of the 3 subjects, escalate to the next dose and begin again
2. If toxicity observed in 2 or more of the 3 subjects, STOP
3. If toxicity observed in exactly 1 subject of the 3, treat 3 additional subjects at this dose. If none of the additional subjects exhibits toxicity, then escalate to the next dose and begin again; otherwise, STOP

**MTD:** Usually defined as one of

- Dose at which trial stops
- Next lowest dose in the sequence
- Some fraction of the last dose

**Determining the initial dose:** The starting dose of the sequence is chosen in a conservative way, e.g.

- From rodent studies, estimate of  $LD_{10}$  is available;  $LD_{10}$  = dose where percentage of rodents exhibiting mortality is 10%
- Use as starting dose 1/10 of the rodent  $LD_{10}$  given on a mg/kg basis (scaled from rodent to human size)
- OR from larger animal (e.g. dog) studies, determine the *toxic dose low* (TDL) = the lowest dose at which any toxicity seen
- Use as starting dose 1/3 of the dog TDL

**Determining the sequence:** Want to select doses in a way that will reveal the “MTD” without requiring an excessive number of dose levels

- Common technique – “modified Fibonacci” sequence
- Usual Fibonacci sequence 1, 1, 2, 3, 5, 8, 13, 21, ...
- Ratio of successive terms  $\rightarrow 1.618$  ( $\approx 62\%$  increase)
- Modify to increase less rapidly with decreasing increments
- Alternative: equally-spaced doses on log scale over range

**Example:** Modified Fibonacci with  $D =$  initial dose

Step	Usual	Modified	% Increment
1	$D$	$D$	–
2	$2 \times D$	$2 \times D$	100
3	$3 \times D$	$3.3 \times D$	67
4	$5 \times D$	$5 \times D$	50
5	$8 \times D$	$7 \times D$	40
6	$13 \times D$	$9 \times D$	29
7	$21 \times D$	$12 \times D$	33
8	$34 \times D$	$16 \times D$	33

**Performance of the standard design: Example**

Dose level	Actual toxicity prob. ( $\pi_i$ ) (unknown in practice)	Prob ( $p_i$ ) of stopping at this dose
1	0.15	0.186
2	0.20	0.237
3	0.25	0.231
4	0.30	0.178
5	0.33	0.096
6	0.50	

- E.g. probability of stopping at dose level 1 ( $\pi_1 = 0.15$ ):

$$\begin{aligned} p_1 &= P[(Z_1 > 1) \cup (Z_1 = 1, Z_2 > 0)] \quad (Z_1, Z_2 \stackrel{iid}{\sim} \text{binomial}(3, \pi_1)) \\ &= P[Z_1 > 1] + P[Z_1 = 1, Z_2 > 0] \\ &= P[Z_1 > 1] + P[Z_1 = 1]P[Z_2 > 0] \\ &= 3\pi_1^2(1 - \pi_1) + \pi_1^3 + 3\pi_1(1 - \pi_1)^2(1 - (1 - \pi_1)^3) = 0.186 \end{aligned}$$

- Given the trial reaches level 2, the probability of stopping at dose level 2 can be calculated in the same way (replace  $\pi_1$  by  $\pi_2$ ), which is 0.2912.
- The probability of stopping at dose level 2 ( $\pi_2 = 0.20$ ):

$$p_2 = 0.2912 \times (1 - p_1) = 0.237.$$

- Chance of ever reaching the 33rd percentile is only 16.8%  
( $1 - p_1 - p_2 - p_3 - p_4$ )
- Given the trial reaches the 33rd percentile (dose level 5), the chance of stopping there is only 57% (calculated similarly to  $p_1$  by replacing  $\pi_1$  with  $\pi_5$ )

**Remarks:** Statistically speaking

- If the true MTD is defined as previously, not clear that the dose announced as MTD is a credible estimate of this quantity
- The design has no intrinsic property that makes it stop at the 33rd or any other percentile of the toxicity distribution
- The announced MTD can only be at or near one of the doses in the sequence used, none of which may be exactly equal to the 33rd percentile
- No basis for accounting for sampling error (standard errors?)
- No appeal to formal statistical model
- Likely to treat most subjects at low doses

**Remarks:**

- The standard design with this method of declaring the MTD is widely used, despite statistical concerns
- The method of MTD determination is favored by investigators (the declared MTD depends on individual patient outcomes)
- The method of MTD determination is of concern to statisticians (sampling error is not taken into account)
- Proposals in the statistical literature for more rigorous analysis and other designs have been made

## Formal approach to analysis of the standard design:

- Assume a statistical model for probability of toxicity at dose  $d$

$$P(Y = 1|d) = f(d, \beta)$$

$Y = 1$  if subject exhibits toxicity, 0 otherwise

- For the  $i$ th group of 3 subjects, let  $Z_i = \#$  of subjects experiencing toxicity, corresponding  $X_i =$  dose level
- It is possible to write out the likelihood for the data  $(Z_i, X_i)$ ,  $i = 1, \dots, n$ , where  $n$  is the (random) sample size
- The MTD as formally defined could then be estimated by maximizing likelihood in  $\beta$  and solving for MTD
- Writing down the likelihood requires a recursive conditioning
- **Important:** Properties of this are not as simple because sampling was *not random*

## Attempts to improve upon the standard design and analysis:

### Two-stage designs (Storer, 1989):

- Stage 1 – use very few patients to get to the dose region where the “action is” (close to MTD)
- Stage 2 – use a version of the standard design and find the MTD
- Allow dose de-escalation (go to lower dose) as well as escalation (“up and down” design)
- Use a statistical model, MLE to estimate MTD formally

## Two-stage designs (Storer, 1989):

- Stage 1:
  - ★ Single subject at a dose
  - ★ If no toxicity, escalate dose for next subject
  - ★ If toxicity, de-escalate dose for next subject
  - ★ Begin second stage at first toxicity
  
- Stage 2:
  - ★ 3 subjects at a dose
  - ★ If no toxicity escalate dose for next group
  - ★ If 1 toxicity, add 3 subjects at this dose
  - ★ If  $> 1$  toxicity, de-escalate dose for next group
  - ★ # groups fixed in advance

## Stochastic approximation methods (Anbar, 1984):

- *Estimate* the next dose sequentially

## Continual reassessment method (CRM, O'Quigley, Pepe, and Fisher, 1990):

- Bayesian approach – specify a prior distribution for the MTD and model for probability of toxicity
- First subject gets dose = prior value of the MTD
- After each subject, use Bayes' rule to update the posterior distribution of the MTD given the data so far
- Use the mode of this distribution (Bayesian “estimate”) as dose for next subject
- Stop after a fixed number of subjects and do a Bayesian analysis to estimate the MTD

## Disadvantages: Both approaches

- Require analysis after each subject
- May be too aggressive
- Require doses that may be difficult to prepare
- Modifications have been suggested; are area of current research and investigation

## Phase I Pharmacokinetic Studies

### Goals of drug therapy:

- Achieve therapeutic objective (cure disease, mitigate symptoms)
- Minimize toxicity
- Minimize difficulty of administration, optimize dose regimen to maximize therapeutic effect

**Pharmacology:** Understand the processes that allow these objectives to be achieved

## Implementation of Drug Therapy:

- Which drug?
- To whom?
- In what form?
- How much?
- How often?
- How long?

**Result:** In Phase I trials, where the drug is given to humans for the first time, it is sensible to begin to understand the pharmacology of a drug in humans

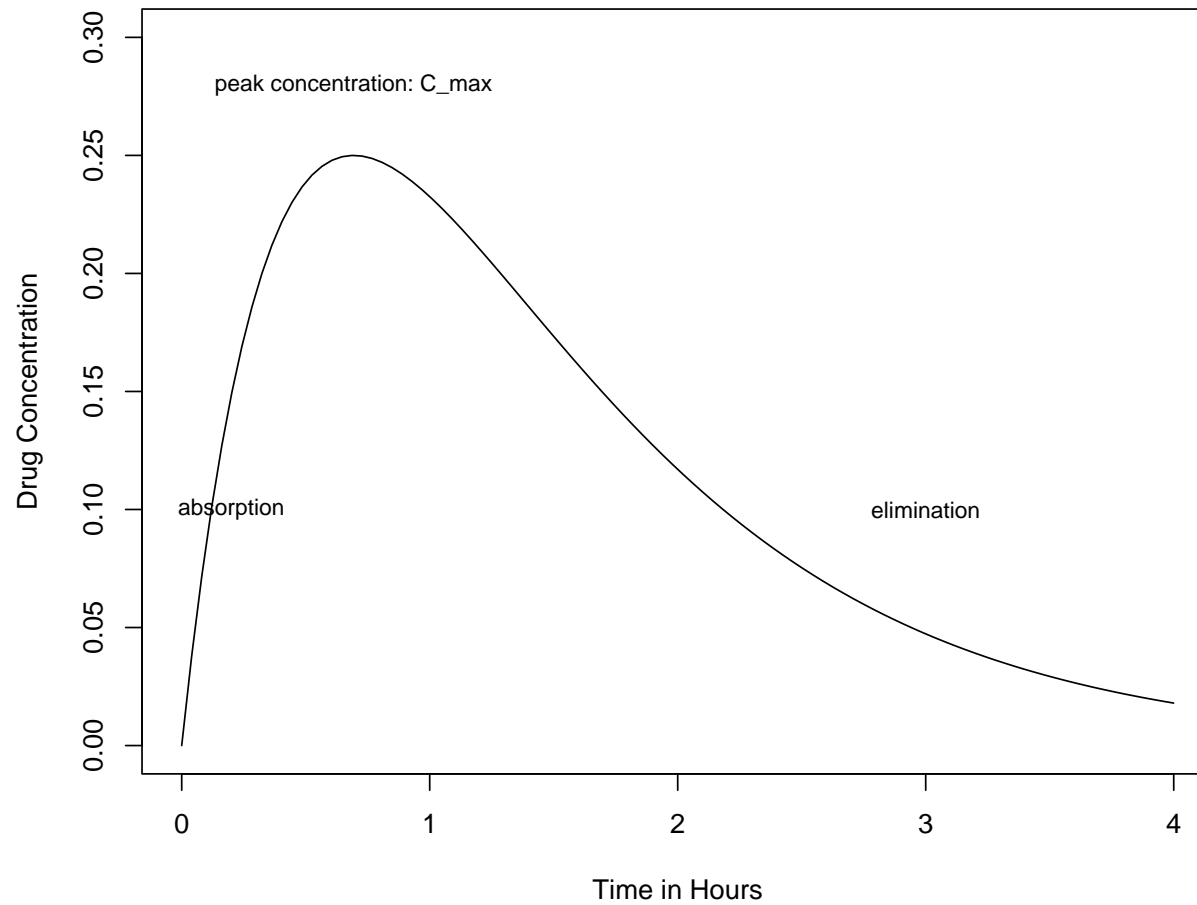
## Basic principles:

- There is a “*site(s) of action*” where drug produces effect(s)
- Magnitudes of desired response, toxicity are functions of drug concentration at site of action
- Drug can not be placed directly at site of action, but must be *absorbed* and then move:

[site of administration]  $\Rightarrow$  [site of action]

- Simultaneously, drug *distributes* to tissues, organs, and is *metabolized, eliminated*
- Drug concentration at site *cannot* be measured directly
- Drug concentration *can* be measured in blood/plasma
- Monitor concentration at site by concentration in blood/plasma

## Drug plasma concentration following oral administration (single dose) at $t = 0$



**Result:** Optimal administration of drugs requires

- Knowledge of mechanisms of ADME:
  - ★ Absorption
  - ★ Distribution
  - ★ Metabolism
  - ★ Elimination
- Understanding of the *kinetics* of these processes
  - ★ Motion of drug in the body over time

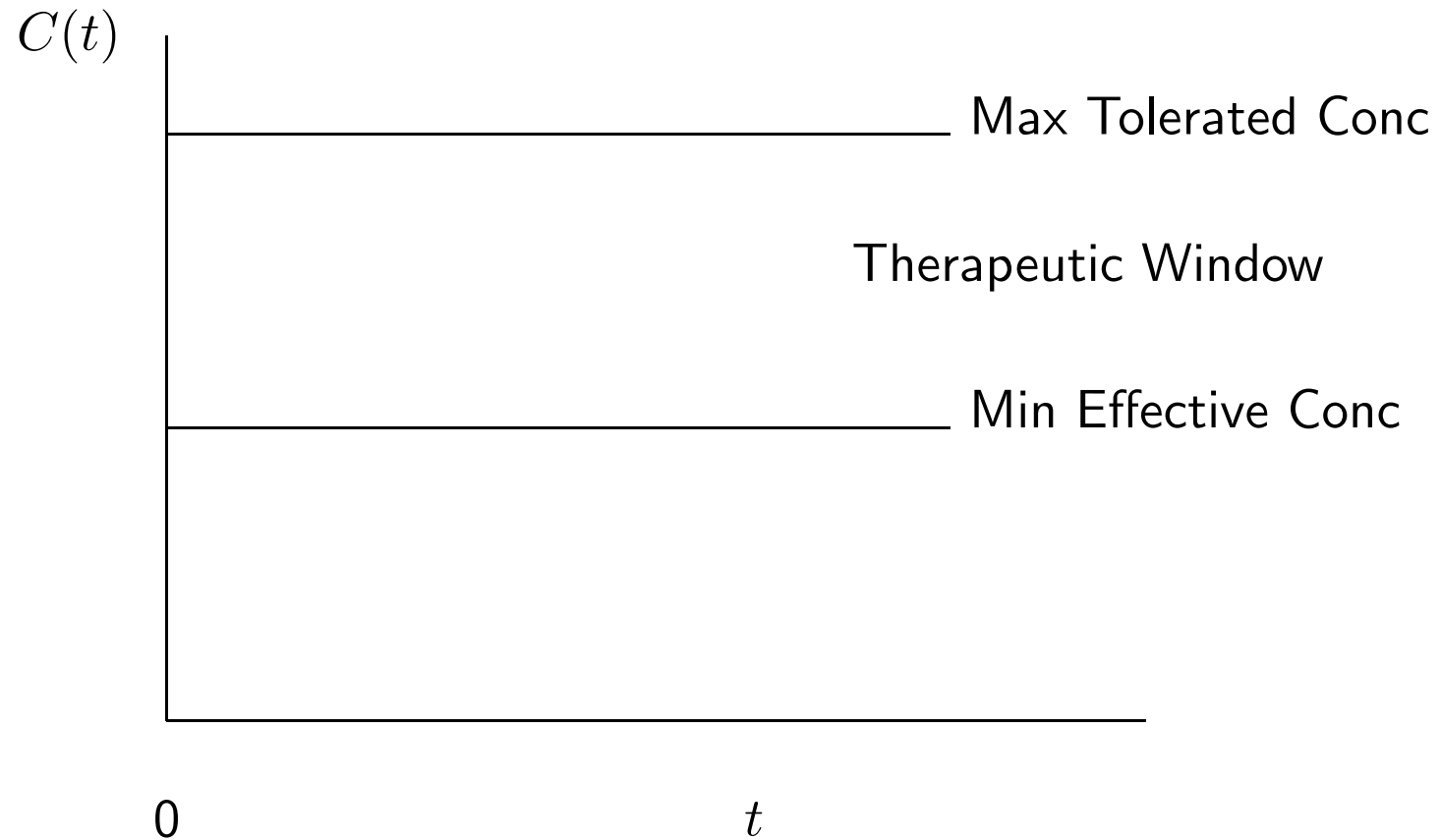
**Pharmacokinetics:** *“What the body does to the drug”*

- Drug concentration in blood/plasma used to monitor concentration at site of action over time
- Relationship between concentration and time/dose
- Kinetics of absorption, distribution, elimination

**Pharmacodynamics:** *“What the drug does to the body”*

- Relationship between drug concentration and pharmacologic response

## Therapeutic window



### Optimal dosage regimen:

- *Loading dose* to achieve concentration within therapeutic window quickly
- Replace drug eliminated by *sustaining doses* to maintain concentration within therapeutic window

### Maintenance dosing: Dose at discrete time intervals

- Route, frequency of administration, amount
- Achieve *steady-state* – amount lost = amount gained

### Governed by:

- Absorption, elimination, distribution, metabolism – *kinetics*
- Width of therapeutic window

## Pharmacokinetic studies:

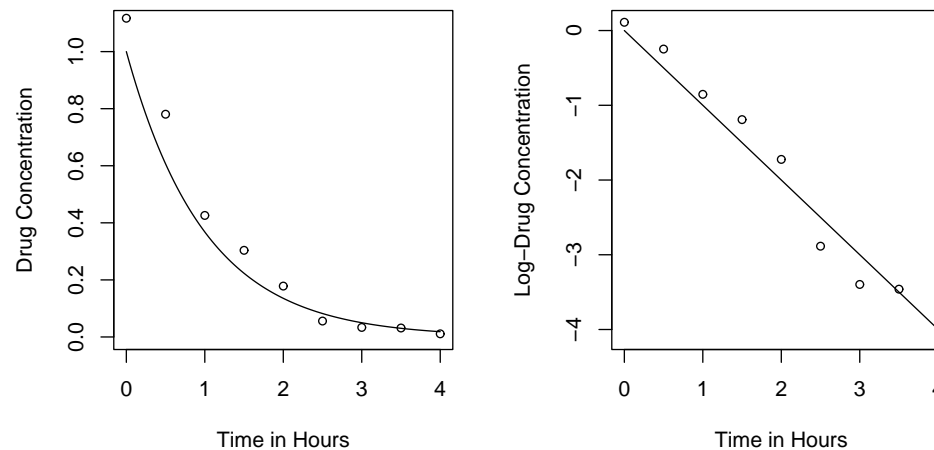
- Collect blood samples over time on each subject following dose
- Use concentration/time data to learn about subject-specific PK
- Use PK from several subjects to infer population behavior (mean, variability)
- The first studies are conducted in Phase I with a small number of subjects

## Pharmacokinetic models:

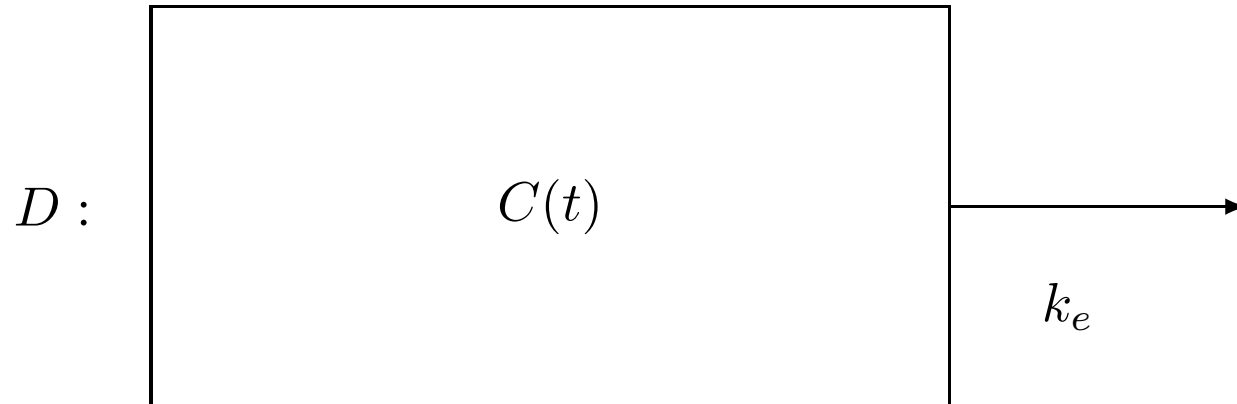
- Formalize notions of absorption, distribution, elimination for each subject within *mathematical* framework
- Describe concentration as a *function* of time/dose
- *Parameters* of function pertain to absorption, distribution, elimination
- *Compartmental models*: represent body as system of compartments, solution of differential equations
- Result: concentration as a *nonlinear* function of time/dose

## Example: Pharmacokinetics following IV administration

- Dose of drug given in rapid (instantaneous) bolus
- No absorption required



Suggests  $\log C(t) = \log C(0) - kt \Rightarrow C(t) = C(0) \exp(-kt)$

**I. One compartment open model: intravenous dose  $D$** 

$$\frac{dC(t)}{dt} = -k_e C(t), \quad C(0) = D/V$$

$$C(t) = (D/V) \exp(-k_e t), \quad k_e = Cl/V$$

## PK parameters:

- *Elimination rate constant* – Fractional rate of removal:

$$k_e = \frac{\text{Drug elimination rate}}{D(t)} (\text{hr}^{-1})$$

$$\left| \frac{dD(t)}{dt} \right| = \text{Drug elimination rate}$$

$D(t)$  = Amount of drug in the body

$$C(t) = \frac{D(t)}{V}$$

- *Clearance* – rate of elimination relative to concentration:

$$Cl = \frac{\text{Drug elimination rate (mg/hr)}}{C(t) \text{ (mg/L)}}$$

$Cl$  (L/hr) = volume of plasma that would be cleared of drug in one time unit

- (*Apparent*) *volume of distribution*  $V$  (L): Not a real physiological

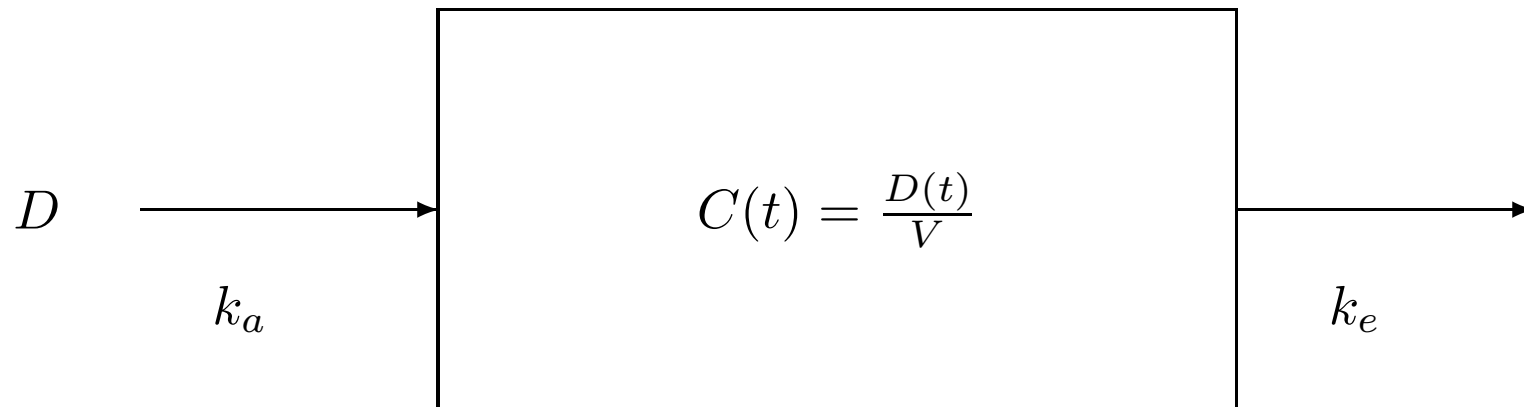
volume, but fluid volume that would be required to account for all drug in body

- *Elimination half-life* – Time for concentration to fall by half

$$t_{1/2} = \log(2)/k_e \text{ (hr)}$$

- **Note:** The above *PK* parameters  $k_e$ ,  $Cl$ ,  $t_{1/2}$  and  $V$  are all subject-specific. Given individual's concentration data, least squares or maximum likelihood method (assuming distribution for the data) can be used to estimate the *PK* parameters. These estimated parameters then can be used to make inference of interest.

## II. One compartment open model with first-order absorption: oral dose $D$



$$\frac{dD(t)}{dt} = k_a D_a(t) - k_e D(t), \quad D(0) = 0$$

$$\frac{dD_a(t)}{dt} = -k_a D_a(t), \quad D_a(0) = FD$$

$F$  = fraction available,  $D_a(t)$  = amount of drug at absorption site

$$D(t) = \frac{k_a DF}{k_a - k_e} \{ \exp(-k_e t) - \exp(-k_a t) \}, \quad k_e = Cl/V$$

$\implies$ 

$$C(t) = \frac{k_a DF}{V(k_a - k_e)} \{ \exp(-k_e t) - \exp(-k_a t) \}.$$

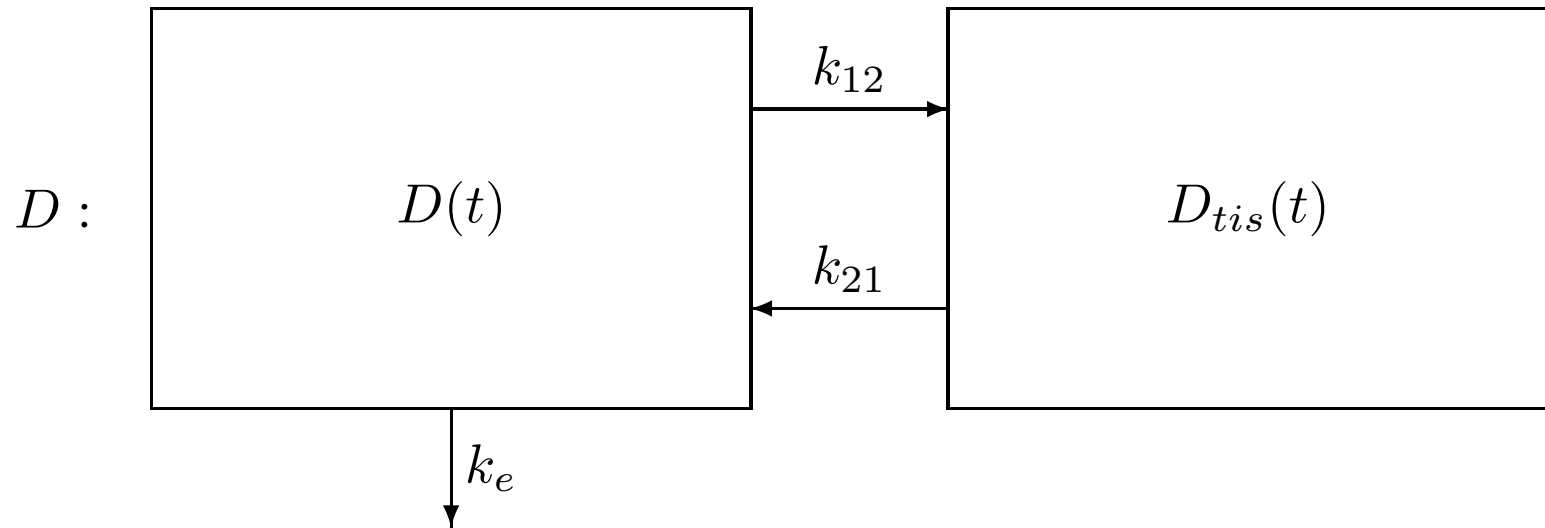
**Note:**

- $k_a > k_e$ . Otherwise, there will be no drug in the body
- Here the PK parameters include  $k_a, k_e, t_{1/2}, C_{max}$  and  $V$ .
- It can be shown that  $C(t)$  reaches its max at

$$t_{max} = \frac{\log(k_a) - \log(k_e)}{k_a - k_e}.$$

- $C_{max}$  is then given by  $C(t_{max})$ .

**Two-compartment open model with IV administration:** A second compartment is required to explain distribution of drug to various tissue groups



$$\frac{dD(t)}{dt} = k_{21}D_{tis}(t) - k_{12}D(t) - k_eD(t), \quad D(0) = D$$

$$\frac{dD_{tis}(t)}{dt} = k_{12}D(t) - k_{21}D_{tis}(t), \quad D_{tis}(0) = 0$$

Solving the above system leads to (bi-exponential)

$$D(t) = \tilde{A}_1 \exp(-\beta_1 t) + \tilde{A}_2 \exp(-\beta_2 t), \quad \beta_1 > \beta_2,$$

where

$$\beta_1 + \beta_2 = ke + k_{12} + k_{21}$$

$$\beta_1 \times \beta_2 = ke \times k_{21}$$

$$\tilde{A}_1 = \frac{D(\beta_1 - k_{21})}{\beta_1 - \beta_2}$$

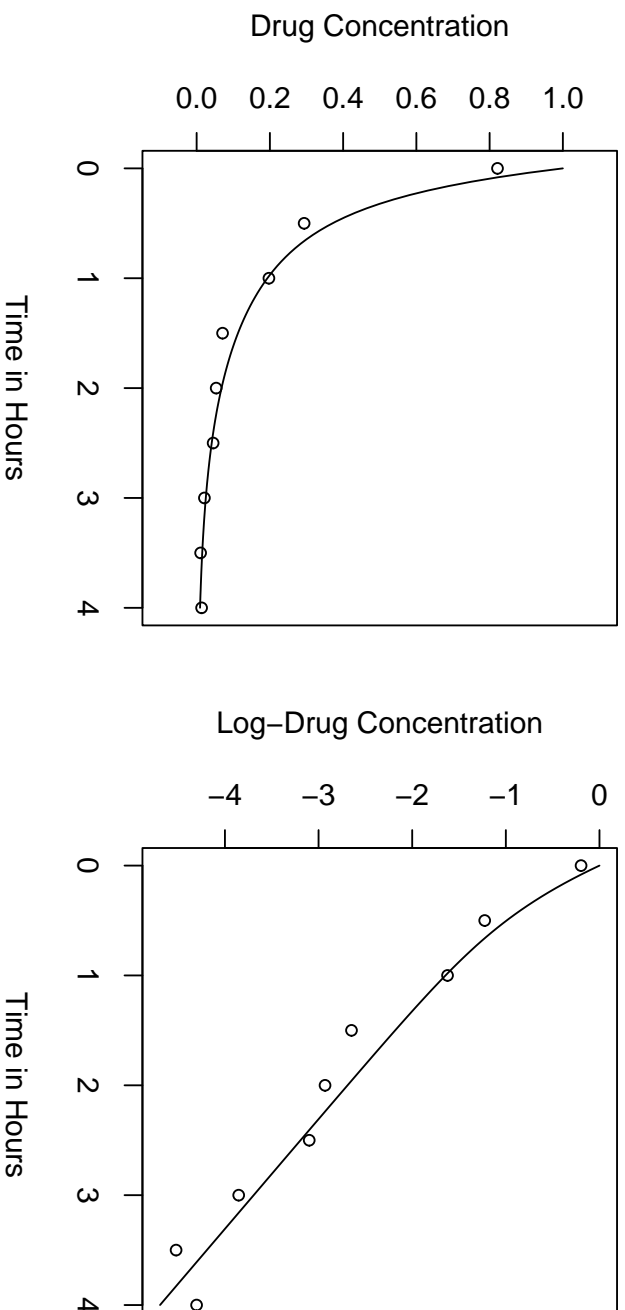
$$\tilde{A}_2 = \frac{D(k_{21} - \beta_2)}{\beta_1 - \beta_2}$$

Obviously,  $\beta_1, \beta_2 > 0$ . Hence,

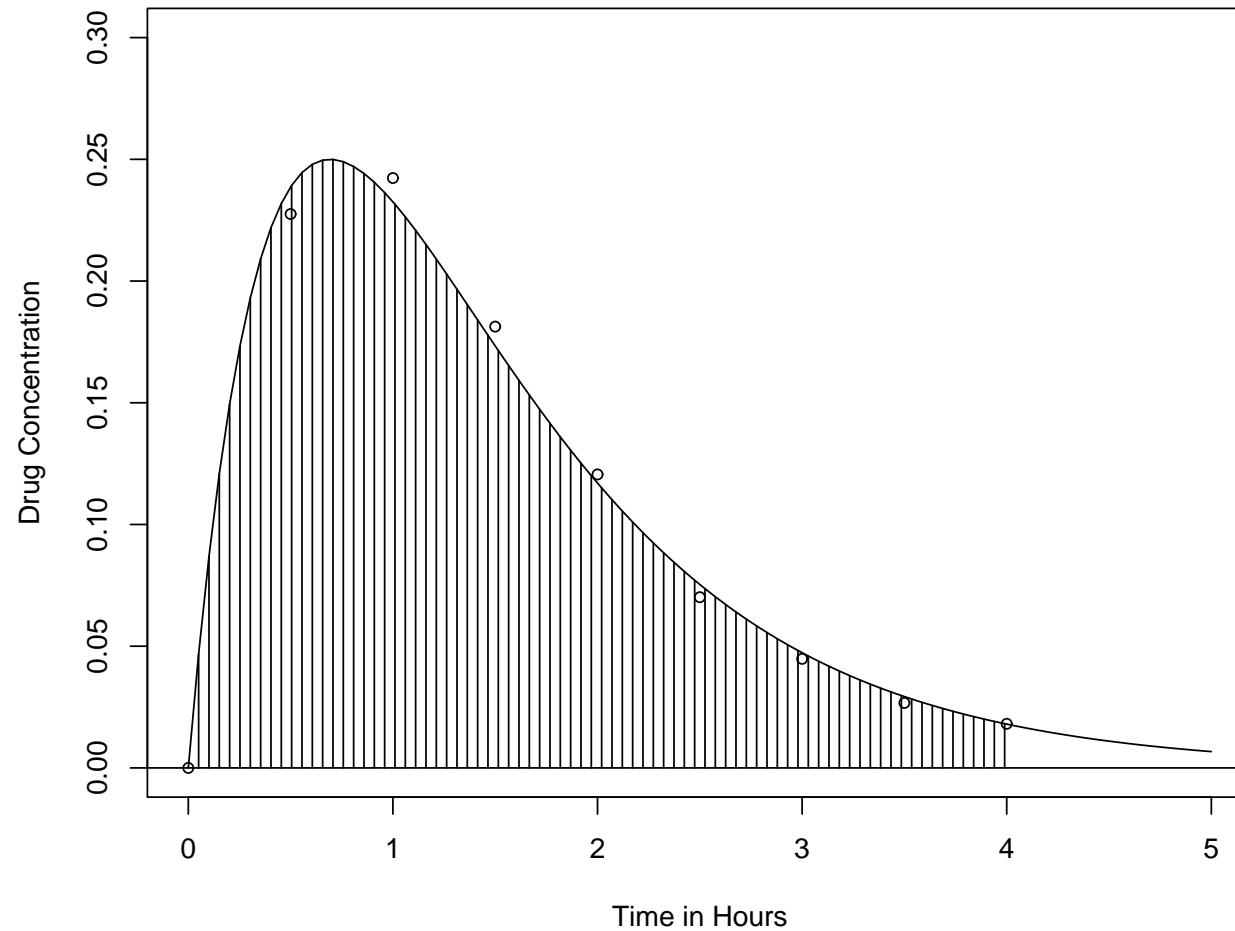
$$C(t) = A_1 \exp(-\beta_1 t) + A_2 \exp(-\beta_2 t), \quad \beta_1 > \beta_2,$$

with  $A_1 = \tilde{A}_1/V$  and  $A_2 = \tilde{A}_2/V$ .

## Graphical representation:



**Area Under the Curve  $AUC$ :** Reflects extent of drug availability and exposure following dose ( $AUC_{0-T}$  or  $AUC_{0-\infty}$ ).



## Pharmacokinetic analysis:

- Use measurements of drug concentrations on subjects to elucidate kinetics of drug absorption and disposition
- *Estimate* values of PK parameters in model for each subject
- *Estimate* typical values of PK parameters like  $Cl$ ,  $V$ ,  $k_a$ ,  $k_e$ ,  $AUC_{0-T}$ ,  $AUC_{0-\infty}$ ,  $t_{1/2}$ ,  $C_{max}$  from all subjects' data (mean, variance)
- Explore association between PK parameters and toxicity and other variables

## Implementation:

- *Weighted nonlinear regression* techniques – fit the PK model to concentration-time data for each subject
- With  $y_{ij}$  = concentration on subject  $i$  at time  $t_{ij}$  following dose  $D_i$

$$y_{ij} = f(t_{ij}, D_i, \beta_i) + e_{ij},$$

$\beta_i$  = PK parameters for subject  $i = 1, \dots, m$

- Errors  $e_{ij}$  with  $\text{var}(e_{ij}) = \sigma^2 f^2(t_{ij}, D_i, \beta_i)$  (constant CV)
- Use estimates  $\hat{\beta}_i$  from  $m$  subjects to estimate mean and variance in population
- *Nonlinear mixed effects models*

## Interaction Studies

**Drug-drug interaction:** Investigate the effect of a commonly used drug on some important *PK* parameters of the new drug.

$A$  = new drug,  $B$  = commonly used drug

- Cross-over design (when wash-out is possible): Some study subjects are randomly assigned to  $A + B$  and then to  $A$ ; other subjects are given  $A$  and then  $A + B$ . Compare important *PK* parameters of  $A$  under  $A + B$  to the *PK* parameters under  $A$ .
  - ★ Linear mixed model for formal statistical inference.
  - ★ Sample size calculation is usually based on precision.

- Parallel design (when wash-out is hard to achieve): Study subjects are randomly assigned to  $A + B$  or  $A$ . Compare important  $PK$  parameters of  $A$  under  $A + B$  to the  $PK$  parameters under  $A$ .
  - ★ Two-sample t-test can be used for formal statistical inference.
  - ★ Sample size calculation is usually based on precision.

**Food effect:** Investigate the food effect on some important  $PK$  parameters of the new drug.

- Design and analysis are similar to the drug-drug interaction study.

**Alcohol interaction:** Investigate the effect of the new drug administered together with alcohol/benzo over the effect of alcohol/benzo taken alone on some important pharmacodynamic (*PD*) parameters such as reaction time, short-term memory, ability to keep body balanced, etc.

- Cross-over design or parallel design
- Longitudinal data for each endpoint
- Mixed models for each endpoint
- Multivariate mixed models for multiple endpoints

## 2.2 Phase II Clinical Trials

**Phase II** clinical trials usually are conducted to assess

- feasibility of treatment
- side effects and toxicity
- logistics of administration and cost
- dose finding (lowest dose level with good efficacy)

**Major issue:** Is there enough evidence of efficacy of the new drug to move to phase III?

**Surrogate markers** are often used.

Usually, one-arm (no comparison)

**Example:** Suppose a new drug is developed for patients with lung cancer. Ultimately, we would like to know whether this drug will extend the life of lung cancer patients as compared to currently available treatments. Establishing the effect of a new drug on survival would require a long study with relatively large number of patients and thus may not be suitable as a screening mechanism. Instead, during phase II, the effect of the new drug may be assessed based on tumor shrinkage in the first few weeks of treatment. If the new drug shrinks tumors sufficiently for a sufficiently large proportion of patients, then this may be used as evidence for further testing.

In this example,

Overall (or disease-free) survival time = **clinical endpoint**

tumor shrinkage = **surrogate markers**

Other examples of surrogate markers are

- Lowering blood pressure or cholesterol for patients with heart disease
- Increasing CD4 counts or decreasing viral load for patients with HIV disease

## Caution in using surrogate markers:

- surrogacy is difficult to establish.
- If the new drug has no effect on the **surrogate markers**, it is probably more likely that the new drug will have no effect on the final clinical endpoint.
- However, sometimes it is possible that the new drug may have effect on the surrogate markers but have no effect on the final clinical endpoint.

## Statistical Issues and Methods:

**Goal:** estimate the effect of the new drug on some endpoint (a surrogate marker, safety endpoint, etc) with enough precision to decide whether we investigate the new drug in phase III.

### Examples:

- probability of a random patient responding to treatment (response has to be unambiguously defined)
- probability that a treated patient has side effects
- average decrease in blood pressure over a two week period

Consider a **binary** endpoint:

whether or not a patient responds to the new drug; whether or not a patient will have side effects, etc.

Suppose  $n$  patients are treated with the new drug:

$$X \sim \text{bin}(n, \pi)$$

- $X$  = total number of patients who respond to the new drug
- $\pi$  = population response rate (if the whole patient population is given the new drug.)

One objective: estimate  $\pi$  with enough precision.

## Properties of a binomial distribution:

- $E(X) = n\pi$ , where  $E(\cdot)$  denotes the expectation of a random variable.
- $Var(X) = n\pi(1 - \pi)$ , where  $Var(\cdot)$  denotes the variance of a random variable.
- $P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$ , where  $P(\cdot)$  denotes the probability of an event, and  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Denote the sample proportion by  $p = X/n$ , then
  - ★  $E(p) = \pi$
  - ★  $Var(p) = \pi(1 - \pi)/n$
- When  $n$  is sufficiently large, the distribution of the sample proportion  $p = X/n$  is well approximated by a normal distribution

with mean  $\pi$  and variance  $\pi(1 - \pi)/n$ .

$$p \sim N(\pi, \pi(1 - \pi)/n)$$

- A large sample  $(1 - \alpha)$  CI of  $\pi$  can be constructed as

$$p \pm z_{\alpha/2} \{p(1 - p)/n\}^{1/2}.$$

- A large sample 95% CI of  $\pi$ :

$$p \pm 1.96 \{p(1 - p)/n\}^{1/2}.$$

- Can be used to calculate sample size  $n$ .

**Example:** Suppose our best guess for the response rate of a new drug is about 35%; if we want the precision of our estimator to be such that the 95% confidence interval is within 15% of the true  $\pi$ , then we need

$$1.96 \left\{ \frac{(.35)(.65)}{n} \right\}^{1/2} = .15,$$

or

$$n = \frac{(1.96)^2 (.35)(.65)}{(.15)^2} = 39 \text{ patients.}$$

## Exact Confidence Intervals

If  $n\pi$  or  $n(1 - \pi)$  is small, then the normal approximation may not be accurate.

$\implies$  **Exact CI.**

**Definition:** A  $(1 - \alpha)$ -th confidence region (interval) for  $\pi$ :  $\mathcal{C}(k)$  ( $k =$  observed # of response) such that

$$P_{\pi}\{\mathcal{C}(X) \supset \pi\} \geq 1 - \alpha, \text{ for all } 0 \leq \pi \leq 1.$$

**Question:** How to find  $\mathcal{C}(k)$  for given  $k$  responses out of  $n$  patients?

Consider testing the following hypothesis:

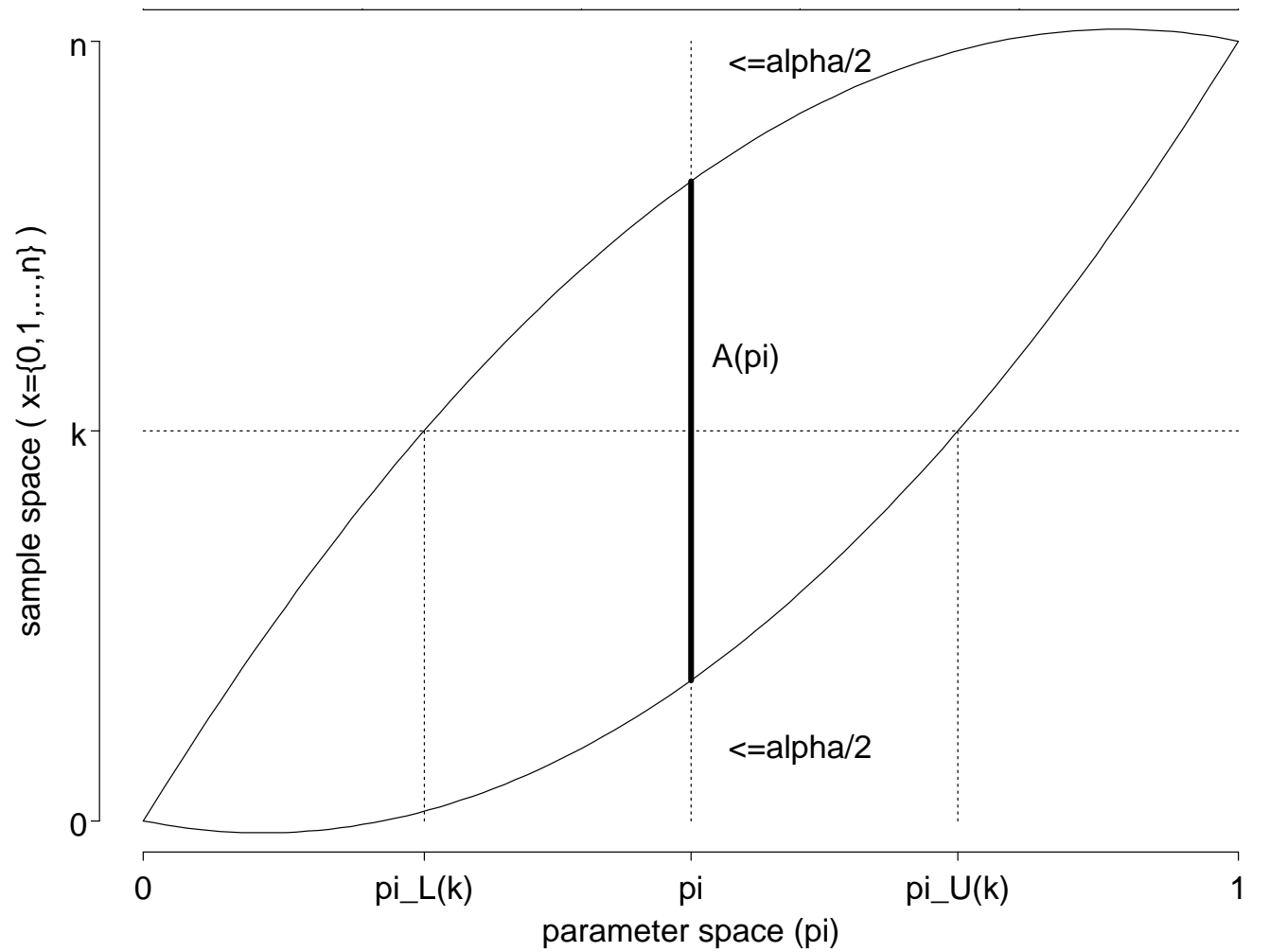
$$H_0 : \pi = \pi_0 \quad v.s. \quad H_a : \pi \neq \pi_0$$

for some  $\pi_0$ .

- Intuitively, we would reject  $H_0$  if  $k$  is too small or too large.
- Equivalently, we would accept  $H_0$  if  $k$  is neither too small nor too large; that is, there is an interval (or region)  $\mathcal{A}(\pi_0)$  (determined by  $\pi_0$ ) such that we would not reject (and hence accept)  $H_0$  if  $k \in \mathcal{A}(\pi_0)$ .
- $\mathcal{A}(\pi_0)$  is called the **acceptance region**.
- If we set the type I error probability of the above testing procedure at  $\alpha$ , then we have:

$$P[X \in \mathcal{A}(\pi_0) | H_0] \geq 1 - \alpha \quad \text{for } X \sim \text{bin}(n, \pi_0).$$

- For given observed  $k$ , solving  $k \in \mathcal{A}(\pi_0)$  will usually give us an interval  $\mathcal{C}(k) = [\pi_L(k), \pi_U(k)]$ , which is the exact  $(1 - \alpha)$  CI of  $\pi$ .
- Suppose we observe  $k$  responses out of  $n$  patients, then for any  $\pi_0 \in [\pi_L(k), \pi_U(k)]$ , we would not reject  $H_0$
- That is, the above CI consists of all values of  $\pi_0$  which is consistent with  $H_0$  given  $k$ .
- **Question:** How to find  $\pi_L(k)$  and  $\pi_U(k)$ ?

Figure 1: *Exact confidence intervals*

- The figure indicates that for given  $k$ ,  $\pi_L(k)$  and  $\pi_U(k)$  have to satisfy

$$P_{\pi_L(k)}[X \geq k] = \sum_{j=k}^n \binom{n}{j} \pi_L(k)^j \{1 - \pi_L(k)\}^{n-j} = \alpha/2,$$
$$P_{\pi_U(k)}[X \leq k] = \sum_{j=0}^k \binom{n}{j} \pi_U(k)^j \{1 - \pi_U(k)\}^{n-j} = \alpha/2.$$

- $\pi_L(k)$  and  $\pi_U(k)$  can be solved using binomial tables or through statistical software.
- When  $k = 0$ , the first equation has no solution. Set  $\pi_L(k) = 0$ .  $\pi_U(k)$  can be solved.
- When  $k = n$ , the second equation has no solution. Set  $\pi_U(k) = 1$ .  $\pi_L(k)$  can be solved.

- **Remark:** Since  $X$  has a discrete distribution, the way we define the  $(1 - \alpha)$ -th confidence interval above will yield

$$P_{\pi}\{\pi \in [\pi_L(k), \pi_U(k)]\} > 1 - \alpha$$

(strict inequality) for most values of  $0 \leq \pi \leq 1$ . Strict equality cannot be achieved because of the discreteness of the binomial random variable.

**Example:** Suppose in a Phase II clinical trial, 3 of 19 patients respond to  $\alpha$ -interferon treatment for multiple sclerosis.

- 95% CI of  $\pi$  based Normal approximation:

$$\frac{3}{19} \pm 1.96 \left( \frac{\frac{3}{19} \times \frac{16}{19}}{19} \right)^{1/2} = [-.006, .322].$$

- Exact 95% CI: need to find out  $\pi_L(3)$  and  $\pi_U(3)$  ( $n = 19, k = 3$ ) such that

$$P_{\pi_L(3)}(X \geq 3) = .025 \iff P_{\pi_L(3)}(X \leq 2) = .975$$

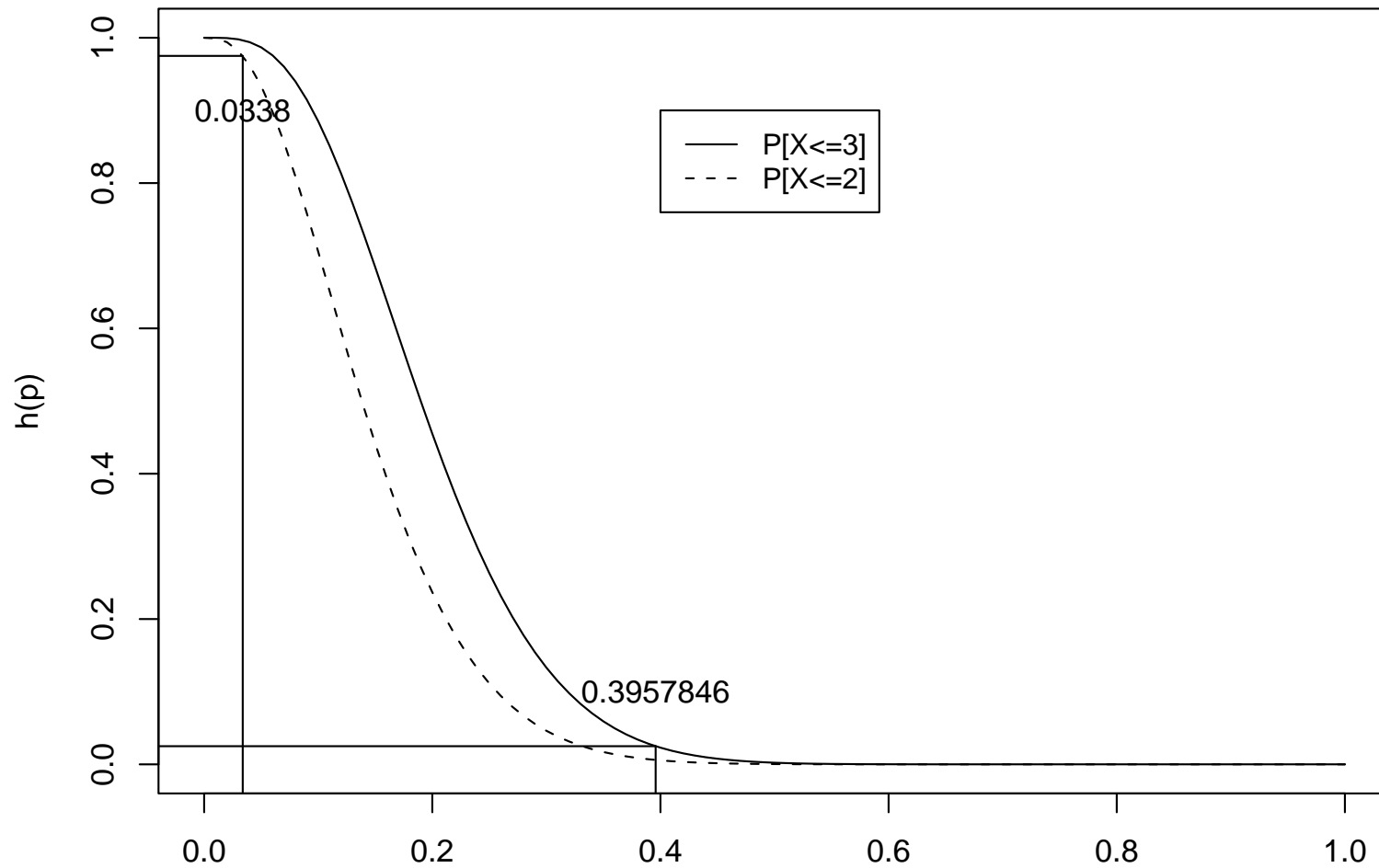
$$P_{\pi_U(3)}(X \leq 3) = .025.$$

$\implies$  (using binomial tables)

$$\pi_L(3) \approx .03, \quad \pi_U(3) \approx .40$$

Figure 2: *Find Exact CI*

$h(p) = P[X \leq 3]$  or  $h(p) = P[X \leq 2]$  for  $X \sim \text{Bin}(19, p)$



- *SAS* function for  $P[X \leq k]$  where  $X \sim \text{bin}(n, p)$ :  
`probbnml(p, n, k)`
- *R* function for  $P[X \leq k]$ :  
`pbinom(k, n, p)`

```
options ls=80 ps=200 nodate;

data binprob;
  do pi=0.35 to 0.45 by 0.01;
    prob = probbnml(pi, 19, 3);
    output;
  end;
run;

proc print data=binprob;
run;
```

Obs	pi	prob
1	0.35	0.059140
2	0.36	0.049483
3	0.37	0.041180
4	0.38	0.034083
5	0.39	0.028053
6	0.40	0.022959
7	0.41	0.018683
8	0.42	0.015115
9	0.43	0.012156
10	0.44	0.009717
11	0.45	0.007719

## Gehan's Two-Stage Design

One goal of a phase II trial is to discard ineffective treatments early.

**Gehan's Two-Stage Design** achieves this goal with 2 stages in a trial:

- Stage I: Give the new treatment to  $n_0$  patients. If no patient responds, declare the treatment ineffective.
- Stage II: If at least one patient responds in stage I, add  $n - n_0$  patients and count the total number of patients responding to the new treatment. Calculate point estimate of  $\pi$  and construct a CI for  $\pi$ .

## How to determine $n_0$ and $n$ ?

- Determine  $n_0$ : denote  $X = \#$  responses out of  $n_0$  patients.

$$P[X = 0] = (1 - \pi)^{n_0}.$$

- $\pi_0 =$  minimal efficacy; that is, if  $\pi \geq \pi_0$ , we want to investigate the new drug in phase III.
- We would like to control the probability of discarding the new drug early if in fact it is promising.
- That is,  $n_0$  has to satisfy:

$$P[X = 0] = (1 - \pi)^{n_0} \leq \alpha_0 \quad \text{for all } \pi \geq \pi_0,$$

where  $\alpha_0$  is our tolerance.

- Since  $P[X = 0] = (1 - \pi)^{n_0}$  is an decreasing function of  $\pi$ , only need

$$(1 - \pi_0)^{n_0} \leq \alpha_0.$$

$\implies$

$$n_0 \log(1 - \pi_0) \leq \log(\alpha_0)$$

$\implies$

$$n_0 \geq \frac{\log(\alpha_0)}{\log(1 - \pi_0)}$$

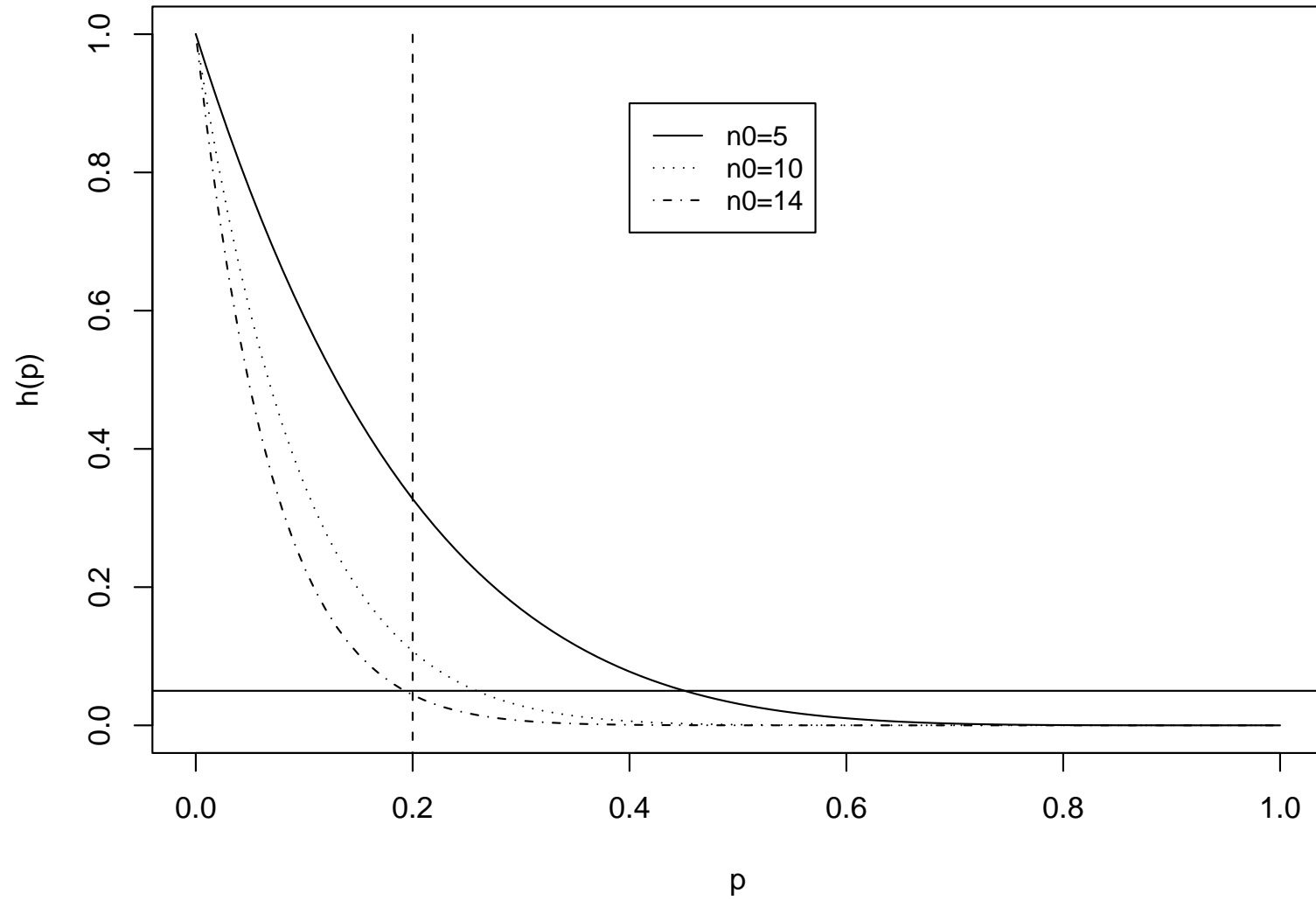
- For example,  $\pi_0 = 0.2$ ,  $\alpha_0 = 0.05$ , then

$$n_0 \geq \frac{\log(0.05)}{\log(1 - 0.2)} = 14(\text{round up}).$$

- Determine  $n$ : based on precision of 95% CI.
- For example, want to be 95% sure that the estimate is within  $\pm 15\%$  of the minimum  $\pi_0 = 0.2$ :

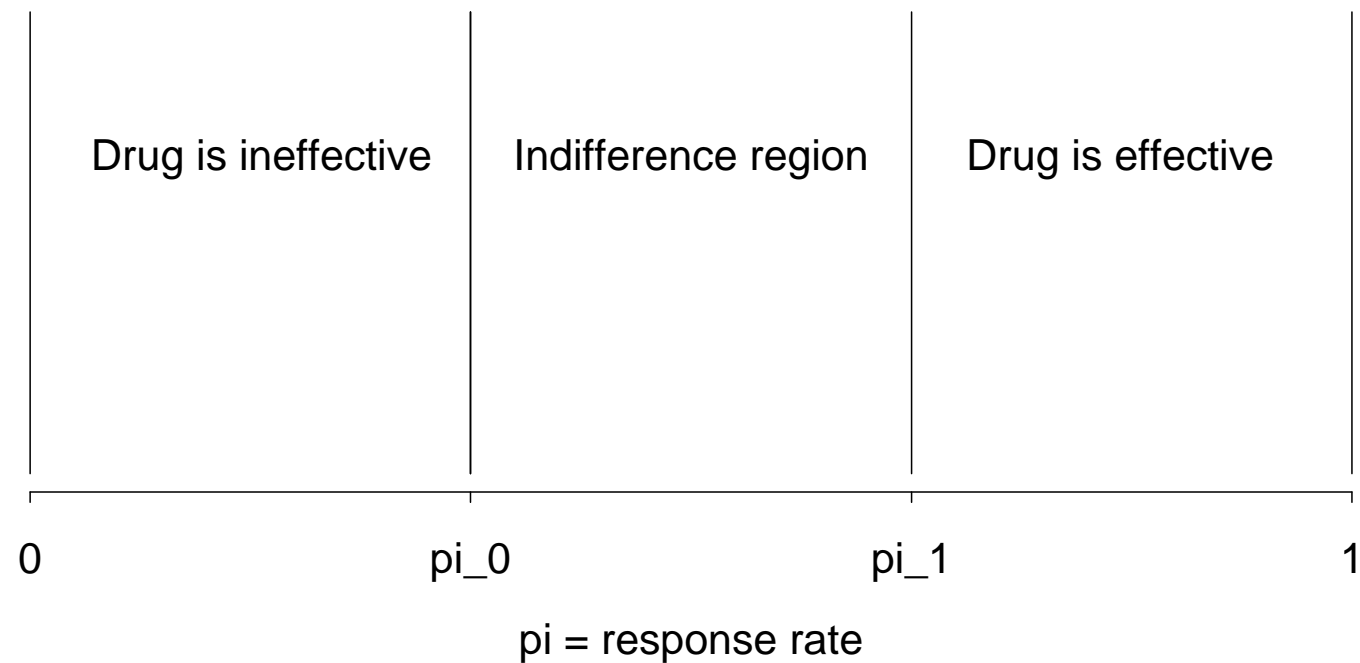
$$1.96 \left( \frac{.2 \times .8}{n} \right)^{1/2} = .15, \text{ or } n = 28.$$

## Probability of no reponse



## Simon's Two-Stage Design

Suppose two values  $\pi_0 < \pi_1$  are pre-specified such that



- If  $\pi \leq \pi_0$ , then we want to declare the drug ineffective with high probability, say  $1 - \alpha$ , where  $\alpha$  is taken to be small.
- If  $\pi \geq \pi_1$ , then we want to consider this drug for further investigation with high probability, say  $1 - \beta$ , where  $\beta$  is taken to be small.

The values of  $\alpha$  and  $\beta$  are generally taken to be between .05 and .20.

Simon's two-stage design proceeds as follows: Integers  $n_1$ ,  $n$ ,  $r_1$ ,  $r$ , with  $n_1 < n$ ,  $r_1 < n_1$ , and  $r < n$  are chosen (to be described later) and

- $n_1$  patients are given treatment in the first stage. If  $r_1$  or less respond, then declare the treatment a failure and stop.
- If more than  $r_1$  respond, then add  $(n - n_1)$  additional patients for a total of  $n$  patients.
- At the second stage, if the total number that respond among all  $n$  patients is greater than  $r$ , then declare the treatment a success; otherwise, declare it a failure.
- Of course, we stop the trial at stage 1 if the number of responses among  $n_1$  patients is greater than  $r$  and declare the treatment a success.

$X_1$  = the number of responses in stage 1 (out of  $n_1$  patients)

$X_2$  = the number of responses in stage 2 (out of  $n_2 = n - n_1$  patients)

$$X_1 \sim b(n_1, \pi), \quad X_2 \sim b(n_2, \pi) \quad (X_1 \text{ and } X_2 \text{ are ind}).$$

- Declare the new drug a failure if

$$(X_1 \leq r_1) \text{ or } \{(X_1 > r_1) \text{ and } (X_1 + X_2 \leq r)\}$$

- The new drug is declared a success if

$$\{(X_1 > r_1) \text{ and } (X_1 + X_2) > r\}.$$

- Design constraints  $\implies$

$$P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)\} \leq \alpha \text{ for all } \pi \leq \pi_0 \quad (2.1)$$

$$P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)\} \geq 1 - \beta \text{ for all } \pi \geq \pi_1 \quad (2.2)$$

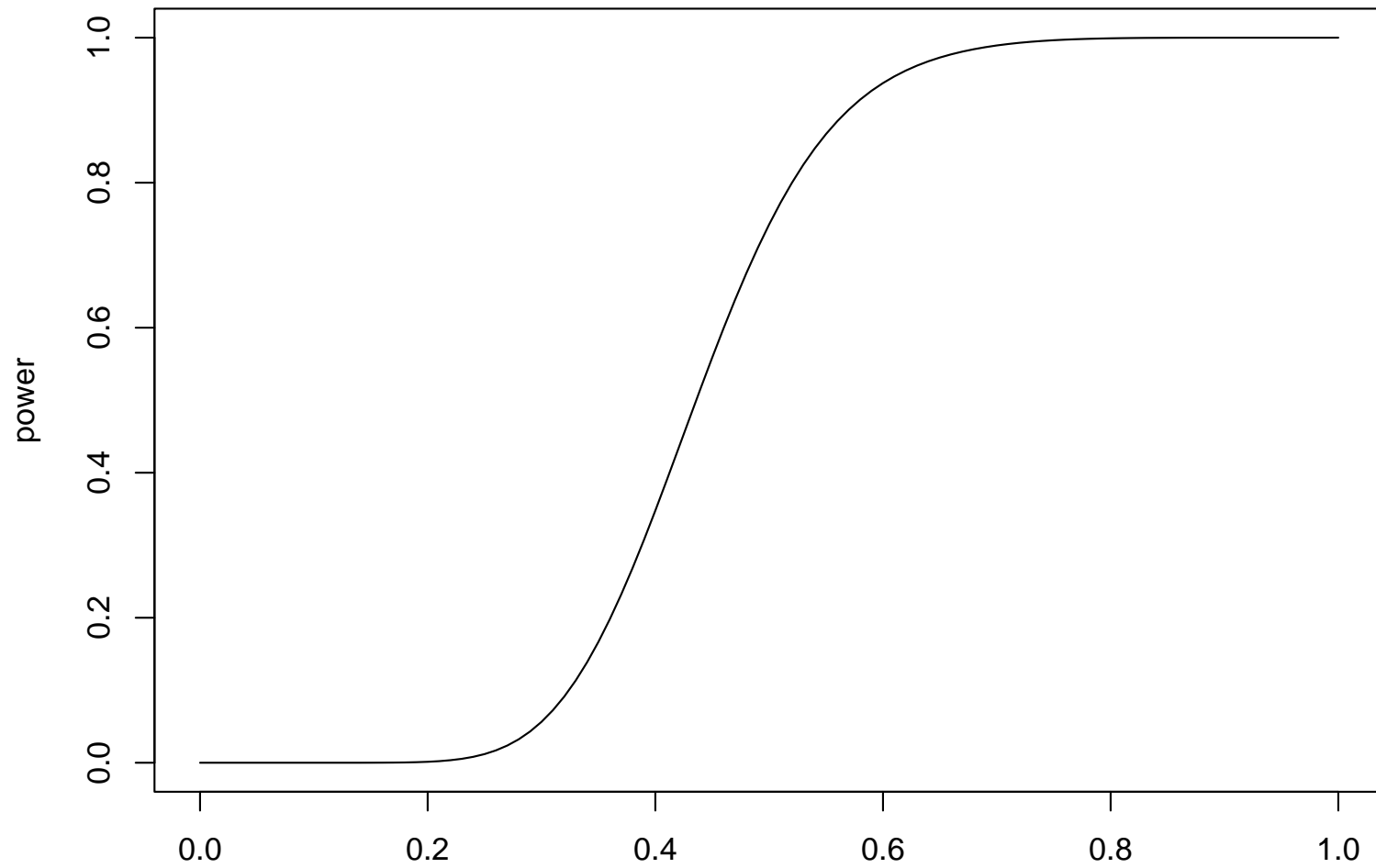
- Denote the power function by

$$h(\pi) = P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)|\pi\}.$$

- It can be shown that  $h(\pi)$  is an increasing function of  $\pi$  for any  $n_1, r_1, n, r$ .
- Therefore, criteria (2.1) and (2.2) are equivalent to

$$P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)|\pi = \pi_0\} = \alpha$$

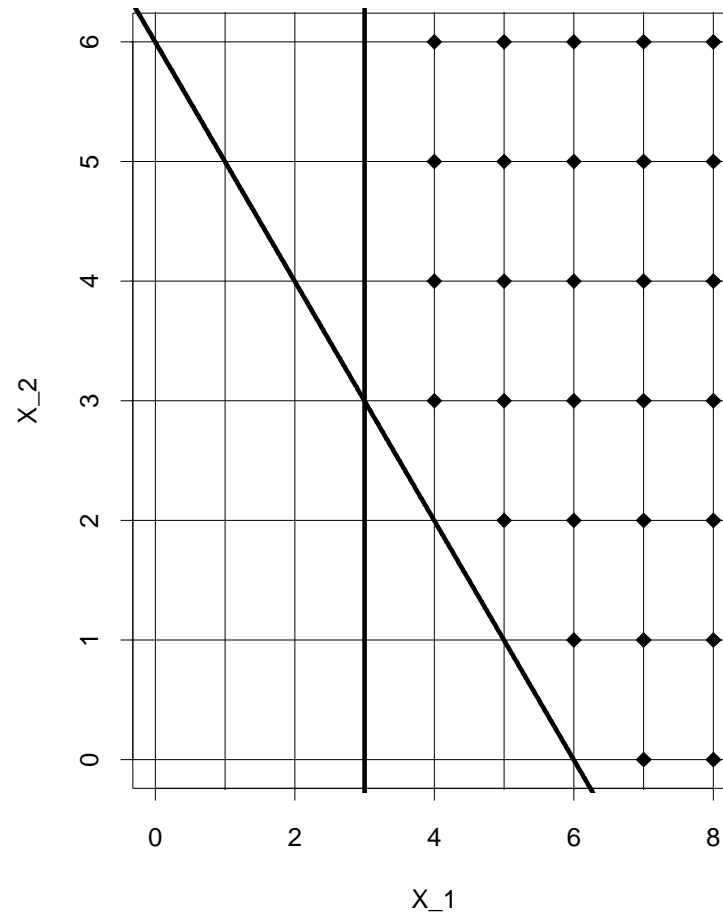
$$P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)|\pi = \pi_1\} = 1 - \beta.$$

Figure 3: *Power Function of Simon's Design*

- How to calculate  $P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)|\pi\}$  for each  $\pi$ ?
- By independence of  $X_1$  and  $X_2$ :

$$\begin{aligned} & P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)|\pi\} \\ = & \sum_{m_1 > r_1, m_1 + m_2 > r} P[X_1 = m_1, X_2 = m_2] \\ = & \sum_{m_1 > r_1, m_1 + m_2 > r} P[X_1 = m_1]P[X_2 = m_2] \end{aligned}$$

Figure 4: *Example:  $n_1 = 8$ ,  $n = 14$ ,  $X_1 > 3$ , and  $X_1 + X_2 > 6$*



- Many combinations of  $(r_1, n_1, r, n)$  satisfy (2.1) and (2.2).
- “**Optimal design**” is the one that has smallest expected sample size when  $\pi = \pi_0$  (when the new drug is ineffective)
- The expected sample size when  $\pi = \pi_0$ :

$$\begin{aligned} & n_1 P(\text{stopping at stage 1}) + n P(\text{did not stop at stage 1}) \\ = & n_1 \{P(X_1 \leq r_1 | \pi = \pi_0) + P(X_1 > r | \pi = \pi_0)\} \\ & + n P(r_1 + 1 \leq X_1 \leq r | \pi = \pi_0) \end{aligned}$$

- Through computer search, the **optimal design** can be identified.
- Simon, R. (1989). Optimal two-stage designs for Phase II clinical trials. *Controlled Clinical Trials*. 10: 1-10.
- Software can be downloaded from Dr. Simon’s website (see class website)

**Table 1** Designs for  $p_1 - p_0 = 0.20^a$ 

$p_0$	$p_1$	Optimal Design				Minimax Design			
		Reject Drug if Response Rate $\leq r/n$		EN( $p_0$ )	PET( $p_0$ )	Reject Drug if Response Rate $\leq r/n$		EN( $p_0$ )	PET( $p_0$ )
		$\leq r_1/n_1$	$\leq r/n$			$\leq r_1/n_1$	$\leq r/n$		
0.05	0.25	0/9	2/24	14.5	0.63	0/13	2/20	16.4	0.51
		0/9	2/17	12.0	0.63	0/12	2/16	13.8	0.54
		0/9	3/30	16.8	0.63	0/15	3/25	20.4	0.46
0.10	0.30	1/12	5/35	19.8	0.65	1/16	4/25	20.4	0.51
		1/10	5/29	15.0	0.74	1/15	5/25	19.5	0.55
		2/18	6/35	22.5	0.71	2/22	6/33	26.2	0.62
0.20	0.40	3/17	10/37	26.0	0.55	3/19	10/36	28.3	0.46
		3/13	12/43	20.6	0.75	4/18	10/33	22.3	0.50
		4/19	15/54	30.4	0.67	5/24	13/45	31.2	0.66
0.30	0.50	7/22	17/46	29.9	0.67	7/28	15/39	35.0	0.36
		5/15	18/46	23.6	0.72	6/19	16/39	25.7	0.48
		8/24	24/63	34.7	0.73	7/24	21/53	36.6	0.56
0.40	0.60	7/18	22/46	30.2	0.56	11/28	20/41	33.8	0.55
		7/16	23/46	24.5	0.72	17/34	20/39	34.4	0.91
		11/25	32/66	36.0	0.73	12/29	27/54	38.1	0.64
0.50	0.70	11/21	26/45	29.0	0.67	11/23	23/39	31.0	0.50
		8/15	26/43	23.5	0.70	12/23	23/37	27.7	0.66
		13/24	36/61	34.0	0.73	14/27	32/53	36.1	0.65
0.60	0.80	6/11	26/38	25.4	0.47	18/27	24/35	28.5	0.82
		7/11	30/43	20.5	0.70	8/13	25/35	20.8	0.65
		12/19	37/53	29.5	0.69	15/26	32/45	35.9	0.48
0.70	0.90	6/9	22/28	17.8	0.54	11/16	20/25	20.1	0.55
		4/6	22/27	14.8	0.58	19/23	21/26	23.2	0.95
		11/15	29/36	21.2	0.70	13/18	26/32	22.7	0.67

<sup>a</sup>For each value of  $(p_0, p_1)$ , designs are given for three sets of error probabilities  $(\alpha, \beta)$ . The first, second and third rows correspond to error probability limits  $(0.10, 0.10)$ ,  $(0.05, 0.20)$ , and  $(0.05, 0.10)$  respectively. For each design, EN( $p_0$ ) and PET( $p_0$ ) denote the expected sample size and the probability of early termination when the true response probability is  $p_0$ .

Table 2 Designs for  $p_1 - p_0 = 0.15^a$ 

$p_0$	$p_1$	Optimal Design				Minimax Design			
		Reject Drug if Response Rate		EN( $p_0$ )	PET( $p_0$ )	Reject Drug if Response Rate		EN( $p_0$ )	PET( $p_0$ )
		$\leq r_1/n_1$	$\leq r/n$			$\leq r_1/n_1$	$\leq r/n$		
0.05	0.20	0/12	3/37	23.5	0.54	0/18	3/32	26.4	0.40
		0/10	3/29	17.6	0.60	0/13	3/27	19.8	0.51
		1/21	4/41	26.7	0.72	1/29	4/38	32.9	0.57
0.10	0.25	2/21	7/50	31.2	0.65	2/27	6/40	33.7	0.48
		2/18	7/43	24.7	0.73	2/22	7/40	28.8	0.62
		2/21	10/66	36.8	0.65	3/31	9/55	40.0	0.62
0.20	0.35	5/27	16/63	43.6	0.54	6/33	15/58	45.5	0.50
		5/22	19/72	35.4	0.73	6/31	15/53	40.4	0.57
		8/37	22/83	51.4	0.69	8/42	21/77	58.4	0.53
0.30	0.45	9/30	29/82	51.4	0.59	16/50	25/69	56.0	0.68
		9/27	30/81	41.7	0.73	16/46	25/65	49.6	0.81
		13/40	40/110	60.8	0.70	27/77	33/88	78.5	0.86
0.40	0.55	16/38	40/88	54.5	0.67	18/45	34/73	57.2	0.56
		11/26	40/84	44.9	0.67	28/59	34/70	60.1	0.90
		19/45	49/104	64.0	0.68	24/62	45/94	78.9	0.47
0.50	0.65	18/35	47/84	53.0	0.63	19/40	41/72	58.0	0.44
		15/28	48/83	43.7	0.71	39/66	40/68	66.1	0.95
		22/42	60/105	62.3	0.68	28/57	54/93	75.0	0.50
0.60	0.75	21/34	47/71	47.1	0.65	25/43	43/64	54.4	0.46
		17/27	46/67	39.4	0.69	18/30	43/62	43.8	0.57
		21/34	64/95	55.6	0.65	48/72	57/84	73.2	0.90
0.70	0.85	14/20	45/59	36.2	0.58	15/22	40/52	36.8	0.51
		14/19	46/59	30.3	0.72	16/23	39/49	34.4	0.56
		18/25	61/79	43.4	0.66	33/44	53/68	48.5	0.81
0.80	0.95	5/7	27/31	20.8	0.42	5/7	27/31	20.8	0.42
		7/9	26/29	17.7	0.56	7/9	26/29	17.7	0.56
		16/19	37/42	24.4	0.76	31/35	35/40	35.3	0.94

# 3 Phase III Clinical Trials

Why are clinical trials needed?

Examples that show anecdotal information may be misleading:

- Blood-letting
- High oxygen concentration for premature infants has harm
- Intermittent positive pressure breathing (very expensive procedure) has no major benefit for patients with chronic obstructive pulmonary disease (COPD).
- Laetrile (a drug extracted from grapefruit seeds) was rumored to be the wonder drug for Cancer patients. Clinical trials showed that Laetrile had no effect.

- The Cardiac Antiarrhythmia Suppression Trial (CAST) documented that commonly used antiarrhythmia drugs were harmful in patients with MI
- More recently, against common belief, it was shown that prolonged use of Hormone Replacement Therapy for women following menopause may have deleterious effects.

## Issues to consider before designing a clinical trial

David Sackett (Pioneer of evidence-based medicine) gives the following six prerequisites

1. The trial needs to be done
  - (a) the disease must have either high incidence and/or serious course and poor prognosis
  - (b) existing treatment must be unavailable or somehow lacking
  - (c) The intervention must have promise of efficacy (pre-clinical as well as phase I-II evidence)
2. The trial question posed must be appropriate and unambiguous
3. The trial architecture is valid. **Random allocation** is one of the best ways that treatment comparisons made in the trial are valid. Other methods such as **blinding** and **placebos** should be considered when appropriate

4. The inclusion/exclusion criteria should strike a balance between efficiency and generalizability. Entering patients at high risk who are believed to have the best chance of response will result in an efficient study. This subset may however represent only a small segment of the population of individuals with disease that the treatment is intended for and thus reduce the study's generalizability
5. The trial **protocol** is feasible
  - (a) The protocol must be attractive to potential investigators
  - (b) Appropriate types and numbers of patients must be available
6. The trial administration is effective.

## Other issues that also need to be considered

- **Applicability:** Is the intervention likely to be implemented in practice?
- **Expected size of effect:** Is the intervention “strong enough” to have a good chance of producing a detectable effect?
- **Obsolescence:** Will changes in patient management render the results of a trial obsolete before they are available?

## Objectives and Outcome Assessment

- Primary objective: What is the primary question to be answered?
  - ★ ideally just one
  - ★ important, relevant to care of future patients
  - ★ capable of being answered
- Primary outcome (endpoint)
  - ★ ideally just one
  - ★ relatively simple to analyze and report
  - ★ should be well defined; objective measurement is preferred to a subjective one. For example, clinical and laboratory measurements are more objective than say clinical and patient impression

- Secondary Questions
  - ★ other outcomes or endpoints of interest
  - ★ subgroup analyses
  - ★ secondary questions should be viewed as exploratory
    - \* trial may lack power to address them
    - \* multiple comparisons will increase the chance of finding “statistically significant” differences even if there is no effect
  - ★ avoid excessive evaluations; as well as problem with multiple comparisons, this may effect data quality and patient support

## Choice of Primary Endpoint

**Example:** new treatment for HIV patients; attacks HIV virus.

1. Increase in CD4 count.
2. Viral RNA reduction. Measures the amount of virus in the body
3. Time to the first opportunistic infection
4. Time to death from any cause
5. Time to death or first opportunistic infection, whichever comes first
  - Outcomes 1 & 2 are good for a phase II trial.
  - Outcome 4 is the ultimate endpoint; may lead to obsolescence.
  - Outcome 3 is not complete.
  - Outcome 5 is good one in a phase III trial.

## Ethical Issues

- No alternative which is superior to any trial intervention is available for each subject
- Equipoise—There should be genuine uncertainty about which trial intervention may be superior for each individual subject before a physician is willing to allow their patient to participate in such a trial
- Exclude patients for whom risk/benefit ratio is likely to be unfavorable
  - ★ pregnant women if possibility of harmful effect to the fetus
  - ★ too sick to benefit
  - ★ if prognosis is good without interventions

## Justice Considerations

- Should not exclude a class of patients for non medical reasons nor unfairly recruit patients from poorer or less educated groups

This last issue is a bit tricky as “equal access” may hamper the evaluation of interventions. For example

- Elderly people may die from diseases other than that being studied
- IV drug users are more difficult to follow in AIDS clinical trials

## The Randomized Clinical Trial

**Goal:** to evaluate a new treatment  $\implies$  compare the new treatment to

- no intervention
- placebo
- best available therapy

### Fundamental Principle in Comparing Treatment Groups

Groups must be alike in all important aspects and only differ in the treatment which each group receives. Otherwise, differences in response between the groups may not be due to the treatments under study, but can be attributed to the particular characteristics of the groups.

## How should the control group be chosen

Here are some examples:

- Literature controls
- Historical controls
- Patient as his/her own control (cross-over design)
- Concurrent control (non-randomized)
- Randomized concurrent control

*Results of Rapid Injection of 5-FU for Treatment of Advanced Carcinoma of the Large Bowel*

---

Group	# of Patients	# of Response	% of Response
1. Sharp and Benefiel	13	11	85
2. Rochlin et al.	47	26	55
3. Cornell et al.	13	6	46
4. Field	37	15	41
5. Weiss and Jackson	37	13	35
6. Hurley	150	46	31
7. ECOG	48	13	27
8. Brennan et al.	183	42	23
9. Ansfield	141	24	17
10. Ellison	87	10	12
11. Knoepp et al.	11	1	9
12. Olson and Greene	12	1	8

---

Suppose there is a new treatment for *advanced carcinoma of the large bowel*.

**Question:** *How should the new treatment be compared to 5-FU using the historical controls?*

- If there is no study-to-study variation, we should pool the studies and compare the new treatment to the pooled result.
- What if there is study-to-study variation? What should we do?

The experiment can be represented by

$$(n_i, X_i, \pi_i) \quad i = 1, 2, \dots, N = 12.$$

where

$n_i$  = sample size of the  $i$ th study

$X_i$  = # of responses from the  $i$ th study

$\pi_i$  = true response rate for the  $i$ th study population

- If there is no study-to-study variation,  $\pi_1 = \pi_2 = \dots = \pi_N = \pi$ , and

$$X_i | n_i \sim \text{bin}(n_i, \pi)$$

Then the best estimate of  $\pi$  is

$$\hat{\pi} = \frac{\sum_{i=1}^N X_i}{\sum_{i=1}^N n_i} = 0.267 \quad (\text{SE} = 0.016)$$

and the new treatment can (and should) be compared to  $\hat{\pi}$ .

- If there is study-to-study variation, we may assume  $\pi_i$  is a random variable with  $E(\pi_i) = \mu_\pi$  and  $\text{var}(\pi_i) = \sigma_\pi^2$ . Then we should compare the new treatment response rate to  $\mu_\pi$ .
- In this case, we have

$$X_i | n_i, \pi_i \sim \text{bin}(n_i, \pi_i)$$

- **Question:** How to estimate  $\sigma_\pi^2$  (and  $\mu_\pi$ )?
- Double expectation theorem:

$$E(X) = E[E(X|Y)]$$

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}[E(X|Y)]$$

For any two random variables (vectors)  $X$  and  $Y$ .

- Denote sample proportion for the  $i$ th study by

$$p_i = X_i/n_i,$$

an estimate of  $\pi_i$ .

- Marginal variation of  $p_i$  contains binomial sampling variation and variation ( $\sigma_\pi^2$ ) among  $\pi_i$ :

$$\text{Var}(p_i) = \text{E}[\text{Var}(p_i|n_i, \pi_i)] + \text{Var}[\text{E}(p_i|n_i, \pi_i)].$$

Since

$$\text{E}(p_i|n_i, \pi_i) = \pi_i,$$

$$\text{Var}(p_i|n_i, \pi_i) = \frac{\pi_i(1 - \pi_i)}{n_i},$$

so

$$\text{Var}(p_i) = \text{E} \left[ \frac{\pi_i(1 - \pi_i)}{n_i} \right] + \sigma_\pi^2.$$

- We can estimate  $\text{Var}(p_i)$  by its sample variance

$$S_p^2 = \frac{1}{N-1} \sum_{i=1}^N (p_i - \bar{p})^2$$

where

$$\bar{p} = \frac{p_1 + p_2 + \dots + p_N}{N}.$$

- If we can also estimate

$$\mathbb{E} \left[ \frac{\pi_i(1 - \pi_i)}{n_i} \right],$$

then we can estimate  $\sigma_\pi^2$ .

- If  $\pi_i$  could be observed, then

$$Z_i = \frac{\pi_i(1 - \pi_i)}{n_i}$$

could also be observed. The above expectation  $E(Z_i)$  could be estimated by its sample mean

$$\frac{Z_1 + Z_2 + \dots + Z_N}{N}.$$

- We would like to find an observable r.v.  $Q_i$  such that  $E(Q_i) = E(Z_i)$ .

- It can be shown that we can take  $Q_i$  as

$$Q_i = \frac{p_i(1 - p_i)}{n_i - 1} :$$

$$\mathbf{E}(Q_i) = \mathbf{E} \left[ \mathbf{E} \left( \frac{p_i(1 - p_i)}{n_i - 1} \middle| n_i, \pi_i \right) \right],$$

and

$$\mathbf{E}[p_i(1 - p_i) | n_i, \pi_i] = \frac{n_i - 1}{n_i} \pi_i(1 - \pi_i).$$

- Therefore,  $\mathbf{E}(Z_i) = \mathbf{E}(Q_i)$  can be estimated by

$$\frac{Q_1 + \dots + Q_N}{N} = \frac{1}{N} \sum_{i=1}^N \frac{p_i(1 - p_i)}{n_i - 1}.$$

- So an estimate of  $\sigma_\pi^2$  is given by

$$\hat{\sigma}_\pi^2 = S_p^2 - \frac{1}{N} \sum_{i=1}^N \frac{p_i(1-p_i)}{n_i-1}.$$

- For our example,

$$S_p^2 = \frac{1}{N-1} \sum_{i=1}^N (p_i - \bar{p})^2 = 0.0496$$

$$\frac{1}{N} \sum_{i=1}^N \frac{p_i(1-p_i)}{n_i-1} = 0.0061$$

$$\implies \hat{\sigma}_\pi^2 = 0.0435.$$

- $\hat{\sigma}_\pi^2$  is relatively large. The new study should be compared to

$$\hat{\mu}_\pi = \frac{p_1 + p_2 + \dots + p_N}{N} = 0.324 \quad \text{with SE} = S_p/\sqrt{N} = 0.064$$

if historical controls are used.

## When should historical data be used for treatment comparison?

- Compare treatment A to the best available treatment B for a disease.
- Population response rates are  $\mu_A$  and  $\mu_B$ .
- Treatment effect

$$\Delta = \mu_A - \mu_B.$$

- Assume  $\mu_B$  can be estimated very precisely.

There are  $n$  patients enrolled in the study.

1. Use historical data:

$$\hat{\Delta}_1 = \frac{X}{n} - \mu_B.$$

2. Use concurrent randomized controls (half patients receive A & B):

$$\hat{\Delta}_2 = \frac{X_1}{n/2} - \frac{X_2}{n/2}.$$

**Question:** Which estimate is better?

1. When historical data are used:

$$X|\pi_A \sim b(n, \pi_A), \quad \mathbf{E}(\pi_A) = \mu_A, \quad \text{var}(\pi_A) = \sigma_\pi^2,$$

where  $\pi_A$  is the true response rate of the subpopulation if treated with A from which  $n$  patients are sampled.

$$\mathbf{E}(\widehat{\Delta}_1) = \mathbf{E}\left(\frac{X}{n} - \mu_B\right) = \mathbf{E}(\pi_A) - \mu_B = \mu_A - \mu_B = \Delta,$$

$$\begin{aligned} \text{var}(\widehat{\Delta}_1) &= \text{var}\left(\frac{X}{n}\right) = \text{var}(\pi_A) + \mathbf{E}\left[\frac{\pi_A(1 - \pi_A)}{n}\right] \\ &= \sigma_\pi^2 + n^{-1}\mathbf{E}[\pi_A(1 - \pi_A)]. \end{aligned}$$

2. When concurrent randomized controls are used:

$$X_1|\pi_A \sim b(n/2, \pi_A), \quad X_2|\pi_B \sim b(n/2, \pi_B),$$

$\pi_B$  = true response rate of the subpopulation if treated with B.

Assume the same treatment effect in every subpopulation:

$$\pi_A - \pi_B = \Delta.$$

Then

$$\mathbf{E}(\widehat{\Delta}_2) = \mathbf{E}\left(\frac{X_1}{n/2} - \frac{X_2}{n/2}\right) = \mathbf{E}(\pi_A - \pi_B) = \Delta,$$

$$\begin{aligned} \text{var}(\widehat{\Delta}_2) &= \text{var}\left(\frac{X_1}{n/2} - \frac{X_2}{n/2}\right) \\ &= \text{var}(\Delta) + \mathbf{E}\left[\frac{\pi_A(1 - \pi_A)}{n/2} + \frac{\pi_B(1 - \pi_B)}{n/2}\right] \\ &\approx 4n^{-1}\mathbf{E}[\pi_B(1 - \pi_B)]. \end{aligned}$$

**Conclusion:**

- Both  $\hat{\Delta}_1$  and  $\hat{\Delta}_2$  are unbiased to  $\Delta$ .
- When  $\sigma_\pi^2 = 0$ , using historical data is always more efficient.
- When  $\sigma_\pi^2 > 0$ ,  $\hat{\Delta}_1$  will be more efficient than  $\hat{\Delta}_2 \iff$

$$\sigma_\pi^2 + n^{-1}\mathbb{E}[\pi_A(1 - \pi_A)] < 4n^{-1}\mathbb{E}[\pi_B(1 - \pi_B)]$$

$$\iff$$

$$n < \frac{3\mathbb{E}[\pi_B(1 - \pi_B)]}{\sigma_\pi^2}$$

Given historical data as in the example,  $\mathbb{E}[\pi_B(1 - \pi_B)]$  can be estimated by

$$\frac{1}{N} \sum_{i=1}^N \frac{n_i p_i (1 - p_i)}{n_i - 1}.$$

# 4 Randomization

## Advantage of Randomization:

- Eliminates conscious bias
  - ★ physician selection
  - ★ patient self selection
- Balances unconscious bias between treatment groups
  - ★ supportive care
  - ★ patient management
  - ★ patient evaluation
  - ★ unknown factors affecting outcome

- Groups are alike on average
  - ★ Allows us to make causal statement for the treatment effect
- Provides a basis for standard methods of statistical analysis such as significance tests

**Design-based Inference:** Randomization allows us to make *design-based* inference rather than *model-based* inference.

- Suppose we are comparing A to B.
- sharp null  $H_0$ : A & B are exactly the same for each patient
- The design allows us to test the above  $H_0$  without assuming a distribution of data

- For example, 4 patients in a clinical trial. Patients 1 & 2 received A, patients 3 & 4 received B. We would reject  $H_0$  if

$$T = \left( \frac{y_1 + y_2}{2} \right) - \left( \frac{y_3 + y_4}{2} \right)$$

is too large or too small.

- **Question:** How do we calculate the P-value of the above test for given observed  $T_{obs}$ ?
- Under the sharp null, the permutational distribution of  $T$  (induced by randomization) can be calculated.

Table 1: *Permutational distribution under sharp null*

patient	1	2	3	4	
response	$y_1$	$y_2$	$y_3$	$y_4$	Test statistic $T$
possible	A	A	B	B	$\left(\frac{y_1+y_2}{2}\right) - \left(\frac{y_3+y_4}{2}\right) = t_1$
treatment	A	B	A	B	$\left(\frac{y_1+y_3}{2}\right) - \left(\frac{y_2+y_4}{2}\right) = t_2$
assignments	A	B	B	A	$\left(\frac{y_1+y_4}{2}\right) - \left(\frac{y_2+y_3}{2}\right) = t_3$
each	B	A	A	B	$\left(\frac{y_2+y_3}{2}\right) - \left(\frac{y_1+y_4}{2}\right) = t_4$
equally	B	A	B	A	$\left(\frac{y_2+y_4}{2}\right) - \left(\frac{y_1+y_3}{2}\right) = t_5$
likely	B	B	A	A	$\left(\frac{y_3+y_4}{2}\right) - \left(\frac{y_1+y_2}{2}\right) = t_6$

The first one ( $t_1$ ) is the observed test statistic.

- Under sharp  $H_0$ ,  $T$  can take any of those 6 values with equal prob=1/6.

- One-sided P-value (the alternative is “A is better than B”):

$$P[T \geq t_1 | \text{sharp } H_0] = \frac{\# \text{ of } t_i \geq t_1}{6}.$$

- Two-sided P-value (the alternative is “A is different than B”):

$$P[|T| \geq |t_1| | \text{sharp } H_0] = \frac{\# \text{ of } |t_i| \geq |t_1|}{6}.$$

- When sample size gets large, the distribution looks like normal.
- **Example:** suppose A:  $y_1 = 8, y_2 = 4$ , B:  $y_3 = 6, y_4 = 2$ .  
P-values=?

- **Remark:** In the permutational distribution, we treat each individual's response as fixed. Randomness is induced by the treatment assignment mechanism.

**Statistical model:**

$$Y_1, Y_2 \text{ are iid } N(\mu_1, \sigma^2)$$

$$Y_3, Y_4 \text{ are iid } N(\mu_2, \sigma^2)$$

and we are testing the null hypothesis

$$H_0 : \mu_1 = \mu_2.$$

Reject  $H_0$  (in favor of  $H_a : \mu_1 > \mu_2$ ) if the observed

$$T = \frac{\bar{Y}_A - \bar{Y}_B}{s_p(n_A^{-1} + n_B^{-1})^{1/2}}$$

is too large.

- Under  $H_0$ ,  $T \sim t_{n_A+n_B-2}$ . The P-value of the above test:

$$\text{P-value} = P[t_{n_A+n_B-2} \geq T_{\text{obs}}].$$

- **Comment:** Distribution-free feature is nice, but the most important aspect of a randomized clinical trial is that it allows us to make **causal** inference. In observational studies such as epidemiological studies, we can only make **associational** statement.

## Disadvantages of Randomization

- Patients or physician may not care to participate in an experiment involving a chance mechanism to decide treatment
- May interfere with physician patient relationship
- Part of the resources are expended in the control group; i.e. If we had  $n$  patients eligible for a study and had good and reliable historical control data, then it could be more efficient to put all  $n$  patients on the new treatment and compare the response rate to the historical controls rather than randomizing the patients into two groups, say,  $n/2$  patients on new treatment and  $n/2$  on control treatment and then comparing the response rates among these two randomized groups.

## How Do We Randomize?

### I. Fixed Allocation Randomization

Consider

- Two treatments A & B.
- If patient population were given A, the average response would be  $\mu_1$ .
- If patient population were given B, the average response would be  $\mu_2$ .
- We are interested in estimating  $\Delta = \mu_1 - \mu_2$  and make inference on  $\Delta$ .
- With fixed allocation, the probability that each patient receives treatment A is a constant  $\pi$ , usually  $\pi = 0.5$ .

- Suppose after treatment allocation,  $n_1$  patients received A and  $n_2$  received B, with  $n = n_1 + n_2$  being the total sample size.

- ★ We will estimate  $\Delta$  using

$$\hat{\Delta} = \bar{Y}_1 - \bar{Y}_2,$$

where  $\bar{Y}_1$  is the sample average response of the  $n_1$  patients receiving treatment A and  $\bar{Y}_2$  is the sample average response of the  $n_2$  patients receiving treatment B

- ★ **When** is  $\hat{\Delta}$  most efficient?

- ★ Assume equal variance, then

$$\text{var}(\hat{\Delta}) = \text{var}(\bar{Y}_1) + \text{var}(\bar{Y}_2) = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \approx \frac{\sigma^2}{n} \left\{ \frac{1}{\pi(1-\pi)} \right\}.$$

- ★ Minimizing the above variance gives  $\pi = 0.5$ , the minimum variance is  $4\sigma^2/n$ .
- ★ If  $\pi \neq 0.5$ , it is less efficient. But the loss is not great. For

example, if  $\pi = 2/3$ , with the same  $n$ , the variance of  $\text{var}(\hat{\Delta})$  will be  $4.5\sigma^2/n$ .

- ★ If we want the same efficiency, we only need to increase sample size by  $4.5/4 - 1 = 0.125$ . That is, a 12.5% increase in sample size.
- Some investigators prefer to put more patients in the new treatment
  - ★ better experience on a drug where there is little information
  - ★ efficiency loss is slight
  - ★ if new treatment is good (as is hoped) more patients will benefit
  - ★ might be more cost efficient
- Putting more patients in the new treatment has disadvantage too
  - ★ might be difficult to justify ethically; It removes equipoise for the participating clinician
  - ★ new treatment may be detrimental

1. Simple randomization: Each patient has probability  $\pi$  to receive A (hence probability  $1 - \pi$  to receive B); usually  $\pi = 0.5$ .

For patient  $i$ , generate a uniform random variable  $U_i \in [0, 1]$

$$\text{If } \begin{cases} U_i \leq \pi \text{ then assign treatment A} \\ U_i > \pi \text{ then assign treatment B.} \end{cases}$$

### Advantages of simple randomization

- easy to implement
- virtually impossible for the investigators to guess what the next treatment assignment will be.
- the properties of many statistical inferential procedures (tests and estimators) are established under the simple randomization assumption (iid)

**Disadvantages of simple randomization:** The major disadvantage is that the number of patients assigned to the different treatments are random. Therefore, the possibility exists of severe treatment imbalance (even with equal allocation probability  $\pi = 0.5$ )

- leads to less efficiency:
- appears awkward and may lead to loss of credibility in the results of the trial

For example, with  $n = 20$ ,

$$P[\text{imbalance of 12:8 or worse} | \pi = 0.5] \approx 0.5.$$

When  $n = 100$ ,

$$P[\text{imbalance of 60:40 or worse} | \pi = 0.5] \approx 0.05.$$

## 2. Permuted block randomization: try to balance A & B.

- (a) Permuted block randomization with a **fixed** block size; for example block size=4; then 6 possible combinations:

A A B B – per1

A B A B – per2

A B B A – per3

B A A B – per4

B A B A – per5

B B A A – per6

for each block of 4 patients, randomly pick up one combination and assign the treatments to those 4 patients in the sequence specified by the combination.

## Ways to Choose a random permutation

- i. Order the 6 permutations by  $per1 - per6$ ; generate a uniform random number  $U_i$  for the  $i$ th block of 4 patients; if  $U_i \in [0, 1/6]$ , then use  $per1$ ; if  $U_i \in [1/6, 2/6]$ , then pick up  $per2$ , etc.
- ii. For AABB, generate a uniform random number for each letter; re-order the random numbers (ascending or descending). Then the re-ordered letters give a permutation as illustrated in the following table:

Treatment	random number	rank
A	0.069	1
A	0.734	3
B	0.867	4
B	0.312	2

This table gives ABAB.

**Potential problem:**

If the block size (such as 4) is known, the physician can guess what treatment next patient is going to receive (with certainty for the last treatment). This may cause bias in estimating treatment effect. Solution is ...

- (b) **Permuted block randomization with varying block size:** choose several block sizes in advance and pick up a block size with some pre-specified probability; after a block size is chosen, pick up the permutation randomly (with each probability).
- 3. Stratified Randomization** (often used with blocking): form strata using prognostic factors; then in each stratum, perform permuted block randomization (with fixed or varying block size).

For example, if age and gender are strong prognostic factors, then we can form following strata:

	Age		
Gender	40-49	50-59	60-69
Male			
Female			

The maximum imbalance between A & B: # of strata  $\times$  (block size)/2.

## Advantages of Stratified Randomization

- Makes the treatment groups appear similar. This can give more credibility to the results of a study
- Blocked randomization within strata may result in more precise estimates of treatment difference; but one must be careful to conduct the appropriate analysis

## Illustration on the effect that blocking within strata has on the precision of estimators

A prognostic factor with 2 strata:

$$S = \begin{cases} 1 = \text{strata 1} \\ 0 = \text{strata 0.} \end{cases}$$

Let

$$X = \begin{cases} 1 = \text{treatment A} \\ 0 = \text{treatment B.} \end{cases}$$

Assume a model for the response  $Y$  for the  $i$ th patient:

$$Y_i = \mu + \alpha S_i + \beta X_i + \epsilon_i$$

$\beta$  is the treatment effect,  $\epsilon_i$  are iid errors with mean 0 and variance  $\sigma^2$ .

Denote sample means:  $\bar{Y}_A$  and  $\bar{Y}_B$ :

$$\bar{Y}_A = \sum_{X_i=1} Y_i/n_A,$$

$$\bar{Y}_B = \sum_{X_i=0} Y_i/n_B,$$

where  $n_A = \sum_{i=1}^n X_i$ , number of patients receiving treatment A,  
 $n_B = n - n_A$ .

We will estimate treatment effect  $\beta$  by

$$\hat{\Delta} = \bar{Y}_A - \bar{Y}_B.$$

Assume

Table 2: *Number of observations falling into the different strata by treatment*

strata	Treatment		total
	A	B	
0	$n_{A0}$	$n_{B0}$	$n_0$
1	$n_{A1}$	$n_{B1}$	$n_1$
total	$n_A$	$n_B$	$n$

Then

$$\begin{aligned}\bar{Y}_A &= \sum_{X_i=1} Y_i/n_A \\ &= \sum_{X_i=1} (\mu + \alpha S_i + \beta X_i + \epsilon_i)/n_A \\ &= (n_A \mu + \alpha \sum_{X_i=1} S_i + \beta \sum_{X_i=1} X_i + \sum_{X_i=1} \epsilon_i)/n_A \\ &= (n_A \mu + \alpha n_{A1} + \beta n_A + \sum_{X_i=1} \epsilon_i)/n_A \\ &= \mu + \alpha \frac{n_{A1}}{n_A} + \beta + \bar{\epsilon}_A,\end{aligned}$$

where  $\bar{\epsilon}_A = \sum_{X_i=1} \epsilon_i/n_A$ .

Similarly,

$$\bar{Y}_B = \mu + \alpha \frac{n_{B1}}{n_B} + \bar{\epsilon}_B,$$

where  $\bar{\epsilon}_B = \sum_{X_i=0} \epsilon_i / n_B$ . Therefore

$$\hat{\Delta} = \bar{Y}_A - \bar{Y}_B = \beta + \alpha \left( \frac{n_{A1}}{n_A} - \frac{n_{B1}}{n_B} \right) + (\bar{\epsilon}_A - \bar{\epsilon}_B).$$

- Under stratified blocked randomization:

$$n_A \approx n_B \approx n/2$$

$$n_{A1} \approx n_{B1} \approx n_1/2$$

$$n_{A0} \approx n_{B0} \approx n_0/2.$$

So

$$E(\hat{\Delta}) = \beta,$$

$$\text{var}(\hat{\Delta}) = \text{var}(\bar{\epsilon}_A) + \text{var}(\bar{\epsilon}_B) = \sigma^2 \left( \frac{2}{n} + \frac{2}{n} \right) = \frac{4\sigma^2}{n}.$$

- Under **simple randomization**:  $n_A, n_B, n_{A1}$  and  $n_{B1}$  are all random, and

$$n_{A1}|n_A, n_B \sim b(n_A, \theta), \quad n_{B1}|n_A, n_B \sim b(n_B, \theta),$$

where  $\theta$  is the probability that a patient is in stratum 1.

So

$$\begin{aligned} E(\widehat{\Delta}) &= E(\bar{Y}_A - \bar{Y}_B) \\ &= \beta + \alpha \left\{ E\left(\frac{n_{A1}}{n_A}\right) - E\left(\frac{n_{B1}}{n_B}\right) \right\} + E(\bar{\epsilon}_A - \bar{\epsilon}_B) \\ &= \beta \end{aligned}$$

and

$$\begin{aligned}\text{var}(\widehat{\Delta}) &= \text{var}(\bar{Y}_A - \bar{Y}_B) \\ &= E\{\text{var}(\bar{Y}_A - \bar{Y}_B | n_A, n_B)\} + \text{var}\{E(\bar{Y}_A - \bar{Y}_B | n_A, n_B)\}.\end{aligned}$$

Since

$$\begin{aligned}& \text{var}(\bar{Y}_A - \bar{Y}_B | n_A, n_B) \\ &= \text{var} \left\{ \beta + \alpha \left( \frac{n_{A1}}{n_A} - \frac{n_{B1}}{n_B} \right) + (\bar{\epsilon}_A - \bar{\epsilon}_B) | n_A, n_B \right\} \\ &= \alpha^2 \left\{ \text{var} \left( \frac{n_{A1}}{n_A} | n_A \right) + \text{var} \left( \frac{n_{B1}}{n_B} | n_B \right) \right\} \\ & \quad + \text{var}(\bar{\epsilon}_A | n_A) + \text{var}(\bar{\epsilon}_B | n_B) \\ &= \alpha^2 \left\{ \frac{\theta(1-\theta)}{n_A} + \frac{\theta(1-\theta)}{n_B} \right\} + \left( \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} \right) \\ &= \{ \sigma^2 + \alpha^2 \theta(1-\theta) \} \left( \frac{1}{n_A} + \frac{1}{n_B} \right).\end{aligned}$$

Therefore

$$\begin{aligned}
 \text{var}(\bar{Y}_A - \bar{Y}_B) &= E\{\text{var}(\bar{Y}_A - \bar{Y}_B | n_A, n_B)\} \\
 &= \{\sigma^2 + \alpha^2\theta(1 - \theta)\} E\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \\
 &= \{\sigma^2 + \alpha^2\theta(1 - \theta)\} E\left(\frac{1}{n_A} + \frac{1}{n - n_A}\right),
 \end{aligned}$$

where  $n_A \sim b(n, 1/2)$ .

It has shown that

$$\frac{1}{n_A} + \frac{1}{n - n_A} \geq \frac{4}{n}.$$

Hence

$$\text{var}(\hat{\Delta}) \geq \frac{4}{n} \{\sigma^2 + \alpha^2\theta(1 - \theta)\} > \frac{4\sigma^2}{n},$$

which is the variance  $\hat{\Delta}$  obtained under stratified blocked randomization.

**Remark:** Suppose we perform stratified blocked randomization, we should take the design into account in data analysis. For example, if we simply use two-sample t-test

$$\frac{\bar{Y}_A - \bar{Y}_B}{s_P \left( \frac{1}{n_A} + \frac{1}{n_B} \right)^{1/2}},$$

where

$$s_P^2 = \left\{ \frac{\sum_{X_i=1} (Y_i - \bar{Y}_A)^2 + \sum_{X_i=0} (Y_i - \bar{Y}_B)^2}{n_A + n_B - 2} \right\}.$$

It turns out that  $s_P^2$  is an unbiased estimator for  $\{\sigma^2 + \alpha^2\theta(1 - \theta)\}$  as it should be for simple randomization. However, with stratified randomization, we showed that the variance of  $(\bar{Y}_A - \bar{Y}_B)$  is  $\frac{4\sigma^2}{n}$ .

Therefore the statistic

$$\frac{\bar{Y}_A - \bar{Y}_B}{s_P \left( \frac{1}{n_A} + \frac{1}{n_B} \right)^{1/2}} \approx \frac{\bar{Y}_A - \bar{Y}_B}{\{\sigma^2 + \alpha^2\theta(1 - \theta)\}^{1/2} \left( \frac{2}{n} + \frac{2}{n} \right)^{1/2}},$$

has variance

$$\frac{4\sigma^2/n}{4\{\sigma^2 + \alpha^2\theta(1 - \theta)\}/n} = \frac{\sigma^2}{\sigma^2 + \alpha^2\theta(1 - \theta)} \leq 1.$$

Hence the statistic commonly used to test differences in means between two populations

$$\frac{\bar{Y}_A - \bar{Y}_B}{s_P \left( \frac{1}{n_A} + \frac{1}{n_B} \right)^{1/2}},$$

does not have a t-distribution if used with a stratified design and  $\alpha \neq 0$  (i.e. some strata effect). In fact, it has a distribution with smaller variance. Thus, if this test were used in conjunction with a stratified randomized design, then the resulting analysis would

be conservative.

The correct analysis would have considered the strata effect in a two-way analysis of variance ANOVA which would then correctly estimate the variance of the estimator for treatment effect.

**In general, if we use permuted block randomization within strata in the design, we need to account for this in the analysis.**

In contrast, if we used simple randomization and the two-sample t-test, we would be making correct inference. Even so, we might still want to account for the effect of strata post-hoc in the analysis to reduce the variance and get more efficient estimators for treatment difference.

**Disadvantage of blocking within strata:** If we use too many prognostic factors to form strata, we might end up with very few (or even zero) patients in some strata. If each stratum has only one patient, we are back to simple randomization.

**II. Adaptive Randomization Procedures:** the allocation probability depends on the treatment allocation of previous patients

- 1. Efron biased coin design:** Choose an integer  $D$  and a probability  $\phi < 0.5$  (for example,  $D = 3$  and  $\phi = 0.25$ ). Assign next patient to treatment A with  $\pi_A$ :

$$\pi_A = .5 \quad \text{if } |n_A - n_B| \leq D$$

$$\pi_A = \phi \quad \text{if } n_A - n_B > D$$

$$\pi_A = 1 - \phi \quad \text{if } n_B - n_A > D$$

- 2. Urn Model (L.J. Wei):** Start with  $m$  balls labeled with A and  $m$  balls labeled with B. Randomly pick a ball for the first patient and assign the treatment indicated by the ball to that patient. If the patient receives A then replace that A ball with a B ball and vice versa.

3. **Minimization Method of Pocock and Simon:** Suppose there are  $K$  prognostic factors, each with  $k_i$  levels ( $i = 1, 2, \dots, K$ ). At any point in time in the study, let us denote by  $n_{Aij}$  the number of patients that are on treatment A for the  $j$ -th level of prognostic factor  $i$ . An analogous definition for  $n_{Bij}$ .

**Note:** If  $n_A$  denotes the total number on treatment A, then

$$n_A = \sum_{j=1}^{k_i} n_{Aij}; \text{ for all } i = 1, \dots, K.$$

Similarly,

$$n_B = \sum_{j=1}^{k_i} n_{Bij}; \text{ for all } i = 1, \dots, K.$$

The measure of marginal discrepancy is given by

$$MD = w_0 |n_A - n_B| + \sum_{i=1}^K w_i \left( \sum_{j=1}^{k_i} |n_{Aij} - n_{Bij}| \right).$$

The weights  $w_0, w_1, \dots, w_K$  are positive numbers which may differ according to the emphasis you want to give to the different prognostic factors. Generally  $w_0 = K, w_1 = \dots = w_K = 1$ .

The next patient that enters the study is assigned either treatment A or treatment B according to whichever makes the subsequent measure of marginal discrepancy smallest. In case of a tie, the next patient is randomized with probability .5 to either treatment. We illustrate with an example. For simplicity, consider two prognostic factors,  $K=2$ , the first with two levels,  $k_1 = 2$  and the second with three levels  $k_2 = 3$ . Suppose after 50 patients have entered the study, the marginal configuration of counts for treatments A and B, by prognostic factors, looks as follows:

Treatment A				Treatment B			
PF1				PF1			
PF2	1	2	Total	PF2	1	2	Total
1		*	13	1		*	12
2			9	2			6
3			4	3			6
Total	16	10	26	Total	14	10	24

If we take the weights to be  $w_0 = 2$  and  $w_1 = w_2 = 1$ , then the measure of marginal discrepancy equals

$$MD = 2|26 - 24| + 1(|16 - 14| + |10 - 10|) + 1(|13 - 12| + |9 - 6| + |4 - 6|) = 12.$$

Suppose the next patient entering the study is at the second level of PF1 and the first level of PF2. Which treatment should that patient be randomized to.?

If the patient were randomized to treatment A, then the result would be

Treatment A				Treatment B			
PF1				PF1			
PF2	1	2	Total	PF2	1	2	Total
1			14	1			12
2			9	2			6
3			4	3			6
Total	16	11	27	Total	14	10	24

and the measure of marginal discrepancy

$$MD = 2|27-24| + 1(|16-14| + |11-10|) + 1(|14-12| + |9-6| + |4-6|) = 16.$$

Whereas, if that patient were assigned to treatment B, then

Treatment A				Treatment B			
PF1				PF1			
PF2	1	2	Total	PF2	1	2	Total
1			13	1			13
2			9	2			6
3			4	3			6
Total	16	10	26	Total	14	11	25

and the measure of marginal discrepancy

$$MD = 2|26 - 25| + 1(|16 - 14| + |10 - 11|) + 1(|13 - 13| + |9 - 6| + |4 - 6|) = 10.$$

Therefore, we would assign this patient to treatment B.

Note that design-based inference is not even possible since the allocation is virtually deterministic.

**III. Response Adaptive Randomization:** allocation probability depends on the **outcome** of the previous patients.

**1. Play-the-Winner Rule (Zelen):**

- First patient is randomized to either treatment A or B with equal probability.
- Next patient is assigned the same treatment as the previous one if the previous patient's response was a success; whereas, if the previous patient's response is a failure, then the patient receives the other treatment. The process calls for staying with the winner until a failure occurs and then switching.

For example,

	Patient ordering							
Treatment	1	2	3	4	5	6	7	8
A	S	F				S	S	F
B			S	S	F			

2. **Urn Model (L.J. Wei)**: Every time there is a success on treatment A add  $r$  A balls into the urn, when there is a failure on treatment A add  $r$  B balls. Similarly for treatment B. The next patient is assigned to whichever ball is drawn at random from this urn.

Response adaptive allocation schemes have the intended purpose of maximizing the number of patients in the trial that receive the superior treatment.

### **Difficulties with response adaptive allocation schemes**

- Information on response may not be available immediately.
- Such strategies may take a greater number of patients to get the desired answer. Even though more patients on the trial may be getting the better treatment, by taking a longer time, this better treatment is deprived from the population at large who may benefit.
- May interfere with the ethical principle of equipoise.

- Results may not be easily interpretable from such a design.

### **ECMO trial:**

Extracorporeal membrane oxygenator was a promising treatment for a neonatal population suffering from respiratory insufficiency. This device oxygenates the blood to compensate for the lung's inability or inefficiency in achieving this task. The mortality rate was very high for this population and due to very promising results of ECMO it was decided to use a play-the-winner rule.

The first child was randomized to the control group and died. The next 10 children were assigned ECMO and all survived at which point the trial was stopped and ECMO declared a success.

It turned out that after further investigation, the first child was the sickest of all the children studied. Controversy ensued and the study had to be repeated using a more traditional design.

Footnote on page 73 of the textbook FFD gives further references.

## Mechanics of Randomization

The following formal sequence of events should take place before a patient is randomized into a phase III clinical trial.

- Patient requires treatment
- Patient is eligible for the trial. Inclusion and exclusion criteria should be checked immediately. For a large multi-center trial, this may be done at a central registration office
- Clinician is willing to accept randomization
- Patient consent is obtained. In the US this is a legal requirement
- Patient formally entered into the trial

After a patient and his/her physician agree to participate in the trial then

- Each patient must be formally identified. This can be done by collecting some minimal information; i.e. name, date of birth, hospital number. This information should be kept on a log (perhaps

at a central office) and given a trial ID number for future identification. This helps keep track of the patient and it helps guard against investigators not giving the allocated treatment.

- The treatment assignment is obtained from a randomization list. Most often prepared in advance
  1. The randomization list could be transferred to a sequence of sealed envelopes each containing the name of the next treatment on the card. The clinician opens the envelope when a patient has been formerly registered onto the trial
  2. If the trial is double-blind then the pharmacist preparing the drugs needs to be involved. They prepare the sequence of drug packages according to the randomization list.
  3. For a multi-center trial, randomization is carried out by the central office by phone or by computer.
  4. For a double-blind multi-center trial, the randomization may need to be decentralized to each center according to (2).

However, central registration is recommended.

## Documentation

- A confirmation form needs to be filled out after treatment assignment which contains name, trial number and assigned treatment. If randomization is centralized then this confirmation form gets sent from the central office to the physician. If it is decentralized then it goes from physician to central office.
- An on-study form is then filled out containing all relevant information prior to treatment such as previous therapies, personal characteristics (age, race, gender, etc.), details about clinical conditions and certain laboratory tests (e.g. lung function for respiratory illness)

All of these checks and balances must take place quickly but accurately prior to the patient commencing therapy.

## 5 Some Additional Issues in Phase III Clinical Trials

**Double blinding:** neither patient, physician nor evaluator are aware which treatment the patient is receiving.

- The patient— psychological benefit
- The treatment team— management and care of patients may be different
- The evaluator— try to make endpoint objective

**Other methods** to reduce bias:

- Make treatments to be compared look, taste, feel similar, etc.
- Use **placebo** when no best treatment is available

**The Hippocratic Oath:** *I swear by Apollo the physician, by Aesculapius, Hygeia and Panacea, and I take to witness all the gods, all the goddesses, to keep according to my ability and my judgment the following Oath...*

**Modern version of The Hippocratic Oath:** *I swear to fulfill, to the best of my ability and judgment, this covenant:*

*I will respect the hard-won scientific gains of those physicians in whose steps I walk, and gladly share such knowledge as is mine with those who are to follow...*

**IRB:** Institutional Review Board or Internal Review Board

1. The risks to the study participants are minimized
2. The risks are reasonable in relation to the anticipated benefit
3. The selection of study patients is equitable
4. Informed consent is obtained and appropriately documented for each participant
5. There are adequate provisions for monitoring data collected to ensure the safety of the study participants
6. The privacy of the participants and confidentiality of the data are protected

## The Protocol Document

**Definition:** a scientific document for a medical study on human subjects; contains background, experimental design, patient population, treatment and evaluation details, and data collection procedures.

### Purposes

1. To assist investigators in thinking through the research
2. To ensure that both patient and study management are considered at the planning stage
3. To provide a sounding board for external comments
4. To orient the staff for preparation of forms and processing procedures
5. To provide a document which can be used by other investigators who wish to confirm (replicate) the results

A protocol generally has the following elements:

1. **Schema:** Depicts the essentials of a study design.

WHI: page 18

2. **Objectives:** The objectives should be few in number and should be based on specific quantifiable endpoints

WHI: pages 14-15 and pages 22-24

3. **Project background:** This section should give the referenced medical/historical background for therapy of these patients.

WHI: pages 2-13

This generally includes

- (a) standard therapy
- (b) predecessor studies (phase I and II if appropriate)
- (c) previous or concurrent studies of a similar nature
- (d) moral justification of the study

4. **Patient Selection:** A clear definition of the patient population to

be studied. This should include clear, unambiguous inclusion and exclusion criteria that are verifiable at the time of patient entry. Each item listed should be verified on the study forms.

WHI: pages 24-28

- 5. Randomization/Registration Procedures** This section spells out the mechanics of entering a patient into the study

WHI: pages 29-38

- 6. Treatment Administration and Patient Management:** How the treatment is to be administered needs to be specified in detail. All practical eventualities should be taken into account, at least, as much as possible. Protocols should not be written with only the study participants in mind. Others may want to replicate this therapy such as community hospitals that were not able to participate in the original study.

WHI: pages 18-22 and 44-49

- 7. Study parameters:** This section gives the schedule of the

required and optional investigations/tests.

WHI: pages 38-39

## 8. Statistical Considerations:

WHI: pages 52-55 and an extensive appendix

- Study outline, stratification and randomization
- Sample size criteria: Motivation for the sample size and duration of the trial needs to be given. This can be based on type I and type II error considerations in a hypothesis testing framework or perhaps based on the desired accuracy of a confidence interval.
- Accrual estimates
- Power calculations
- Brief description of the data analysis that will be used
- Interim monitoring plans

## 9. Informed Consent The consent form needs to be included.

- (a) an explanation of the procedures to be followed and their

purposes

- (b) a description of the benefits that might reasonably be expected
- (c) a description of the discomforts and risks that could reasonably be expected
- (d) a disclosure of any appropriate alternative procedures that might be advantageous
- (e) a statement that the subject is at liberty to abstain from participation in the study and is free to withdraw at any time

**10. Study Management Policy:** This section includes how the study will be organized and managed, when the data will be summarized and the details of manuscript development and publication

WHI: pages 58-61

## 6 Sample Size Calculations

A major responsibility of a statistician: sample size calculation.

**Hypothesis Testing:** compare treatment 1 (new treatment) to treatment 2 (standard treatment); Assume continuous endpoints.

- $\Delta =$  treatment effect, parameter of interest. For example  $\Delta = \mu_1 - \mu_2$ .
- nuisance parameters:  $\theta = (\mu_2, \sigma^2)$
- $H_0 : \Delta \leq 0$ : stay with the standard treatment
- $H_A : \Delta > 0$ : switch to the new treatment
- Data  $(z_1, \dots, z_n)$ ,  $z_i =$  realization of  $Z_i = (Y_i, A_i)$ , where

$$Y_i | A_i = 1 \sim N(\mu_1, \sigma^2), \quad Y_i | A_i = 2 \sim N(\mu_2, \sigma^2)$$

- Construct a test statistic

$$T = T_n(z_1, \dots, z_n).$$

For example, two-sample t-test statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{n_1^{-1} + n_2^{-1}}}.$$

The statistic  $T$  should be constructed in such a way that

1. Larger values of  $T$  are evidence against  $H_0$  (in favor of  $H_A$ ).
  2. The distribution of  $T$  can be (approximately) evaluated at the **border** between  $H_0$  and  $H_A$ ; i.e. at  $\Delta = 0$ .
- Given observed test statistic  $T_{obs}$ , p-value for testing  $H_0 : \Delta \leq 0$  vs.  $H_A : \Delta > 0$  is calculated as

$$P_{\Delta=0}(T \geq T_{obs}).$$

For given type I error prob  $\alpha$  (usually 0.025 or 0.05), reject  $H_0$  if

p-value  $< \alpha$ .

- **Note:**

1. Most often,  $P_{\Delta}(T \geq x)$  increases as  $\Delta$  increases, for all  $x$ . So

$$P_{\Delta=0}[T \geq T_{obs}] \leq \alpha \implies P_{\Delta}[T \geq T_{obs}] \leq \alpha \text{ for all } \Delta \in H_0.$$

2. The distribution of  $T$  at  $\Delta = 0$  is known and

$$T \stackrel{(\Delta=0)}{\sim} N(0, 1).$$

$$\text{P-value} \leq \alpha \iff T_{obs} \geq z_{\alpha}.$$

3. For the two-sample t-test statistic, we have

$$T \stackrel{(\Delta=0)}{\sim} t_{n-2} \approx N(0, 1).$$

- **Remark on two-sided tests:**

$$H_0 : \Delta = 0 \text{ vs. } \Delta \neq 0$$

★ Reject  $H_0$  if  $|T|$  is large

★ P-value

$$P_{\Delta=0}(|T| \geq |T_{obs}|) = P_{\Delta=0}[T \geq |T_{obs}|] + P_{\Delta=0}[T \leq -|T_{obs}|].$$

★ For given  $\alpha$ , reject  $H_0$  if  $|T| \geq z_{\alpha/2}$ .

● rejection region

For one-sided level  $\alpha$  tests, the rejection region is

$$\{(z_1, \dots, z_n) : T_n(z_1, \dots, z_n) \geq \mathcal{Z}_\alpha\},$$

and for two-sided level  $\alpha$  tests, the rejection region is

$$\{(z_1, \dots, z_n) : |T_n(z_1, \dots, z_n)| \geq \mathcal{Z}_{\alpha/2}\}.$$

● **Power**

For one-sided tests:

$$P_{\Delta=\Delta_A}[T \geq z_\alpha], \quad \text{for } \Delta_A \in H_A.$$

---

Usually we would like to have high power (0.9) to detect a clinically

important  $\Delta_A$ .

- Often time,  $T$  has (approximate) normal distribution under  $\Delta = \Delta_A$ :

$$T \stackrel{H_A=(\Delta_A, \theta)}{\underset{\sim}{\approx}} N(\phi(n, \Delta_A, \theta), \sigma_*^2(\Delta_A, \theta)).$$

Usually,  $\sigma_*^2(\Delta_A, \theta) = 1$ . In this case,  $\phi(n, \Delta_A, \theta)$  is called the non-centrality parameter.

For example, the two-sample t-test statistic:

$$T_n = \frac{\bar{Y}_1 - \bar{Y}_2}{s_Y \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{1/2}} \approx \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma_Y \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{1/2}}$$

Then

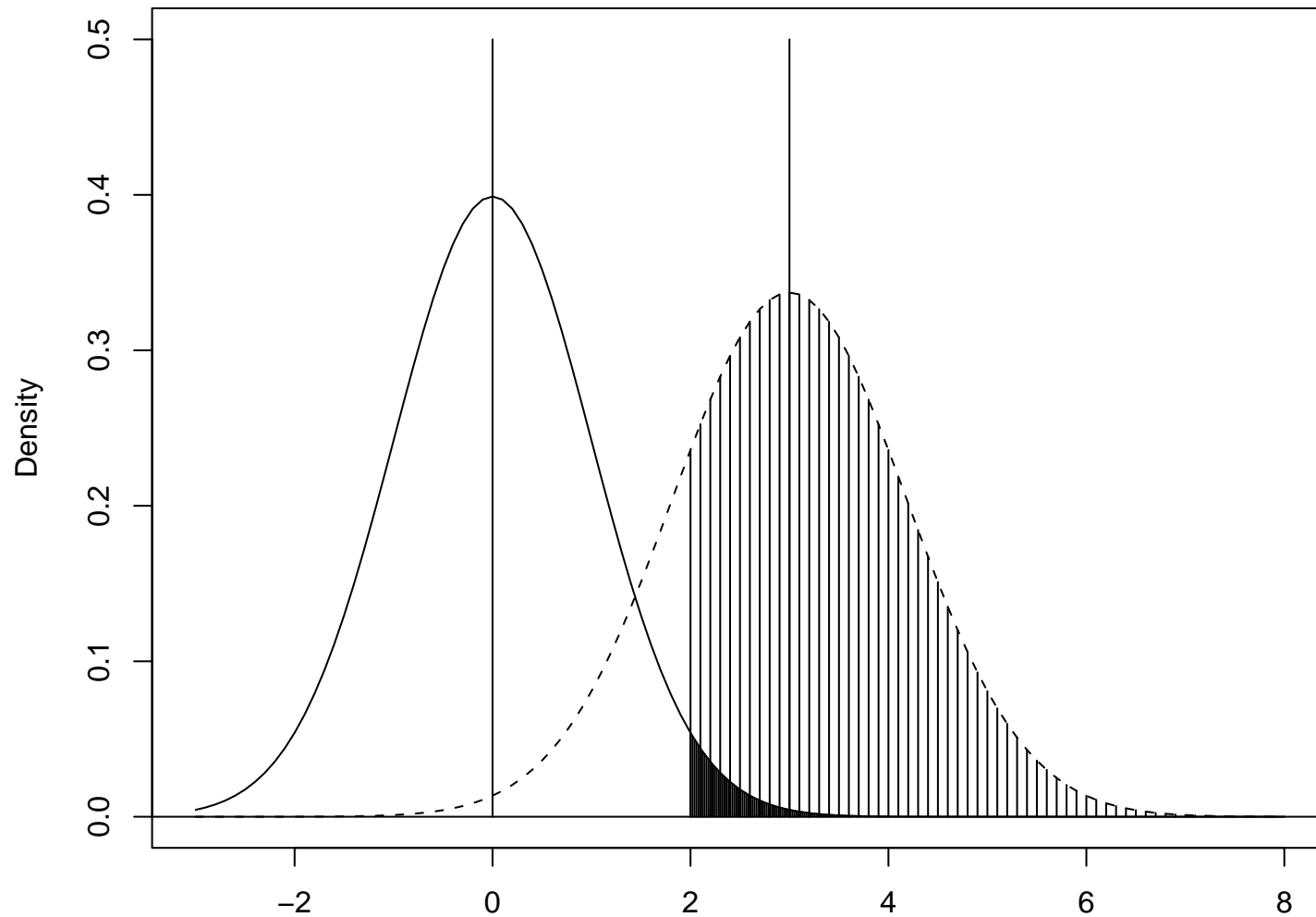
$$\phi(n, \Delta_A, \theta) = \frac{\Delta_A}{\sigma_Y \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{1/2}}, \quad \sigma_*^2(\Delta_A, \theta) = 1.$$

## Deriving sample size to achieve desired power

### Design characteristics:

- Use the above testing procedure
- Type I error probability  $\alpha$ .
- Power  $1 - \beta$  to detect clinically important treatment effect  $\Delta_A$
- Nuisance parameter  $\theta$  is known

### How to find sample size?

Figure 1: *Distributions of  $T$  under  $H_0$  and  $H_A$* 

- The figure indicates the equation:

$$\phi(n, \Delta_A, \theta) = \{z_\alpha + z_\beta \sigma_*(\Delta_A, \theta)\}. \quad (6.1)$$

For the two-sample t-test, if we do equal allocation ( $n_1 = n_2 = n/2$ ), then

$$\frac{\Delta_A}{\sigma_Y \left(\frac{2}{n} + \frac{2}{n}\right)^{1/2}} = z_\alpha + z_\beta$$

$\implies$

$$n = \left\{ \frac{(z_\alpha + z_\beta)^2 \sigma_Y^2 \times 4}{\Delta_A^2} \right\}.$$

- **Note:** For two-sided tests we replace  $z_\alpha$  by  $z_{\alpha/2}$ .

- **Example:** find the sample size necessary to detect a difference in mean response of 20 units between two treatments with 90% power using a  $t$ -test (two-sided) at the .05 level of significance. We assume population standard deviation of response  $\sigma_Y$  is expected to be about 60 units.

$$z_{\alpha/2} = z_{.025} = 1.96, z_{\beta} = z_{0.1} = 1.28, \Delta_A = 20, \sigma_Y = 60,$$

$$n = \frac{(1.96 + 1.28)^2 (60)^2 \times 4}{(20)^2} \approx 378 \text{ (rounding up),}$$

or about 189 per each treatment.

- **How large** should  $\hat{\Delta}$  be so that we will have a significant p-value (p-value=0.05) for the calculated sample size?

$$P_{\Delta=0}[T \geq |T_{obs}|] + P_{\Delta=0}[T \leq -|T_{obs}|] = 0.05$$

 $\Leftrightarrow$ 

$$P_{\Delta=0}[T \geq |T_{obs}|] = 0.025$$

 $\Leftrightarrow$ 

$$T_{obs} = z_{0.025} = 1.96$$

 $\Leftrightarrow$ 

$$\frac{\hat{\Delta}}{\sigma_Y \left(\frac{2}{n} + \frac{2}{n}\right)^{1/2}} = z_{0.025} = 1.96$$

 $\Rightarrow$ 

$$\hat{\Delta} = 1.96 \times \sigma_Y \left(\frac{4}{n}\right)^{1/2} = 1.96 \times 60 \times 2/\sqrt{378} = 12.1 < 20$$

- If the study result turns out to be what we expected, what P-value will be expected?

$$T_{obs} = \frac{\hat{\Delta}}{\sigma_Y \left(\frac{2}{n} + \frac{2}{n}\right)^{1/2}} = \frac{20}{60 \times 2 / \sqrt{378}} = 3.24$$

$$P - \text{value} = 2P_{\Delta=0}[T > |T_{obs}|] = 2P_{\Delta=0}[T > 3.24] = 0.001.$$

## Comparing two response rates

- $\pi_1 =$  response rate of treatment 1,  $\pi_2 =$  response rate of treatment 2
- Treatment effect  $\Delta = \pi_1 - \pi_2$
- $n_1$  patients are to be assigned to treatment 1,  $n_2$  patients are to be assigned to treatment 2 (usually,  $n_1 = n_2$ )
- Wish to test  $H_0 : \Delta \leq 0$  ( $\pi_1 \leq \pi_2$ ) versus  $H_A : \Delta > 0$  ( $\pi_1 > \pi_2$ ).
- Data from each treatment:

$$X_1 \sim \text{bin}(n_1, \pi_1), \quad X_2 \sim \text{bin}(n_2, \pi_2)$$

- $p_1 = X_1/n_1, p_2 = X_2/n_2$  best estimates of  $\pi_1$  and  $\pi_2$

$$E(p_1) = \pi_1, \text{ var}(p_1) = \frac{\pi_1(1 - \pi_1)}{n_1},$$

$$E(p_2) = \pi_2, \text{ var}(p_2) = \frac{\pi_2(1 - \pi_2)}{n_2}.$$

- Test statistic for testing  $H_0$ :

$$T = \frac{p_1 - p_2}{\left\{ \bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{1/2}},$$

where  $\bar{p} = (X_1 + X_2)/(n_1 + n_2)$ , best estimate of  $\pi_1(\pi_2)$  under  $\pi_1 = \pi_2$ .

**Note:** The  $T^2$  is the usual chi-square test used to test equality of proportions.

- We can write

$$\bar{p} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = p_1 \left( \frac{n_1}{n_1 + n_2} \right) + p_2 \left( \frac{n_2}{n_1 + n_2} \right).$$

So

$$\bar{p} \approx \pi_1 \left( \frac{n_1}{n_1 + n_2} \right) + \pi_2 \left( \frac{n_2}{n_1 + n_2} \right) = \bar{\pi},$$

$\bar{\pi}$  is a weighted average of  $\pi_1$  and  $\pi_2$ .

Therefore,

$$T \approx \frac{p_1 - p_2}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{1/2}}.$$

- The mean and variance of  $T$  under  $\Delta = 0$ :

$$\begin{aligned}
 E_{\Delta=0}(T) &\approx E_{\Delta=0} \left\{ \frac{p_1 - p_2}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{1/2}} \right\} \\
 &= \frac{E_{\Delta=0}(p_1 - p_2)}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{1/2}} = 0 \\
 \text{var}_{\Delta=0}(T_n) &\approx \frac{\{\text{var}_{\Delta=0}(p_1) + \text{var}_{\Delta=0}(p_2)\}}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}} \\
 &= \frac{\left\{ \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} \right\}}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}} = 1.
 \end{aligned}$$

So under  $\Delta = 0$ ,

$$T \stackrel{(\Delta=0)}{\sim} N(0, 1)$$

- Under  $H_A : \Delta = \Delta_A$ :

$$T \stackrel{(\Delta=\Delta_A)}{\sim} N(\phi(n, \Delta_A, \theta), \sigma_*^2)$$

where

$$\begin{aligned} \phi(n, \Delta_A, \theta) &= E_{H_A}(T) \approx \frac{(\pi_1 - \pi_2)}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{1/2}} \\ &= \frac{\Delta_A}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{1/2}}, \\ \sigma_*^2 &= \text{var}_{H_A}(T) \approx \frac{\left\{ \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} \right\}}{\left\{ \bar{\pi}(1 - \bar{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}}. \end{aligned}$$

- With equal allocation of treatment,  $n_1 = n_2 = n/2$ , then

$$\phi(n, \Delta_A, \theta) = \frac{\Delta_A}{\{\bar{\pi}(1 - \bar{\pi})\frac{4}{n}\}^{1/2}},$$

and

$$\sigma_*^2 = \frac{\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)}{2\bar{\pi}(1 - \bar{\pi})},$$

where  $\pi_1 = \pi_2 + \Delta_A$ .

- If we want to have power  $1 - \beta$  to detect an increase of  $\Delta_A$  with significance level  $\alpha$  using one-sided test (and equal allocation), the sample size  $n$  have to satisfy

$$\frac{n^{1/2}\Delta_A}{\{4\bar{\pi}(1 - \bar{\pi})\}^{1/2}} = Z_\alpha + Z_\beta \left\{ \frac{\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)}{2\bar{\pi}(1 - \bar{\pi})} \right\}^{1/2}.$$

So the sample size is given by

$$n = \frac{\left\{ Z_{\alpha} + Z_{\beta} \left\{ \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{2\bar{\pi}(1-\bar{\pi})} \right\}^{1/2} \right\}^2 4\bar{\pi}(1-\bar{\pi})}{\Delta_A^2}. \quad (6.2)$$

**Note:** For two-sided tests we replace  $Z_{\alpha}$  by  $Z_{\alpha/2}$ .

**Example:** Suppose the standard treatment of care (treatment 2) has a response rate of about .35 (best guess). After collaborations with your clinical colleagues, it is determined that a clinically important difference for a new treatment is an increase in .10 in the response rate. That is, a response rate of .45 or larger. If we are to conduct a clinical trial where we will randomize patients with equal allocation to either the new treatment (treatment 1) or the standard treatment, then how large a sample size is necessary to detect a clinically important difference with 90% power using a one-sided test at the .025 level of significance?

Note for this problem

- ★  $\alpha = .025$ ,  $Z_\alpha = 1.96$
- ★  $\beta = .10$  (power = .9),  $Z_\beta = 1.28$
- ★  $\Delta_A = .10$
- ★  $\pi_2 = .35$ ,  $\pi_1 = .45$ ,  $\bar{\pi} = .40$

Substituting these values into (6.2) we get

$$n = \frac{\left\{ 1.96 + 1.28 \left\{ \frac{.45 \times .55 + .35 \times .65}{2 \times .40 \times .60} \right\}^{1/2} \right\}^2 4 \times .40 \times .60}{(.10)^2} \approx 1,004,$$

or about 502 patients on each treatment arm.

## Arcsine square root transformation

- One **problem** with the above test statistic is that it does not have equal variance under  $\Delta = 0$  and  $\Delta = \Delta_A$ , because  $p_i$ 's variance depends on  $\pi_i$ .
- Variance stabilization transformation:  $p = X/n$ ,  $E(p) = \pi$ ,  $\text{var}(p) = \pi(1 - \pi)/n$ .  
Want to find a monotone function  $g(x)$  such that  $\text{var}(g(p))$  is a constant.
- Using delta method,

$$\text{var}(g(p)) \approx [g'(\pi)]^2 \frac{\pi(1 - \pi)}{n}$$

- If  $g(x)$  satisfies

$$g'(x) = \frac{c}{\sqrt{x(1-x)}},$$

then  $\text{var}(g(p)) \approx \text{a constant}$

- It can be shown that one such  $g(x)$  is given by

$$g(x) = \sin^{-1} \sqrt{x}$$

such that  $\text{var}(g(p)) \approx 1/(4n)$ .

- This  $g(x) = \sin^{-1} \sqrt{x}$  is a monotone function. Therefore,

$$H_0 : \pi_1 = \pi_2 \iff H_0 : \sin^{-1}(\sqrt{\pi_1}) = \sin^{-1}(\sqrt{\pi_2}),$$

and

$$H_A : \pi_1 > \pi_2 \iff H_A : \sin^{-1}(\sqrt{\pi_1}) > \sin^{-1}(\sqrt{\pi_2})$$

- The test statistic testing  $H_0$  would be:

$$T = \frac{\sin^{-1}(\sqrt{p_1}) - \sin^{-1}(\sqrt{p_2})}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)^{1/2}},$$

- By what we derived,

$$T \stackrel{\Delta}{\sim} N(0, 1),$$

and

$$T \stackrel{\Delta}{\sim} N(\phi(n, \Delta_A, \theta), 1),$$

where

$$\phi(n, \Delta_A, \theta) = E_{\Delta=\Delta_A}(T) = \frac{\sin^{-1}(\sqrt{\pi_1}) - \sin^{-1}(\sqrt{\pi_2})}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)^{1/2}}$$

and

$$\Delta_A = \sin^{-1}(\pi_1)^{1/2} - \sin^{-1}(\pi_2)^{1/2}.$$

- With equal allocation,  $n_1 = n_2 = n/2$ , the non-centrality parameter is

$$\phi(n, \Delta_A, \theta) = E_{\Delta=\Delta_A}(T) = \sqrt{n}(\sin^{-1}(\sqrt{\pi_1}) - \sin^{-1}(\sqrt{\pi_2}))$$

In this case, the sample size  $n$  has to satisfy

$$n^{1/2}(\sin^{-1}(\sqrt{\pi_1}) - \sin^{-1}(\sqrt{\pi_2})) = (\mathcal{Z}_\alpha + \mathcal{Z}_\beta),$$

That is,

$$n = \frac{(\mathcal{Z}_\alpha + \mathcal{Z}_\beta)^2}{\Delta_A^2},$$

where

$$\Delta_A = \sin^{-1}(\pi_1)^{1/2} - \sin^{-1}(\pi_2)^{1/2}.$$

**Note:** Replace  $\mathcal{Z}_\alpha$  by  $\mathcal{Z}_{\alpha/2}$  for a two-sided test.

- Going back to our previous example,

$$n = \frac{(1.96 + 1.28)^2}{\{\sin^{-1}(.45)^{1/2} - \sin^{-1}(.35)^{1/2}\}^2} = \frac{(1.96 + 1.28)^2}{(.7353 - .6331)^2} = 1004,$$

the same result. This is because when sample size is this big, normal approximation is pretty good so it does not matter whether or not we do variance stabilization transformation.

# 7 Comparing More Than Two Treatments

Comparing  $K$  treatments, response = binary (1/0)

- Assign  $n_i$  patients to treatment  $i = 1, 2, \dots, K$
- $X_i = \#$  of responses from treatment  $i$ ,

$$X_i \sim \text{bin}(n_i, \pi_i), \quad i = 1, \dots, K.$$

- Wish to test

$$H_0 : \pi_1 = \pi_2 = \dots = \pi_K.$$

- Test statistic has to be based on

$$p_i = X_i/n_i, \quad i = 1, \dots, K.$$

**Question:** How to construct a test statistic for testing  $H_0$ ?

## Testing equality using independent normally distributed estimators

- Assume

$$\hat{\theta}_i \sim N(\theta_i, \sigma_i^2), \quad i = 1, \dots, K$$

where  $\sigma_i^2$  is either known or can be estimated well (consistently) using the data.

- Wish to test

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_K.$$

- Best estimate of common  $\theta$  under  $H_0$ ?

★ Consider the estimate of common  $\theta$  in the following form:

$$\hat{\theta} = \sum_{i=1}^K a_i \hat{\theta}_i.$$

We would like to find the best estimate in this class (in terms of

the first 2 moments)

★ Unbiasedness:

$$\mathbb{E}(\hat{\theta}) = \sum_{i=1}^K a_i \theta = \theta \implies \sum_{i=1}^K a_i = 1.$$

★ Minimum variance:

$$\text{var}(\hat{\theta}) = \sum_{i=1}^K a_i^2 \sigma_i^2$$

Minimizing  $\text{var}(\hat{\theta})$  subject to  $\sum_{i=1}^K a_i = 1 \implies$

$$a_i = \frac{w_i}{\sum_{j=1}^K w_j}, \quad w_i = \frac{1}{\sigma_i^2}.$$

★ The best estimate under  $H_0$ :

$$\hat{\theta} = \frac{\sum_{i=1}^K w_i \hat{\theta}_i}{\sum_{i=1}^K w_i}, \quad w_i = \frac{1}{\sigma_i^2}.$$

- If  $\theta$  were known under  $H_0$ , we might use

$$T = \sum_{i=1}^K w_i (\hat{\theta}_i - \theta)^2$$

to test  $H_0$ . Since  $T \stackrel{H_0}{\sim} \chi_K^2$ , we reject  $H_0$  if  $T \geq \chi_{\alpha, K}^2$ .

- Since  $\theta$  is unknown under  $H_0$ , we replace  $\theta$  by its best estimate under  $H_0$  and get:

$$T = \sum_{i=1}^K w_i (\hat{\theta}_i - \hat{\theta})^2. \quad (7.1)$$

It can be shown that  $T \stackrel{H_0}{\sim} \chi_{K-1}^2$ . Therefore

Reject  $H_0$  if  $T \geq \chi_{\alpha, K-1}^2$ .

- If  $H_0$  is not true (i.e., not all  $\theta_i$ 's are equal), then

$$T = \sum_{i=1}^K w_i (\hat{\theta}_i - \hat{\theta})^2$$

has a non-central  $\chi_{K-1}^2$  distribution with non-centrality parameter equal to

$$\sum_{i=1}^K w_i (\theta_i - \bar{\theta})^2, \quad (7.2)$$

where

$$\bar{\theta} = \frac{\sum_{i=1}^K w_i \theta_i}{\sum_{i=1}^K w_i}.$$

**Note:** The non-centrality parameter is based on the true population parameters.

## Testing equality of dichotomous response rates

- Wish to test:

$$H_0 : \pi_1 = \dots = \pi_K.$$



$$H_0 : \sin^{-1} \sqrt{\pi_1} = \dots = \sin^{-1} \sqrt{\pi_K}.$$

- We have

$$\sin^{-1} \sqrt{p_i} \overset{a}{\sim} N \left( \sin^{-1} \sqrt{\pi_i}, \frac{1}{4n_i} \right).$$

Letting

- ★  $\sin^{-1} \sqrt{p_i}$  take the role of  $\hat{\theta}_i$
- ★  $\sin^{-1} \sqrt{\pi_i}$  take the role of  $\theta_i$
- ★  $\frac{1}{4n_i}$  be  $\sigma_i^2$ ; hence  $w_i = 4n_i$

then by (7.1), the test statistic

$$T_n = \sum_{i=1}^K 4n_i (\sin^{-1} \sqrt{p_i} - \bar{A}_p)^2, \quad (7.3)$$

where

$$\bar{A}_p = \frac{\sum_{i=1}^K n_i \sin^{-1} \sqrt{p_i}}{\sum_{i=1}^K n_i}.$$

- Under  $H_0$ ,  $T \stackrel{a}{\sim} \chi_{K-1}^2$ . So

Reject  $H_0$  if  $T \geq \chi_{\alpha, K-1}^2$ .

## Power and Sample size Calculation with Binary Response

- Under  $H_A : \pi_1 = \pi_{1A}, \dots, \pi_K = \pi_{KA}$ , the test statistic

$$T_n = \sum_{i=1}^K 4n_i (\sin^{-1} \sqrt{p_i} - \bar{A}_p)^2$$

has  $\chi_{K-1}^2$  with non-centrality parameter equal to

$$\phi^2 = \sum_{i=1}^K 4n_i (\sin^{-1} \sqrt{\pi_{iA}} - \bar{A}_{\pi A})^2, \quad (7.4)$$

where

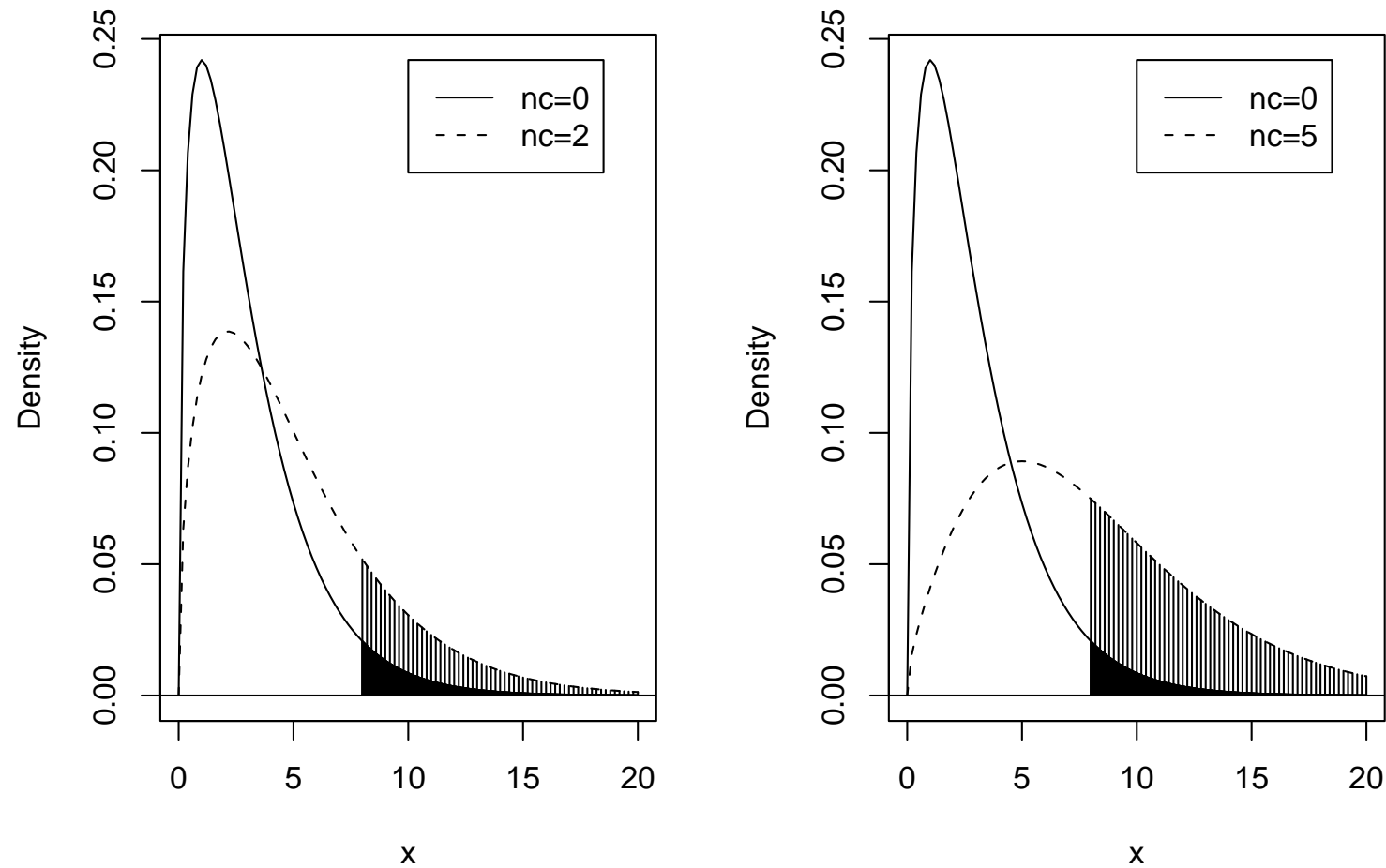
$$\bar{A}_{\pi A} = \frac{\sum_{i=1}^K n_i \sin^{-1} \sqrt{\pi_{iA}}}{\sum_{i=1}^K n_i}.$$

- Suppose we would like to have power  $1 - \beta$  to detect  $H_A$ , how to find  $n_i$ ?
- Assume equal allocation, then  $n_i = n/K$  and

$$\phi^2 = \frac{4n}{K} \left\{ \sum_{i=1}^K (\sin^{-1} \sqrt{\pi_{iA}} - \bar{A}_{\pi_A})^2 \right\}, \quad (7.5)$$

where

$$\bar{A}_{\pi_A} = K^{-1} \sum_{i=1}^K \sin^{-1} \sqrt{\pi_{iA}}.$$

Figure 1: *Distributions of  $T_n$  under  $H_0$  and  $H_A$* 

- ★ Given  $\alpha, \beta$ , define by

$$\phi^2(\alpha, \beta, K - 1),$$

the non-centrality parameter value of a non-central  $\chi_{K-1}^2$  r.v.  $X$  such that

$$P[X \geq \chi_{\alpha, K-1}^2] = 1 - \beta.$$

The values of  $\phi^2(\alpha, \beta, K - 1)$  are given in the class website.

- ★ So for our  $\alpha$  level test to have power  $1 - \beta$  to detect  $H_A$ , its non-centrality parameter has to satisfy:

$$\phi^2 = \frac{4n}{K} \left\{ \sum_{i=1}^K (\sin^{-1} \sqrt{\pi_{iA}} - \bar{A}_{\pi_A})^2 \right\} \geq \phi^2(\alpha, \beta, K - 1).$$

- ★ Solving this we can get the total sample size  $n$ :

$$n = \frac{K \phi^2(\alpha, \beta, K - 1)}{4 \left\{ \sum_{i=1}^K (\sin^{-1} \sqrt{\pi_{iA}} - \bar{A}_{\pi_A})^2 \right\}}. \quad (7.6)$$

Table 1: Non-centrality parameter  $\phi^2(\alpha, \beta, df)$ 

$df$	$\alpha = 0.01$			$\alpha = 0.05$		
	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.05$
1	11.678968	14.879387	17.814164	7.848861	10.507419	12.994709
2	13.8807	17.426689	20.649919	9.634689	12.653936	15.443236
3	15.457657	19.247424	22.674256	10.902563	14.171487	17.169898
4	16.749281	20.736953	24.329097	11.935286	15.405052	18.571649
5	17.869336	22.027506	25.762062	12.827607	16.469464	19.780134
6	18.871852	23.181829	27.043175	13.624286	17.418826	20.857284
7	19.787411	24.235423	28.212039	14.350527	18.283551	21.837879
8	20.63529	25.210653	29.293604	15.022138	19.082696	22.7437
9	21.428557	26.122679	30.304773	15.649798	19.829118	23.589436

## Choosing clinically important alternatives

- ★ Would like to detect  $H_A$  where at least two treatment response rates differ by  $\Delta_A$  (in  $\sin^{-1}(\sqrt{\pi})$  scale).
- ★ The least favorable configuration of the above class of alternatives: smallest  $\phi^2$ , equivalently,

$$\min \sum_{i=1}^K (\sin^{-1} \sqrt{\pi_{iA}} - \bar{A}_{\pi_A})^2$$

subject to

$$\text{at least two } |\sin^{-1} \sqrt{\pi_{iA}} - \sin^{-1} \sqrt{\pi_{jA}}| = \Delta_A.$$

- ★ It can be shown that the above minimum is given by

$$\min \sum_{i=1}^K (\sin^{-1} \sqrt{\pi_{iA}} - \bar{A}_{\pi_A})^2 = \frac{\Delta_A^2}{2}.$$

The sample size for the least favorable configuration:

$$n = \frac{K\phi^2(\alpha, \beta, K - 1)}{2\Delta_A^2}. \quad (7.7)$$

- **Example:** Suppose a standard treatment has a response rate of about .30. Another three treatments have been developed and it is decided to compare all of them in a head to head randomized clinical trial. Equal allocation of patients to the four treatments is used so that approximately  $n_1 = n_2 = n_3 = n_4 = n/4$ . We want the power of a test, at the .05 level of significance, to be at least 90% if any of the other treatments has a response rate greater than or equal to .40. What should we choose as the sample size?

- ★  $\alpha = .05$
- ★  $1 - \beta = .90$ , or  $\beta = .10$
- ★  $K = 4$
- ★  $\Delta_A = \sin^{-1}\sqrt{.40} - \sin^{-1}\sqrt{.30} = .1051$
- ★  $\phi^2(.05, .10, 3) = 14.171$  (derived from the tables provided)

Therefore by (7.7), we get

$$n = \frac{4 \times 14.171}{2(.1051)^2} = 2567,$$

or about  $2567/4 = 642$  patients per treatment arm.

## Multiple comparisons

- After the clinical trial, we test  $H_0 : \pi_1 = \pi_2 = \dots = \pi_K$  using the test statistic developed before.
- If  $H_0$  is not rejected, we usually don't look at individual pair of treatments.
- If  $H_0$  is rejected, we usually want to find out which two treatments are different  $\implies$  multiple comparisons.
- There are  $\binom{K}{2} = K(K-1)/2$  comparisons. If we use the same level  $\alpha$  to compare every two treatments, the overall type I error probability will be greater than  $\alpha$ .

- Suppose we use

$$T_{nij} = \frac{2(\sin^{-1} \sqrt{p_i} - \sin^{-1} \sqrt{p_j})}{\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}}$$

to compare treatment  $i$  and treatment  $j$ . That is

Declare treatments  $i$  and  $j$  are different if  $|T_{nij}| \geq \mathcal{Z}_{\alpha/2}$ .

- Then the overall type I error probability

$$P_{H_0} \left\{ \bigcup_{i < j, i, j = 1, \dots, K} (|T_{nij}| \geq \mathcal{Z}_{\alpha/2}) \right\} > P_{H_{0ij}} (|T_{nij}| \geq \mathcal{Z}_{\alpha/2}) = \alpha.$$

**Solution:**

1. Bonferroni correction: Use level  $2\alpha/\{K(K-1)\}$  for every of  $K(K-1)/2$  comparisons. Then the overall type I error is controlled at  $\alpha$ . That is,

Declare treatments  $i$  and  $j$  are different if  $|T_{nij}| \geq Z_{\alpha/\{K(K-1)\}}$ .

- (a) Bonferroni method can be very conservative (low power).
- (b) In some cases, if we are only interested in certain number (say,  $m$ ) of comparisons, then we can use level  $\alpha/m$  for those comparisons.

2. Hochberg's (*Biometrika*, 1988) approach:

- (a) Order  $m$  ( $m = K(K - 1)/2$  or the number of comparisons of interest) p-values in an ascending order  $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$  and denote the hypothesis corresponding to  $P_{(k)}$  as  $H_{(k)}$ .
- (b) If  $P_{(m)} \leq \alpha$ , then all individual hypotheses are rejected.
- (c) If  $P_{(m)} > \alpha$ ,  $H_{(m)}$  is not rejected; we then compare  $P_{(m-1)}$  to  $\alpha/2$ . If  $P_{(m-1)} \leq \alpha/2$ , then all individual hypotheses  $H_{(1)}, H_{(2)}, \cdots, H_{(m-1)}$  are rejected.
- (d) If  $P_{(m-1)} > \alpha/2$ ,  $H_{(m)}$  and  $H_{(m-1)}$  are not rejected; we then compare  $P_{(m-2)}$  to  $\alpha/3$ . If  $P_{(m-2)} \leq \alpha/3$ , then all individual hypotheses  $H_{(1)}, H_{(2)}, \cdots, H_{(m-2)}$  are rejected.
- (e) The procedure keeps iterating until all remaining hypotheses are rejected or no more hypothesis needs to examine, in which case no individual hypothesis is rejected.

## Alternative Chi-square tests for comparing response rates in $K$ treatments

Testing  $H_0 : \pi_1 = \pi_2 = \dots = \pi_K$  can be done using traditional  $\chi^2$  test for contingency tables ( $2 \times K$  table) using the test statistic

$$\chi^2 = \sum_{\text{cells}} \frac{(O_j - E_j)^2}{E_j}.$$

Reject  $H_0$  if  $\chi^2 \geq \chi_{\alpha, K-1}^2$ .

This will be very close to the test statistic developed using arcsin transformation.

- **Example:** Suppose we have observed data:

Table 2: Observed Counts

Response	<u>Treatment</u>				Total
	1	2	3	4	
yes	206	273	224	275	978
no	437	377	416	364	1594
Total	643	650	640	639	2572

- The expected counts under  $H_0$ :

Table 3: Expected Counts

Response	<u>Treatment</u>				Total
	1	2	3	4	
yes	244.5	247.2	243.4	243	978
no	398.5	402.8	396.6	396	1594
Total	643	650	640	639	2572

The chi-square test is equal to

$$\frac{(206 - 244.5)^2}{244.5} + \frac{(437 - 398.5)^2}{398.5} + \dots + \frac{(364 - 396)^2}{396} = 23.43.$$

$$\text{P-value} = P[\chi_3^2 \geq 23.43] < 0.005$$

- Use the statistic developed here:

<u>Treatment <math>i</math></u>	<u><math>p_i</math></u>	<u><math>\sin^{-1}\sqrt{p_i}</math></u>
1	206/643=.32	.601
2	273/650=.42	.705
3	224/640=.35	.633
4	275/639=.43	.715

$$\bar{A}_p = \frac{643 \times .601 + 650 \times .705 + 640 \times .633 + 639 \times .715}{2572} = .664,$$

and

$$T_n = 4\{643(.601 - .664)^2 + 650(.705 - .664)^2 + 640(.633 - .664)^2 + 639(.715 - .664)^2\} = 23.65.$$

## Example of pairwise comparisons

- Suppose that treatment 1 is standard, treatments 2-4 are new treatments. We only want to see if any of the new treatments are better than the standard one. Therefore,  $m = 3$ . Use  $\alpha = 0.05$  as the overall type I error probability.
- If we use Bonferroni correction, then individual level  $= 0.05/3 = 0.0167$  and  $Z_{.0167/2} = 2.385$ . That is  
Declare treatment  $j$  different from treatment 1 if  $|T_{n1j}| \geq 2.385$ ,  
where

$$T_{n1j} = 2 \left( \frac{n_1 n_j}{n_1 + n_j} \right)^{1/2} (\sin^{-1} \sqrt{p_j} - \sin^{-1} \sqrt{p_1}).$$

Substituting the data above we get

$$T_{n12} = 2 \left( \frac{643 \times 650}{643 + 650} \right)^{1/2} (.705 - .601) = 3.73*$$

$$T_{n13} = 2 \left( \frac{643 \times 640}{643 + 640} \right)^{1/2} (.633 - .601) = 1.14$$

$$T_{n14} = 2 \left( \frac{643 \times 639}{643 + 639} \right)^{1/2} (.715 - .601) = 4.08*$$

- Hochberg's approach:

$$P_{1,2} = 2P[Z \geq 3.73] = 0.0001914798,$$

$$P_{1,3} = 2P[Z \geq 1.14] = 0.2542863,$$

$$P_{1,4} = 2P[Z \geq 4.08] = 0.000045.$$

So  $P_{1,4} < P_{1,2} < P_{1,3}$

★ Since  $P_{1,3} > \alpha = 0.05$ , we don't reject  $H_{1,3}$ .

★ Since  $P_{1,2} < \alpha/2 = 0.025$ , we reject  $H_{1,2}$  and  $H_{1,4}$ .

## K-sample tests for continuous response

- Data  $(Y_i, A_i), i = 1, \dots, n, A_i = 1, 2, \dots, K$  is treatment indicator:

$$E(Y_i | A_i = j) = \mu_j, \quad j = 1, \dots, K$$

and

$$\text{var}(Y_i | A_i = j) = \sigma_{Y_j}^2, \quad j = 1, \dots, K.$$

**Note:**

1. Often, we assume  $\sigma_{Y_1}^2 = \dots = \sigma_{Y_K}^2 = \sigma_Y^2$ ; but this assumption is not necessary.
2. It is also often assumed that

$$(Y_i | A_i = j) \sim N(\mu_j, \sigma_Y^2), \quad j = 1, \dots, K.$$

Again, this assumption is not necessary.

- We wish to test

$$H_0 : \mu_1 = \dots = \mu_K.$$

- We know that the treatment-specific sample mean

$$\bar{Y}_j = \sum_{i=1}^{n_j} Y_{ij} / n_j$$

is an unbiased estimator for  $\mu_j$  and that asymptotically

$$\bar{Y}_j \sim N\left(\mu_j, \frac{\sigma_{Y_j}^2}{n_j}\right), \quad j = 1, \dots, K$$

and  $\sigma_{Y_j}^2$  can be well-estimated by treatment-specific sample variance:

$$s_{Y_j}^2 = \frac{\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{n_j - 1}.$$

If we assume equal variance, then the common variance can be estimated by the pooled sample variance:

$$s_Y^2 = \frac{\sum_{j=1}^K \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{n - K}.$$

- Let
  - ★  $\bar{Y}_j$  take the role of  $\hat{\theta}_j$
  - ★  $\mu_j$  take the role of  $\theta_j$
  - ★  $\frac{s_{Y_j}^2}{n_j}$  take the role of  $\sigma_j^2$ ; hence  $w_j = \frac{n_j}{s_{Y_j}^2}$
- Using (7.1), we construct the test statistic

$$T_n = \sum_{j=1}^K w_j (\bar{Y}_j - \bar{\bar{Y}})^2, \quad (7.8)$$

where

$$\bar{\bar{Y}} = \frac{\sum_{j=1}^K w_j \bar{Y}_j}{\sum_{j=1}^K w_j},$$

and  $w_j = \frac{n_j}{s_{Y_j}^2}$ .

- If we assume equal variance, then we would use

$$T_n = \frac{\sum_{j=1}^K n_j (\bar{Y}_j - \bar{\bar{Y}})^2}{s_Y^2}, \quad (7.9)$$

where

$$\bar{\bar{Y}} = \frac{\sum_{j=1}^K n_j \bar{Y}_j}{n}.$$

- Under  $H_0$ ,  $T_n \stackrel{a}{\sim} \chi_{K-1}^2$ . So reject  $H_0$  if  $T_n \geq \chi_{\alpha, K-1}^2$ .

- Under  $H_A : \mu_1 = \mu_{1A}, \dots, \mu_K = \mu_{KA}$ ,  $T_n$  is approximately distributed as a non-central  $\chi_{K-1}^2$  with non-centrality parameter

$$\phi^2 = \sum_{j=1}^K w_j (\mu_{jA} - \bar{\mu}_A)^2, \quad (7.10)$$

where

$$\bar{\mu}_A = \frac{\sum_{j=1}^K w_j \mu_{jA}}{\sum_{j=1}^K w_j},$$

and  $w_j = \frac{n_j}{\sigma_{Y_j}^2}$ .

The non-centrality parameter can be simplified if equal variance is assumed:

$$\phi^2 = \frac{\sum_{j=1}^K n_j (\mu_{jA} - \bar{\mu}_A)^2}{\sigma_Y^2}, \quad (7.11)$$

where

$$\bar{\mu}_A = \frac{\sum_{j=1}^K n_j \mu_{jA}}{n}.$$

**Note:**  $T_n$  given by (7.9) =  $(K - 1) \times F$ , where  $F$  is the  $F$  statistic in one-way ANOVA table for comparing  $K$  treatments.

When sample size gets larger, the denominator of  $F$  converges to  $\sigma_Y^2$ .

## Sample size computations for continuous response

- Assume equal allocation:  $n_1 = n_2 = \dots = n_K = n/K$ , clinically important difference in means  $\Delta_A$ . Then the non-centrality parameter of our test statistic for the least favorable configuration:

$$\phi^2 = \frac{n\Delta_A^2}{2K\sigma_Y^2}. \quad (7.12)$$

- Type I error probability  $\alpha$ .
- Power  $1 - \beta$  to detect the alternative that at least two treatment means differ by  $\Delta_A$ .

- The non-centrality parameter of the test statistic has to satisfy

$$\frac{n\Delta_A^2}{2K\sigma_Y^2} = \phi^2(\alpha, \beta, K - 1),$$

or

$$n = \frac{2K\sigma_Y^2\phi^2(\alpha, \beta, K - 1)}{\Delta_A^2}. \quad (7.13)$$

- **Example:**

We expand on the example used for two-sample comparisons given on slide 216, but now we consider  $K = 4$  treatments. What is the sample size necessary to detect a significant difference with 90% power or greater if any pairwise difference in mean treatment response is at least 20 units using the K-sample test above at the .05 level of significance? We posit that the standard deviation of response, assumed equal for all treatments, is  $\sigma_Y = 60$  units. Substituting into formula (7.13), we get that

$$n = \frac{2 \times 4 \times (60)^2 \times 14.171}{(20)^2} \approx 1020,$$

or about  $1021/4=255$  patients per treatment arm.

## Non-inferiority Trials

- Standard treatment in the market (treatment 2, response rate  $\pi_2$ )
- Want to show the new treatment (treatment 1) may be little bit worse than the standard treatment but within our tolerance limit  $\Delta_A$ .
- Therefore, our hypothesis testing problem is:

$$H_0 : \pi_1 \leq \pi_2 - \Delta_A \text{ versus } H_A : \pi_1 > \pi_2 - \Delta_A.$$

- The test statistic:

$$T_n = \frac{p_1 - p_2 + \Delta_A}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}},$$

where  $n_1$  and  $n_2$  denote the number of patients allocated to treatments 1 and 2 respectively.

- On  $H_0 \cap H_A$ , i.e.,  $\pi_1 = \pi_2 - \Delta_A$ ,

$$T_n \stackrel{(\pi_1 = \pi_2 - \Delta_A)}{\sim} N(0, 1).$$

So for the given level  $\alpha$ , we reject  $H_0$  if

$$T_n \geq Z_\alpha.$$

Using this strategy, the new drug will not be approved with high probability ( $\geq 1 - \alpha$ ) when in fact it is worse than the standard treatment by at least  $\Delta_A$ .

**Remark:** we didn't use the arcsin square-root transformation here. Because the arcsin square-root is non-linear; thus, a fixed difference of  $\Delta_A$  in response probabilities between two treatments (hypothesis of interest) does not correspond to a fixed difference on the arcsin square-root scale.

## Sample size calculations for non-inferiority trials

- We usually want to have high power to detect if the new drug is at least as good as the standard treatment. That is, the power of our test is calculated at  $\pi_1 = \pi_2$ .
- Under  $\pi_1 = \pi_2$ , our test statistic

$$T_n \stackrel{a}{\sim} N(\phi, 1),$$

where

$$\phi = E(T_n) \stackrel{(\pi_1 = \pi_2 = \pi)}{\approx} \frac{\Delta_A}{\sqrt{\pi(1 - \pi) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

- If we do equal allocation,  $n_1 = n_2 = n/2$ , then

$$\phi = \frac{\Delta_A}{\sqrt{\pi(1 - \pi) \left( \frac{4}{n} \right)}}.$$

- In order to have power  $1 - \beta$  to detect that the new drug is at least as good as the standard treatment, the non-centrality parameter  $\phi$  has to satisfy

$$\frac{\Delta_A}{\sqrt{\pi(1 - \pi) \left(\frac{4}{n}\right)}} = Z_\alpha + Z_\beta$$

or

$$n = \frac{(Z_\alpha + Z_\beta)^2 \times 4\pi(1 - \pi)}{\Delta_A^2}. \quad (7.14)$$

**Note:** Since usually  $\Delta_A$  is very small, the sample size will be much larger than the ones from superiority trials.

- **Example:** Suppose the response rate of the standard treatment is about 30% and a 95% CI of the treatment effect (difference of response rates between the standard treatment and placebo) in a clinical trial is [0.1, 0.25], and we want to show a new treatment is not inferior to the standard one.

- ★  $\Delta_A = 0.1/2 = 0.05$

- ★  $\alpha = 0.05$ , so  $Z_{0.05} = 1.64$

- ★ Good power:  $1 - \beta = 0.9$ ,  $\beta = 0.1$ ,  $Z_{0.1} = 1.28$

- ★  $\pi_1 = 0.3$ .

- The total sample size  $n$ :

$$n = \frac{(1.64 + 1.28)^2 \times 4 \times .3 \times .7}{(.05)^2} = 2864,$$

or 1432 patients per treatment arm.

# 8 Causality, Non-compliance and Intent-to-treat

## 8.1 Causality and Counterfactual Random Variables

- **Two treatments:** treatment 1 (new treatment), treatment 0 (standard treatment)
- For each individual, we can imagine 2 variables ( $Y_0^*$ ,  $Y_1^*$ ):
  - ★  $Y_0^*$ : outcome if the individual takes treatment 0
  - ★  $Y_1^*$ : outcome if the individual takes treatment 1
- $U = Y_1^* - Y_0^*$  = causal effect (treatment 1 over treatment 0).
- Since  $Y_0^*$  and  $Y_1^*$  cannot be observed simultaneously, the individual level causal effect  $U$  is not observable!

- Concede to consider

$$\Delta = \mathbf{E}[Y_1^* - Y_0^*] = \mathbf{E}[Y_1^*] - \mathbf{E}[Y_0^*]$$

which is the **population** level causal effect.

- If covariate information  $x$  (age, sex, etc.) is available, we may consider

$$\Delta(x) = \mathbf{E}[Y_1^* - Y_0^* | X = x] = \mathbf{E}[Y_1^* | X = x] - \mathbf{E}[Y_0^* | X = x],$$

which is the causal effect for the population with  $X = x$ .

- Given data  $(Y_i, A_i, X_i), i = 1, \dots, n$ , can we estimate  $\Delta$  or  $\Delta(x)$ ?
  - ★  $Y_i =$  response
  - ★  $A_i =$  treatment indicator
  - ★  $X_i =$  covariates
- Assume  $Y_i$  is related to the potential outcomes  $Y_{1i}^*, Y_{0i}^*$  of individual  $i$  as follows:

$$Y_i = Y_{1i}^* I(A_i = 1) + Y_{0i}^* I(A_i = 0), \quad (8.1)$$

where  $I(E) = 1$  if  $E$  occurs, otherwise  $I(E) = 0$ .

- Under randomization, we can estimate  $\Delta$  or  $\Delta(x)$ !
- Randomization  $\implies$

$$A_i \perp Y_{1i}^*, Y_{0i}^*, X_i$$

**Note:**  $A_i$  is **not** independent of  $Y_i$  (value after randomization).

All  $Y_{1i}^*, Y_{0i}^*, X_i$  are considered pre-randomization random variables.

- Therefore,

$$\begin{aligned} F_{Y_1^*}(y) &= P[Y_{1i}^* \leq y] \\ &= P[Y_{1i}^* \leq y | A_i = 1] \quad (A_i \perp Y_{1i}^*) \\ &= P[Y_i \leq y | A_i = 1], \end{aligned}$$

which can be estimated using  $Y_i$  from individuals whose  $A_i = 1$ .

- In particular,

$$E(Y_1^*) = E[Y_i | A_i = 1],$$

which can be estimated using the sample mean from treatment 1.

- Similarly,

$$\begin{aligned}F_{Y_0^*}(y) &= P[Y_{0i}^* \leq y] \\ &= P[Y_i \leq y | A_i = 0],\end{aligned}$$

which can be estimated using  $Y_i$  from individuals whose  $A_i = 0$ .

- And

$$E(Y_0^*) = E[Y_i | A_i = 0],$$

which can be estimated using the sample mean from treatment 0.

- Therefore, the causal effect  $\Delta$  can be estimated by

$$\hat{\Delta} = \bar{Y}_1 - \bar{Y}_0,$$

where  $\bar{Y}_1$  and  $\bar{Y}_0$  are sample means for treatments 1 and 0.

- Similarly

$$\begin{aligned}F_{Y_{1i}^*|x}(y) &= P[Y_{1i}^* \leq y | X_i = x] \\&= P[Y_{1i}^* \leq y | A_i = 1, X_i = x] \quad (A_i \perp Y_{1i}^*, X_i) \\&= P[Y_i \leq y | A_i = 1, X_i = x],\end{aligned}$$

and

$$F_{Y_{0i}^*|x}(y) = P[Y_i \leq y | A_i = 0, X_i = x]$$

- Therefore

$$\Delta(x) = E[Y_i | A_i = 1, X = x] - E[Y_i | A_i = 0, X = x]$$

can be estimated from data.

- Of course, we can assume a model such as

$$Y_i = \beta_0 + A_i\beta_1 + X_i\beta_2 + A_iX_i\beta_3 + \epsilon_i,$$

to estimate the above conditional expectations.

In this case

$$\mathbb{E}[Y_i | A_i = 1, X = x] = \beta_0 + \beta_1 + x\beta_2 + x\beta_3$$

$$\mathbb{E}[Y_i | A_i = 0, X = x] = \beta_0 + x\beta_2,$$

so

$$\Delta(x) = \beta_1 + x\beta_3$$

and its estimate is

$$\hat{\Delta}(x) = \hat{\beta}_1 + x\hat{\beta}_3.$$

## 8.2 Noncompliance and Intent-to-treat analysis

- Form of noncompliance
  - ★ A refusal by the patient to start or continue the assigned treatment, perhaps because of side effects or a belief that the treatment is ineffective
  - ★ A failure to comply with detailed instructions, for example, drug dosage, or to attend examinations when requested to do so
  - ★ A change of treatment imposed by the physician for clinical reasons, usually occurrence of adverse effects or deterioration of patient's health
  - ★ An administrative error. In its most extreme form this may be the implementation of the wrong treatment.

- Two analysis approaches when there is noncompliance:
  - ★ **Intent-to-Treat Analysis** (As randomized)  
Everyone is included in the analysis and the comparison of treatments is based on the difference of the average response between the randomized groups ignoring the fact that some patients were non-compliant.
  - ★ **As-treated analysis**  
This type of analysis takes on various forms, but the general idea is to compare only those patients who fully complied with their assigned treatment regimen and exclude non compliers from the analysis.

## General Dogma of Clinical Trials

The exclusion of patients from the analysis should not allow the potential of bias in the treatment comparisons. Thus, exclusions based on post-randomization considerations, such as noncompliance, are **not** allowed for the primary analysis.

- **Example:** Coronary Drug Project (*New England Journal of Medicine*, October, 1980, 303: 1038-1041): a double-blind placebo-controlled clinical trial comparing Clofibrate to Placebo.

Table 1: *Intent-to-Treat Analysis*

	<u>Clofibrate</u>	<u>Placebo</u>
5 year mortality rate	.18	.19
number of patients	1065	2695

Table 2: *Clofibrate Patients Only*

<u>Adherence</u> <u>(% of capsules taken)</u>	<u>5 year</u> <u>mortality rate</u>	<u>Number</u> <u>patients</u>
Poor (< 80%)	.25	357
Good (> 80%)	.15	708

p-value=.001

Table 3: *Clofibrate and Placebo Patients*

	Clofibrate		Placebo	
<u>Adherence</u>	<u>5 year mortality</u>	<u>number patients</u>	<u>5 year mortality</u>	<u>number patients</u>
Poor (< 80%)	.25	357	.28	882
Good (> 80%)	.15	708	.15	1813

- Implication: compliers and non-compliers are prognostically different!

## 8.3 A Causal Model with Noncompliance

- Simple example:
  - ★ Patients are randomized with equal probability to active drug (treatment 1) or placebo (control) (treatment 0)
  - ★ Response is dichotomous; i.e. a patient either responds or not
  - ★ The main goal of the clinical trial is to estimate the difference in the probability of response between active drug and placebo
  - ★ Patients may not comply with their assigned treatment
  - ★ For simplicity, we assume that everyone either takes their assigned treatment or not (partial compliance is not considered)
  - ★ A simple assay can be conducted on patients that were randomized to receive active drug to see if they complied or not
  - ★ Patients assigned to placebo have no access to the study drug
  - ★ Compliance cannot be determined for patients randomized to placebo

- **Counterfactual and observable random variables**
  - ★  $(Y_1^*, Y_0^*)$ , responses if received and took treatment 1 or 0.
  - ★  $C$  compliance indicator (1/0) if offered new treatment (treatment 1); COM = complier, NC = Non-complier
- **Assume**
  - ★  $\theta = P(C = 1)$  denotes the population probability of complying
  - ★  $\pi_1^{COM} = P(Y_1^* = 1|C = 1)$  denotes the population probability of response among compliers if given active drug,
  - ★  $\pi_1^{NC} = P(Y_1^* = 1|C = 0)$  denotes the population probability of response among noncompliers if given active drug
  - ★  $\pi_0^{COM} = P(Y_0^* = 1|C = 1)$  denotes the population probability of response among compliers if given placebo
  - ★  $\pi_0^{NC} = P(Y_0^* = 1|C = 0)$  denotes the population probability of response among noncompliers if given placebo

Table 4: *Hypothetical Population*

	COMPLIERS	NONCOMPLIERS
	$\theta$	$(1 - \theta)$
Treatment	$\pi_1^{COM}$	$\pi_1^{NC}$
Placebo	$\pi_0^{COM}$	$\pi_0^{NC}$
Difference	$\Delta^{COM}$	$\Delta^{NC}$

**Note:**  $(Y_1^*, Y_0^*)$  may not be independent of  $C$ . Thus, we would not expect

$$\pi_1^{COM} = \pi_1^{NC}$$

or

$$\pi_0^{COM} = \pi_0^{NC}.$$

- Total probability law  $\implies$

$$\begin{aligned}\pi_1 &= E(Y_1^*) = P(Y_1^* = 1) \\ &= P(Y_1^* = 1|C = 1)P(C = 1) \\ &\quad + P(Y_1^* = 1|C = 0)P(C = 0) \\ &= \pi_1^{COM}\theta + \pi_1^{NC}(1 - \theta),\end{aligned}$$

and

$$\begin{aligned}\pi_0 &= E(Y_0^*) = P(Y_0^* = 1) \\ &= P(Y_0^* = 1|C = 1)P(C = 1) \\ &\quad + P(Y_0^* = 1|C = 0)P(C = 0) \\ &= \pi_0^{COM}\theta + \pi_0^{NC}(1 - \theta),\end{aligned}$$

- Causal effect:

$$\Delta = \pi_1 - \pi_0 = \Delta^{COM}\theta + \Delta^{NC}(1 - \theta).$$

- Can we estimate  $\Delta$ ?
- Observable  $Y$  is related to  $Y_1^*$ ,  $Y_0^*$ ,  $C$  and  $A$  as follows:

$$Y = Y_0^* I(A = 0) + Y_1^* I(A = 1, C = 1) + Y_0^* I(A = 1, C = 0).$$

- Randomization  $\implies$

$$A \perp (Y_1^*, Y_0^*, C).$$

- Therefore,

$$P(C = 1|A = 1) = P(C = 1) = \theta$$

$$P(Y = 1|A = 0) = P(Y_0^* = 1|A = 0) = P(Y_0^* = 1) = \pi_0$$

$$P(Y = 1|A = 1, C = 1) = P(Y_1^* = 1|A = 1, C = 1)$$

$$= P(Y_1^* = 1|C = 1) = \pi_1^{COM}$$

$$P(Y = 1|A = 1, C = 0) = P(Y_0^* = 1|A = 1, C = 0)$$

$$= P(Y_0^* = 1|C = 0) = \pi_0^{NC}$$

- We can estimate  $\theta$ ,  $\pi_0$ ,  $\pi_1^{COM}$  and  $\pi_0^{NC}$ .
- Since

$$\pi_0 = \pi_0^{COM} \theta + \pi_0^{NC} (1 - \theta),$$

so

$$\pi_0^{COM} = \frac{\pi_0 - \pi_0^{NC} (1 - \theta)}{\theta},$$

is also estimable.

- The only parameter that we cannot estimate is  $\pi_1^{NC}$ .

## Intent-to-treat analysis

- Estimate treatment effect by

$$\hat{\Delta}_{ITT} = \bar{Y}_1 - \bar{Y}_0.$$

- $\hat{\Delta}_{ITT}$  estimate  $P(Y = 1|A = 1) - P(Y = 1|A = 0)$ , where

$$\begin{aligned} & P(Y = 1|A = 1) \\ = & P(Y = 1|A = 1, C = 1)P(C = 1|A = 1) \\ & + P(Y = 1|A = 1, C = 0)P(C = 0|A = 1) \\ = & P(Y_1^* = 1|A = 1, C = 1)P(C = 1|A = 1) \\ & + P(Y_0^* = 1|A = 1, C = 0)P(C = 0|A = 1) \\ = & P(Y_1^* = 1|C = 1)P(C = 1) \\ & + P(Y_0^* = 1|C = 0)P(C = 0) \\ = & \pi_1^{COM} \theta + \pi_0^{NC} (1 - \theta), \end{aligned}$$

and

$$\begin{aligned}
 & P(Y = 1|A = 0) \\
 &= P(Y_0^* = 1|A = 0) \\
 &= P(Y_0^* = 1) = \pi_0 = \pi_0^{COM}\theta + \pi_0^{NC}(1 - \theta).
 \end{aligned}$$

- So  $\hat{\Delta}_{ITT}$  estimates

$$\begin{aligned}
 & \pi_1^{COM}\theta + \pi_0^{NC}(1 - \theta) - \{\pi_0^{COM}\theta + \pi_0^{NC}(1 - \theta)\} \\
 &= \theta(\pi_1^{COM} - \pi_0^{COM}) = \theta\Delta^{COM},
 \end{aligned}$$

some fraction of the complier causal effect.

- The complier causal effect can be estimated by

$$\frac{\bar{Y}_1 - \bar{Y}_0}{\hat{\theta}}.$$

## Remarks

- ★ If the null hypothesis of no treatment effect is true; namely

$$H_0 : \Delta^{COM} = \Delta^{NC} = \Delta = 0$$

- \* The intent to treat analysis, which estimates  $(\Delta^{COM}\theta)$ , gives an unbiased estimator of treatment difference (under  $H_0$ ) and can be used to compute a valid test of the null hypothesis
- ★ If we were interested in estimating the causal parameter  $\Delta^{COM}$ , the difference in response rate between treatment and placebo among compliers only, then
  - \* the intent to treat analysis gives an underestimate of this population causal effect
- ★ Since there are no data available to estimate  $\pi_1^{NC}$ , we are not able to estimate  $\Delta^{NC}$  or  $\Delta$

## As-treated analysis

- One version uses:

$$\widehat{\Delta}_{AT} = \bar{Y}_{A=1, C=1} - \bar{Y}_{A=0}.$$

- This  $\widehat{\Delta}_{AT}$  estimates

$$\begin{aligned} & P[Y = 1|A = 1, C = 1] - P[Y = 1|A = 0] \\ = & P[Y_1^* = 1|A = 1, C = 1] - P[Y_0^* = 1|A = 0] \\ = & P[Y_1^* = 1|C = 1] - P[Y_0^* = 1] \\ = & \pi_1^{COM} - \pi_0 \\ = & \theta\pi_1^{COM} + (1 - \theta)\pi_1^{COM} - \pi_0 \\ = & \pi_1 - (1 - \theta)\pi_1^{NC} + (1 - \theta)\pi_1^{COM} - \pi_0 \\ = & \Delta + (1 - \theta)(\pi_1^{COM} - \pi_1^{NC}). \end{aligned}$$

- As indicated by the example,  $\pi_1^{COM}$  and  $\pi_1^{NC}$  are usually not the same, so as-treated analysis is **biased!**
- **Some Additional Remarks about Intention-to-Treat (ITT) Analyses**
  - ★ By not allowing any exclusions, we are preserving the integrity of randomization
  - ★ With the use of **ITT**, we are comparing the policy of using treatment A where possible to the policy of using treatment B (control) where possible
  - ★ If the intended treatments are always used, there is of course no problem
  - ★ If the treatments are rarely used, then the clinical trial will carry little information about the true effect of A versus B, but a great deal of information about the difficulties to use them
  - ★ The approach of comparing policies of intentions rather than rigorously standardized regimens may be a more realistic

statement of the purpose of the investigation

- \* This is the pragmatic approach to a clinical trial
- \* As compared to the explanatory approach which looks for a measure of effectiveness rather than efficacy.
- ★ The estimate of the causal effect  $\Delta^{COM}$  is larger than the intent-to-treat estimator  $\Delta_{ITT}$ , but it also has proportionately larger standard deviation. Thus use of this estimator as a basis for a test of the null hypothesis yields the same significance level as a standard intent-to-treat analysis

# 9 Survival Analysis in Phase III Clinical Trials

**Primary endpoint:** Time to an event

- survival time (time from birth to death)
- time from treatment of lung cancer to death among patients with lung cancer
- among patients with an infection that are treated with an antibiotic, the time from treatment until eradication of infection

## 9.1 Describing the Distribution of Time to Event

$T$  : Time to an event

- The cumulative distribution function (CDF):

$$F(t) = P(T \leq t) = \int_0^t f(u)du$$

- The survival function:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(u)du$$

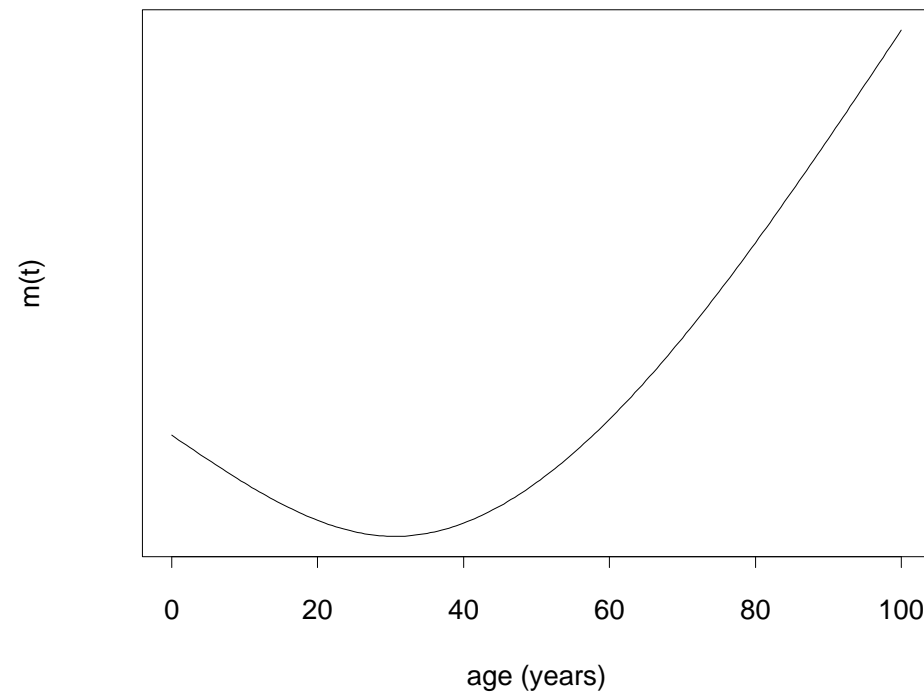
$S(t)$  = the probability a (random) patient will still be alive at time  $t$ .

- **Hazard rate**

- ★ mortality rate at time  $t$ :

$$m(t) = P(t \leq T < t + 1 | T \geq t)$$

Figure 1: *A typical mortality pattern for human*



★ Definition of hazard function  $\lambda(t)$ :

$$\begin{aligned}\lambda(t) &= \lim_{h \rightarrow 0} \left\{ \frac{P(t \leq T < t + h | T \geq t)}{h} \right\} \\ &= \lim_{h \rightarrow 0} \left\{ \frac{P(t \leq T < t + h)/h}{P(T \geq t)} \right\} = \frac{f(t)}{S(t)} \\ &= \frac{-\frac{dS(t)}{dt}}{S(t)} = \frac{-d \log\{S(t)\}}{dt}\end{aligned}$$

$\implies$

$$-\log\{S(t)\} = \int_0^t \lambda(u) du = \Lambda(t) \text{ (Cumulative hazard)}$$

$$\begin{aligned}S(t) &= \exp \left\{ - \int_0^t \lambda(u) du \right\} \\ &= \exp \{-\Lambda(t)\}.\end{aligned}$$

- ★ The mortality rate is related to the hazard rate (function):

$$\begin{aligned} m(t) &= \frac{P(T \geq t) - P(T \geq t + 1)}{P(T \geq t)} \\ &= 1 - \frac{P(T \geq t + 1)}{P(T \geq t)} \\ &= 1 - \frac{\exp\{-\Lambda(t + 1)\}}{\exp\{-\Lambda(t)\}} \\ &= 1 - \exp\left\{-\int_t^{t+1} \lambda(u) du\right\}. \end{aligned}$$

- ★ If  $\lambda(u)$  does not change too much and is small in  $[t, t + 1)$ , then

$$\begin{aligned} m(t) &= 1 - \exp\left\{-\int_t^{t+1} \lambda(u) du\right\} \approx 1 - \left\{1 - \int_t^{t+1} \lambda(u) du\right\} \\ &= \int_t^{t+1} \lambda(u) du \approx \lambda(t). \end{aligned}$$

- Therefore, the distribution of  $T$  can be described by any one of the following:

$$S(t), F(t), f(t), \lambda(t).$$

- **Exponential distribution:**

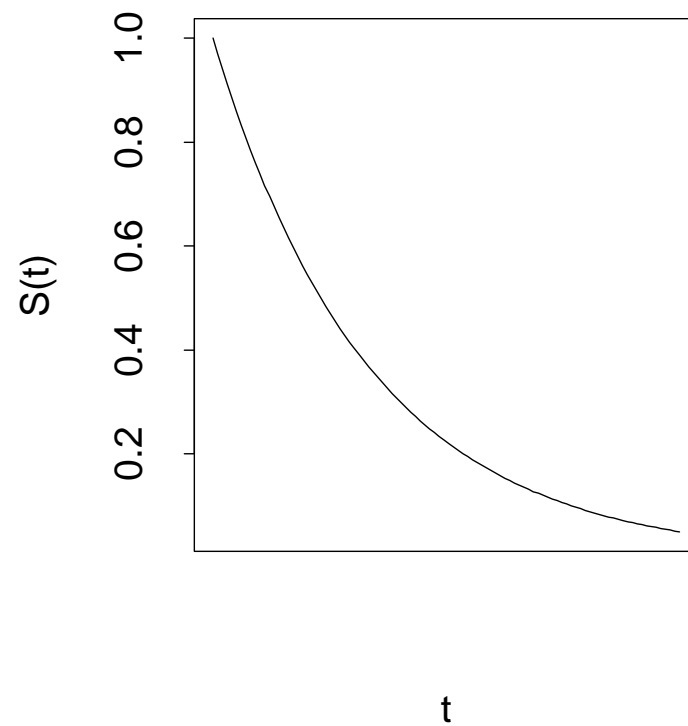
$$\lambda(t) = \lambda$$

$$S(t) = \exp \left\{ - \int_0^t \lambda(u) du \right\} = \exp(-\lambda t)$$

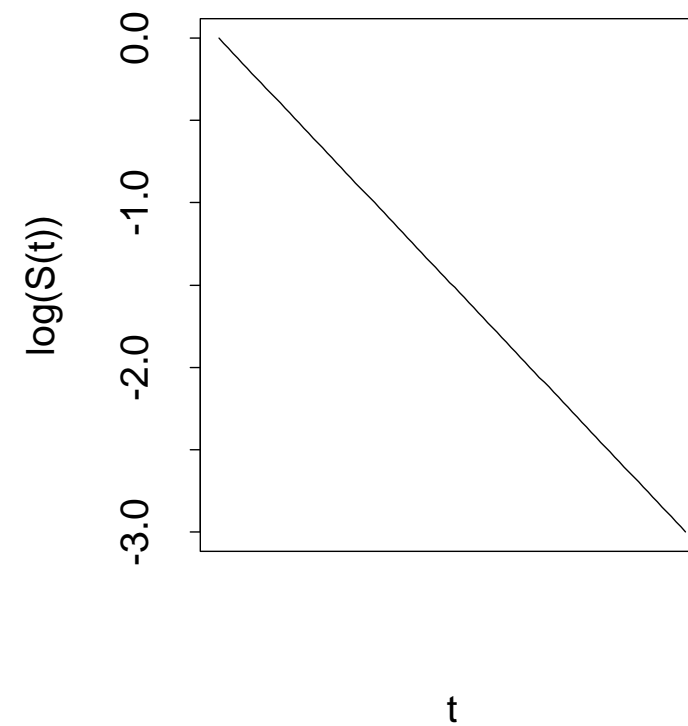
$$\log\{S(t)\} = -\lambda t$$

Figure 2: *The survival function of an exponential distribution on two scales*

Survival function on original scale



Survival function on a log scale



$T$  follows an exponential distribution with hazard rate  $\lambda$ , then

★ the median survival time  $m$ :

$$P(T \geq m) = .5$$

$$\implies \exp(-\lambda m) = .5$$

$$m = -\log(.5)/\lambda = \log(2)/\lambda = .6931/\lambda$$

★ the mean survival time

$$E(T) = \int_0^{\infty} t\lambda \exp(-\lambda t) dt = \lambda^{-1}.$$

- **Weibull distribution:**

$$\lambda(t) = \lambda t^{\gamma-1}; \lambda, \gamma > 0$$

$$S(t) = \exp\left(-\frac{\lambda t^\gamma}{\gamma}\right)$$

- **Gompertz-Makeham distribution** (good for modeling the hazard function of human populations especially later in life):

$$\lambda(t) = \theta + \beta e^{\gamma t}$$

$\theta, \beta, \gamma$  need to be chosen in such way that  $S(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

- **Log-normal distribution:**

$$\log(T) \sim N(\mu, \sigma^2)$$

- **Gamma distribution:**

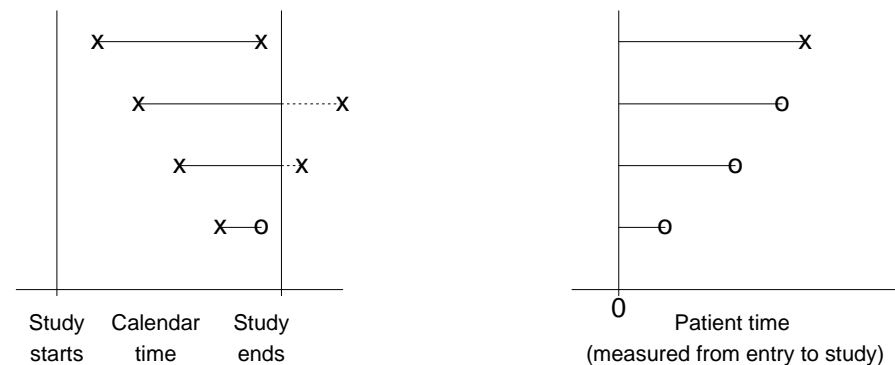
$f(t)$  is proportional to  $t^\rho e^{-\lambda t}$

## 9.2 Censoring and Life-Table Methods

Two issues:

1. Some individuals are still alive (event of interest has not occurred) at the time of analysis. This results in right censored data.
2. The length of follow-up varies due to staggered entry.

Figure 3: *Illustration of censored data*



In addition to censoring occurring because of insufficient follow-up, it may also occur for other reasons such as

- loss to follow-up (patient drops out of the study, stops coming to the clinic or moves away)
- death from other causes  
(competing risks; e.g. gets run over by a bus)

### **Right random censoring**

- **Example.** 146 patients with previous history of heart disease treated with a new anti-hypertensive drug. The study was carried out over a ten year period and the data are grouped into one year intervals.

---

Year since entry	Number at risk at beginning	Number dying in interval	Number cens. in interval
0-1	146	27	3
1-2	116	18	10
2-3	88	21	10
3-4	57	9	3
4-5	45	1	3
5-6	41	2	11
6-7	28	3	5
7-8	20	1	8
8-9	11	2	1
9-10	8	2	6

---

**Objective:** estimate  $S(5) = P[T \geq 5]$ .

Two estimates:

1.  $\hat{F}(5) = \frac{76 \text{ deaths in } [0,5]}{146 \text{ individuals}} = .521$ ,  $\hat{S}(5) = 1 - \hat{F}(5) = .479$
2.  $\hat{F}(5) = \frac{76 \text{ deaths in } [0,5]}{146 - 29 \text{ (withdrawn)}} = .650$ ,  $\hat{S}(5) = 1 - \hat{F}(5) = .350$ .

**Problem** with above two estimates:

1. The first  $\hat{F}(5)$  is too small, so  $\hat{S}(5) = 0.479$  is too optimistic.
2. The second  $\hat{F}(5)$  is too big (it treats un-censored patients as a random sample), so  $\hat{S}(5) = 1 - \hat{F}(5) = 0.350$  is too pessimistic.

## Better method?

$$\begin{aligned}
 S(5) &= P[T \geq 5] \\
 &= P[(T \geq 5) \cap (T \geq 4)] \\
 &= P[T \geq 5|T \geq 4]P[T \geq 4] \\
 &= (1 - P[T < 5|T \geq 4])P[T \geq 4] \\
 &= (1 - P[4 \leq T < 5|T \geq 4])P[T \geq 4] \\
 &= \dots \\
 &= (1 - P[4 \leq T < 5|T \geq 4])(1 - P[3 \leq T < 4|T \geq 3]) \\
 &\quad (1 - P[2 \leq T < 3|T \geq 2])(1 - P[1 \leq T < 2|T \geq 1]) \\
 &\quad (1 - P[0 \leq T < 1|T \geq 0]) \\
 &= \{1 - m(0)\}\{1 - m(1)\}\{1 - m(2)\}\{1 - m(3)\}\{1 - m(4)\}.
 \end{aligned}$$

So if we can estimate  $m(i)$ 's well, we can estimate  $S(5)$  well.  $\implies$   
 Life-table estimate of  $S(5)$ .

Assume censoring occurs at the right of each yearly interval

Time	$n_r$	d	w	$m^R = d/n_r$	$1 - m^R$	$\hat{S}^R = \prod(1 - m^R)$
0-1	146	27	3	.185	.815	.815
1-2	116	18	10	.155	.845	.689
2-3	88	21	10	.239	.761	.524
3-4	57	9	3	.158	.842	.441
4-5	45	1	3	.022	.978	.432

5 year survival probability estimate = .432

5 year death probability estimate = .568

Assume censoring occurs at the left of each interval

time	$n_r$	d	w	$m^L = d/(n_r - w)$	$1 - m^L$	$\hat{S} = \prod(1 - m^L)$
0-1	146	27	3	.189	.811	.811
1-2	116	18	10	.170	.830	.673
2-3	88	21	10	.269	.731	.492
3-4	57	9	3	.167	.833	.410
4-5	45	1	3	.024	.976	.400

5 year survival probability estimate = .400

5 year death probability estimate = .600

Assume censoring occurs somewhere of each interval

time	$n_r$	d	w	$m = d/(n_r - w/2)$	$1 - m$	$\hat{S} = \prod(1 - m)$
0-1	146	27	3	.187	.813	.813
1-2	116	18	10	.162	.838	.681
2-3	88	21	10	.253	.747	.509
3-4	57	9	3	.162	.838	.426
4-5	45	1	3	.023	.977	.416

5 year survival probability estimate = .416

5 year death probability estimate = .584

- Standard Error estimate of Life-table estimate (Greenwood's formula):

$$\text{se}\{\hat{S}(t)\} = \hat{S}(t) \left\{ \sum_{j=1}^t \frac{d_j}{(n_{rj} - w_j/2)(n_{rj} - d_j - w_j/2)} \right\}^{1/2}.$$

- $(1 - \alpha)^{\text{th}}$  confidence interval for  $S(t)$  can be approximated by

$$\hat{S}(t) \pm \mathcal{Z}_{\alpha/2}[\text{se}\{\hat{S}(t)\}],$$

where  $\mathcal{Z}_{\alpha/2}$  is the  $(1 - \alpha/2)$  quantile of the standard normal distribution.

In our example

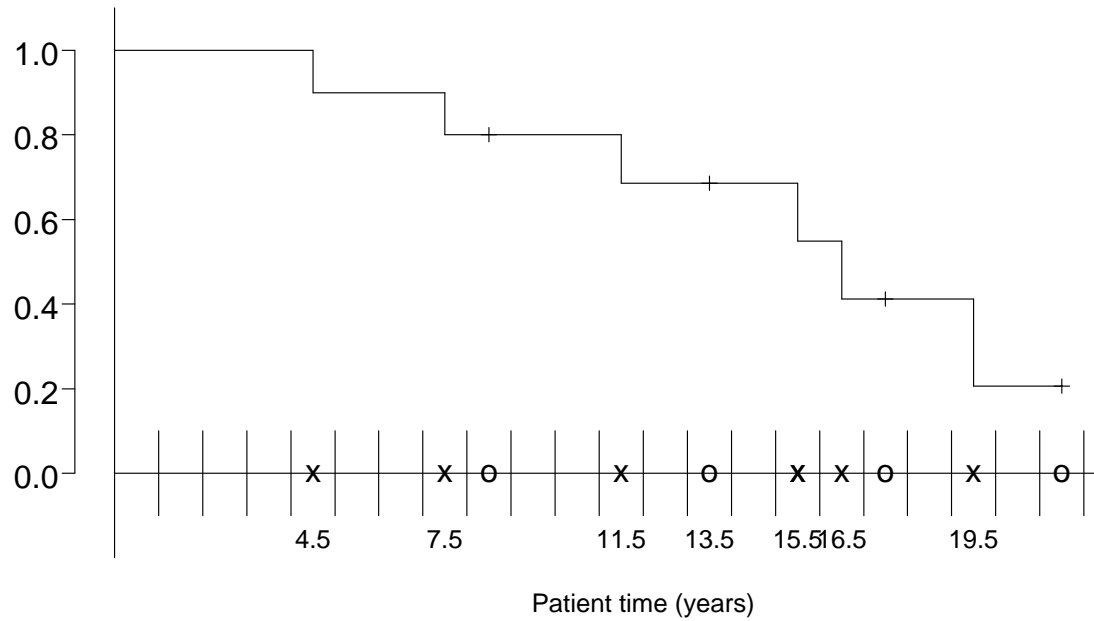
time	$n_r$	d	w	$\hat{S}$	$\frac{d}{(n_r - w/2)(n_r - d - w/2)}$	$\sum \frac{d}{(n - w/2)(n - d - w/2)}$
0-1	146	27	3	.813	.00159	.00159
1-2	118	18	10	.681	.00168	.00327
2-3	88	21	10	.509	.00408	.00735
3-4	57	9	3	.426	.00345	.01084
4-5	45	1	3	.417	.00054	.01138

The 95% confidence interval for  $S(5)$  is given as  $.417 \pm 1.96(.044)$   
 $= (.331 - .503)$ .

## 9.3 Kaplan-Meier or Product-Limit Estimator

- When we have precise information about each patient's survival time or censoring time, we need to use all the information to estimate  $S(t)$ .
- For any partition of  $[0, \infty)$ , life-table method can be used to estimate  $S(t)$ .
- The limit of the above estimate is **Kaplan-Meier** or **Product-limit** estimate of  $S(t)$ .

Figure 4: *An illustrative example of Kaplan-Meier estimator*



$1 - \hat{m}(x) :$	1	$\frac{9}{10}$	1	1	$\frac{8}{9}$	1	1	1	$\frac{6}{7}$	1	1	1	$\frac{4}{5}$	$\frac{3}{4}$	1	1	$\frac{1}{2}$	1	1
$\hat{S}(t) :$	1	$\frac{9}{10}$	.	.	$\frac{8}{10}$	.	.	.	$\frac{48}{70}$	.	.	.	$\frac{192}{350}$	$\frac{144}{350}$	.	.	$\frac{144}{700}$	.	.

where

$$\begin{aligned}
 m = d/n_r &= \frac{\text{number of deaths in an interval}}{\text{number at risk at beginning of interval}} \\
 &= \begin{cases} \frac{1}{n_r} & \text{a death occurred} \\ 0 & \text{no death occurred} \end{cases} \\
 (1 - m) &= \begin{cases} 1 - \frac{1}{n_r} & \text{a death occurred} \\ 1 & \text{no death occurred} \end{cases}
 \end{aligned}$$

- In the limit, the Kaplan-Meier (product-limit) estimator will be a step function taking jumps at times where a failure occurs.

$$\prod_{\text{all deaths}} \left( 1 - \frac{1}{\text{number at risk}} \right)$$

- By convention, the Kaplan-Meier estimator is taken to be right-continuous.

## Non-informative Censoring

In order that life-table estimators give unbiased results, there is an implicit assumption that individuals who are censored have the same risk of subsequent failure as those who are alive and uncensored. The risk set at any point in time (individuals still alive and uncensored) should be representative of the entire population alive at the same time in order that the estimated mortality rates reflect the true population mortality rates.

## Proc lifetest in SAS

```
Data example;
  input survtime censcode;
  cards;
4.5 1
7.5 1
8.5 0
11.5 1
13.5 0
15.5 1
16.5 1
17.5 0
19.5 1
21.5 0
;

Proc lifetest;
  time survtime*censcode(0);
run;
```

Part of the output:

The LIFETEST Procedure

Product-Limit Survival Estimates

SURVTIME	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	10
4.5000	0.9000	0.1000	0.0949	1	9
7.5000	0.8000	0.2000	0.1265	2	8
8.5000*	.	.	.	2	7
11.5000	0.6857	0.3143	0.1515	3	6
13.5000*	.	.	.	3	5
15.5000	0.5486	0.4514	0.1724	4	4
16.5000	0.4114	0.5886	0.1756	5	3
17.5000*	.	.	.	5	2
19.5000	0.2057	0.7943	0.1699	6	1
21.5000*	.	.	.	6	0

\* Censored Observation

## 9.4 Two-sample Log-rank Tests

- Want to compare two treatments where the primary end points are survival time (or time to an event):

$$H_0 : \lambda_1(t) = \lambda_0(t) \quad v.s. \quad H_A : \lambda_1(t) \leq \lambda_0(t) \quad \text{or} \quad \lambda_1(t) \geq \lambda_0(t)$$

- Data:  $(U_i, \Delta_i, A_i), i = 1, \dots, n$ 
  - ★  $U_i = \min(T_i, C_i)$ 
    - \*  $T_i$  denotes the latent failure time
    - \*  $C_i$  denotes the latent censoring time
  - ★  $\Delta_i = I(T_i \leq C_i)$  denotes failure indicator
  - ★  $A_i$  denotes treatment indicator

- Define

- ★  $n_j = \sum_{i=1}^n I(A_i = j)$  denotes the number of patients assigned treatment  $j = 0, 1$ ;  $n = n_0 + n_1$
- ★  $n_j(u) = \sum_{i=1}^n I(U_i \geq u, A_i = j)$  denotes the number at risk at time  $u$  from treatment  $j = 0, 1$
- ★  $n(u) = n_0(u) + n_1(u)$  denotes the total number at risk at time  $u$  from both treatments
- ★  $d_j(u) = \sum_{i=1}^n I(U_i = u, \Delta_i = 1, A_i = j)$  denotes the number of observed deaths at time  $u$  from treatment  $j = 0, 1$
- ★  $d(u) = d_0(u) + d_1(u)$  denotes the number of observed deaths at time  $u$  from both samples

## The log-rank test

- The numerator of the log-rank test statistic:

$$\sum_{\text{all death times } u} \left\{ d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right\}.$$

- At any time  $u$  where  $d(u) \geq 1$ , the data can be treated as

	treatment		total
	1	0	
number of deaths	$d_1(u)$	$d_0(u)$	$d(u)$
number alive	$n_1(u) - d_1(u)$	$n_0(u) - d_0(u)$	$n(u) - d(u)$
number at risk	$n_1(u)$	$n_0(u)$	$n(u)$

- ★ The observed number of deaths at time  $u$  from treatment 1 is  $d_1(u)$
- ★ The expected number of deaths from treatment 1 at time  $u$  if the null hypothesis were true is  $\frac{d(u)}{n(u)}n_1(u)$
- ★ Thus the observed minus expected number of deaths at time  $u$  is  $\left\{ d_1(u) - \frac{d(u)}{n(u)}n_1(u) \right\}$

- The numerator of the log-rank test statistic:

$$\text{sum of } \left\{ d_1(u) - \frac{d(u)}{n(u)}n_1(u) \right\} \text{ over } K \text{ } 2 \times 2 \text{ tables}$$

- Under  $H_0 : \lambda_1(t) = \lambda_0(t)$ ,

$$d_1(u) - \frac{d(u)}{n(u)}n_1(u) \text{ has mean zero}$$

- Under  $H_A : \lambda_1(t) < \lambda_0(t)$ ,

$$d_1(u) - \frac{d(u)}{n(u)}n_1(u) \text{ will be } < 0 \text{ on average}$$

- Standardize the above statistic leads to the log-rank test statistic:

$$T_n = \frac{\sum \left\{ d_1(u) - \frac{d(u)}{n(u)} n_1(u) \right\}}{\left[ \sum \frac{n_1(u)n_0(u)d(u)\{n(u)-d(u)\}}{n^2(u)\{n(u)-1\}} \right]^{1/2}}. \quad (9.1)$$

- The quantity

$$\frac{n_1(u)n_0(u)d(u)\{n(u)-d(u)\}}{n^2(u)\{n(u)-1\}}$$

is the variance of  $d_1(u)$  conditional on all margins.

$d_1(u)$	$\cdot$	$d(u)$
$\cdot$	$\cdot$	$n(u) - d(u)$
$n_1(u)$	$n_0(u)$	$n(u)$

The value  $d_1(u)$ , under the null hypothesis, conditional on the marginal totals, has a hypergeometric distribution with mean

$$\frac{d(u)}{n(u)} n_1(u)$$

and variance

$$\frac{n_1(u)n_0(u)d(u)\{n(u) - d(u)\}}{n^2(u)\{n(u) - 1\}}.$$

- Therefore, under  $H_0 : \lambda_1(t) = \lambda_0(t)$ ,

$$T_n \stackrel{H_0}{\sim} N(0, 1).$$

Consequently,

- ★ A two-sided level  $\alpha$  test would reject the null hypothesis when  $|T_n| \geq \mathcal{Z}_{\alpha/2}$ .
- ★ If testing  $H_0 : \lambda_1(t) = \lambda_0(t)$  v.s.  $H_A : \lambda_1(t) < \lambda_0(t)$ , a level  $\alpha$  test would reject the null hypothesis when  $T_n \leq -\mathcal{Z}_\alpha$ .
- ★ If testing  $H_0 : \lambda_1(t) = \lambda_0(t)$  v.s.  $H_A : \lambda_1(t) > \lambda_0(t)$ , a level  $\alpha$  test would reject the null hypothesis when  $T_n \geq \mathcal{Z}_\alpha$ .
- **Note:** The distribution of  $T_n$  under  $H_0$  does not depend on the shape of  $\lambda_1(t)$  and  $\lambda_0(t)$ , so  $T_n$  is a nonparametric test.

- **Example:** in CALGB 8541 (clinical trial on breast cancer), the major focus was to compare treatment 1 (Intensive CAF) to treatment 2 (Low dose CAF), where CAF is the combination of the drugs Cyclophosphamide, Adriamycin and 5 Fluorouracil.

```
data trt12; set bcancer;  
  if (trt=1) or (trt=2);  
run;
```

```
title "Log-rank test comparing treatments 1 and 2";  
proc lifetest data=trt12 notable;  
  time years*censor(0);  
  strata trt;  
run;
```

Part of the output from the above SAS program:

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for years over Strata

Rank Statistics

trt	Log-Rank	Wilcoxon
1	-30.030	-23695
2	30.030	23695

Covariance Matrix for the Log-Rank Statistics

trt	1	2
1	91.3725	-91.3725
2	-91.3725	91.3725

Covariance Matrix for the Wilcoxon Statistics

trt	1	2
1	54635903	-5.464E7
2	-5.464E7	54635903

## Test of Equality over Strata

Test	Chi-Square	DF	Pr >
			Chi-Square
Log-Rank	9.8692	1	0.0017
Wilcoxon	10.2763	1	0.0013
-2Log(LR)	9.5079	1	0.0020

## 9.5 Power and Sample Size Based on the Log-rank Test

- Need to know the distribution of  $T_n$  under  $H_A : \lambda_1(t) \neq \lambda_0(t)$ .
- Infinitely many parameters under  $H_A$ !
- Consider proportional hazards (PH) alternative

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\gamma), \text{ for all } t \geq 0. \quad (9.2)$$

The hazard ratio  $\exp(\gamma)$  can be viewed as a relative risk and used for purposes of testing the null hypothesis of no treatment difference

- ★  $\gamma > 0$  implies that individuals on treatment 1 have worse survival (i.e. die faster)
- ★  $\gamma = 0$  implies the null hypothesis
- ★  $\gamma < 0$  implies that individuals on treatment 1 have better survival (i.e. live longer)

- Under PH:  $\lambda_1(t) = \lambda_0(t) \exp(\gamma)$ , we have

$$-\frac{d \log S_1(t)}{dt} = -\frac{d \log S_0(t)}{dt} \exp(\gamma),$$

or

$$-\log S_1(t) = -\log S_0(t) \exp(\gamma).$$

Consequently,

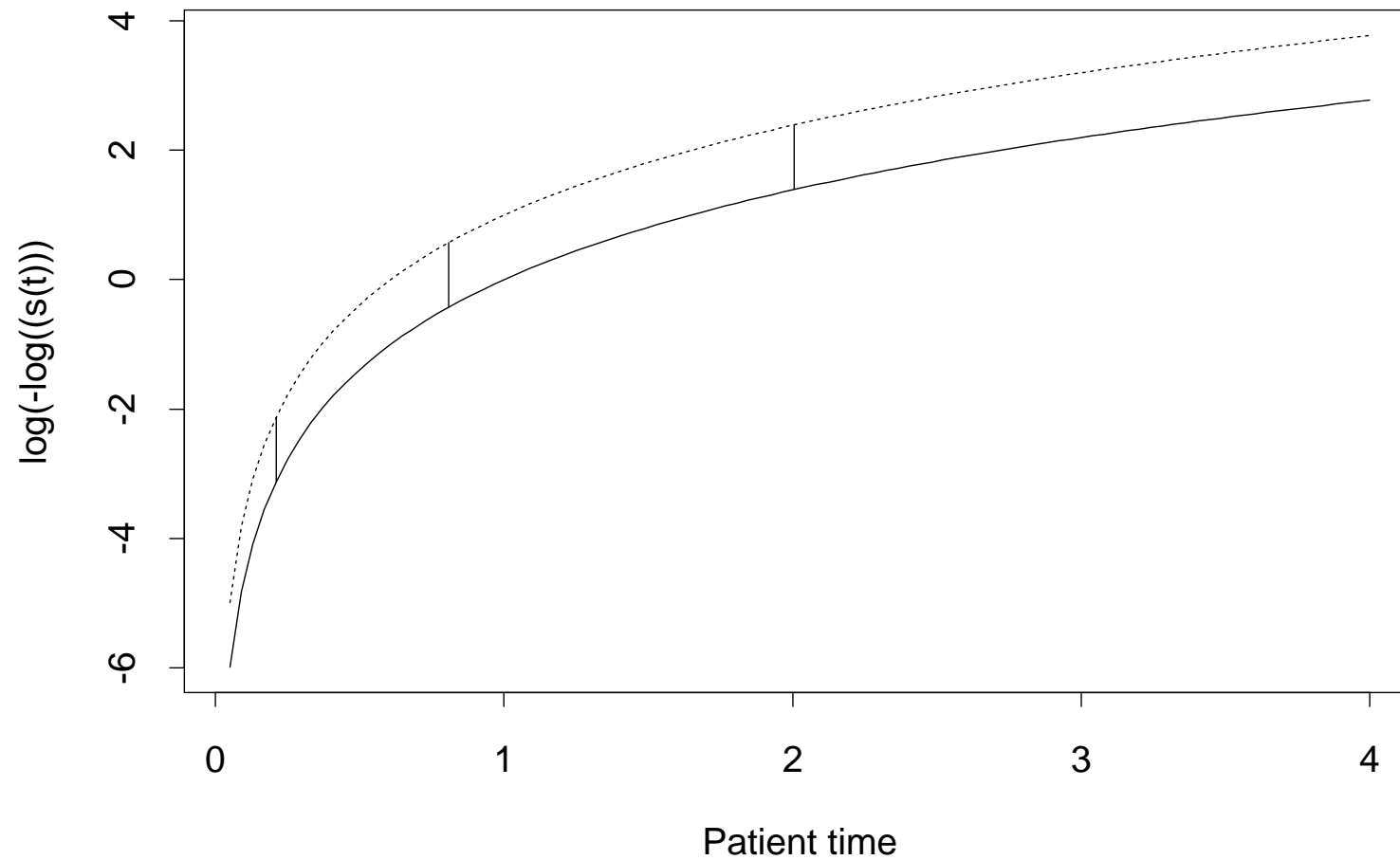
$$S_1(t) = \{S_0(t)\}^{\exp(\gamma)},$$

and

$$\log\{-\log S_1(t)\} = \log\{-\log S_0(t)\} + \gamma.$$

If we estimate  $S_1(t)$  and  $S_0(t)$  by their KM estimates  $\hat{S}_1(t)$  and  $\hat{S}_0(t)$ , then we can plot  $\log\{-\log \hat{S}_1(t)\}$  v.s.  $\log\{-\log \hat{S}_0(t)\}$  to see if they are roughly parallel.

Figure 5: *Two survival functions with proportional hazards on log[-log] scale*



- If we assume exponential distributions for each treatments, then  $\lambda_1(t) = \lambda_1$  and  $\lambda_0(t) = \lambda_0$ ; automatic PH.
  - ★ Medians:  $m_1 = \log(2)/\lambda_1$ ,  $m_0 = \log(2)/\lambda_0$ , so

$$\frac{\lambda_1}{\lambda_0} = \frac{m_0}{m_1}$$

- ★ Means:  $\mu_1 = 1/\lambda_1$ ,  $\mu_0 = 1/\lambda_0$ , so

$$\frac{\lambda_1}{\lambda_0} = \frac{\mu_0}{\mu_1}$$

- Under  $H_A$ :

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\gamma), \text{ for all } t \geq 0,$$

$T_n$  is asymptotically distributed as

$$T_n \stackrel{H_A}{\approx} N(\{d\theta(1 - \theta)\}^{1/2}\gamma_A, 1),$$

where  $d$  denotes the total number of deaths (events), and  $\theta$  denotes the proportion randomized to treatment 1 (generally .5). That is, under a proportional hazards alternative, the logrank test is distributed approximately as a normal random variable with variance 1 and noncentrality parameter

$$\{d\theta(1 - \theta)\}^{1/2}\gamma_A.$$

When  $\theta = .5$ , the noncentrality parameter is

$$\gamma_A d^{1/2} / 2.$$

- In order that a level  $\alpha$  test (say, two-sided) have power  $1 - \beta$  to detect the alternative

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\gamma_A),$$

then the noncentrality parameter must equal  $\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}$ . That is,

$$\gamma_A d^{1/2} / 2 = \mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta},$$

or

$$d = \frac{4(\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta})^2}{\gamma_A^2}. \quad (9.3)$$

- If we take  $\alpha = .05$  (two-sided), power  $(1 - \beta) = .90$ , and  $\theta = .5$ , then

$$d = \frac{4(1.96 + 1.28)^2}{\gamma_A^2}. \quad (9.4)$$

- The required number of deaths for some hazard ratios:

<u>Hazard Ratio <math>\exp(\gamma_A)</math></u>	<u>Number of deaths <math>d</math></u>
2.00	88
1.50	256
1.25	844
1.10	4623

## Sample Size Considerations

- Strategy 1: Continue a clinical trial until we obtain the required number of failures. For example,
  - ★ Median survival time of patients with advanced lung cancer: 6 months
  - ★ Median survival time has to be 9 months for a new treatment to be clinically important.
  - ★ Use two-sided log-rank test at 0.05 significance level, want to detect 9 month median survival time with power 90%.
  - ★ Assume exponential distributions.

the clinically important hazard ratio:

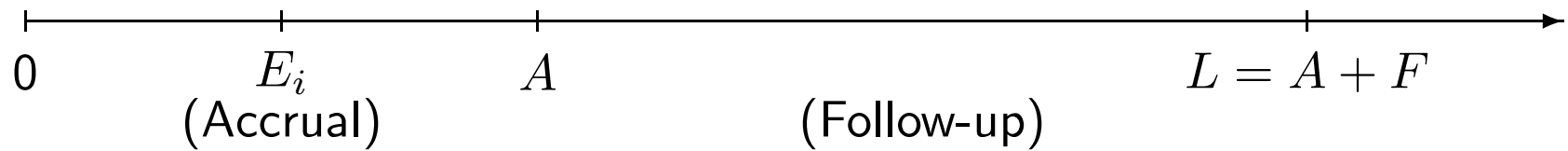
$$\frac{\lambda_1}{\lambda_0} = \frac{m_0}{m_1} = \frac{6}{9} = 2/3.$$

Hence  $\gamma_A = \log(2/3) = -.4055$ . Formula (9.4) gives  $d = 256$  deaths.

Since, for this example, patients do not survive for very long, we can continue the study until 256 deaths occur, or enter more patients (say 350) and analyze the data after 256 deaths.

- Need to better plan the study so to allocate resource.
  - ★ number of patients
  - ★ accrual period
  - ★ follow-up time

Figure 6: *Illustration of accrual and follow-up*



- Assume end-of-study censoring (“administrative censoring”) and define
  - ★  $A$  = accrual period: calendar time period for patients to enter.
  - ★  $F$  = additional follow-up period
  - ★  $L = A + F$  total study length
  - ★ Accrual rate at calendar time  $u$ :  $a(u)$ ; more precisely,

$$a(u) = \lim_{h \rightarrow 0} \left\{ \frac{\text{Expected \# of patients entering in } [u, u + h)}{h} \right\}.$$

Usually, we assume  $a(u) = a$ , say 100 patients/month

Totally # of patients in study:

$$\int_0^A a(u) du \{= Aa \text{ if } a(u) = a\}.$$

- If we do equal allocation, then total expected # of deaths for treatment  $j = 0, 1$ :

$$d_j = \int_0^A \frac{a(u)}{2} F_j(L - u) du, j = 0, 1, \quad (9.5)$$

where  $F_j(t)$  is the cumulative distribution function of the survival time for treatment  $j$ .

- Then for the design characteristics, we need

$$d_1 + d_0 = \frac{4(\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta})^2}{\gamma_A^2}.$$

We can solve the above equation to get  $A$ ,  $F$  and  $L$ .

- If we assume  $a(u) = a$  and exponential distributions with  $\lambda_1$  and  $\lambda_0$ , then

$$\begin{aligned} d_j &= \int_0^A \frac{a}{2} [1 - \exp\{-\lambda_j(L - u)\}] du \\ &= \frac{a}{2} \left[ A - \frac{\exp(-\lambda_j L)}{\lambda_j} \{\exp(\lambda_j A) - 1\} \right]. \end{aligned}$$

- **Example:**

- ★ Median survival time for treatment 0 (standard): 4 years
- ★ Median survival time for treatment 1 (new) is expected = 6 years
- ★ Want to have 90% power to detect this using 2-sided log-rank test at  $\alpha = 0.05$ .
- ★ Assume  $a = 100/\text{year}$
- ★ Assume exponential distributions:  $\lambda_0 = \log(2)/4 = .173$ ,  $\lambda_1 = \log(2)/9 = .116$ .

- Hazard ratio =  $2/3$ ,  $\gamma_A = \log(2/3)$ . Using equation (9.3), the total number of deaths necessary is 256. Hence, we need

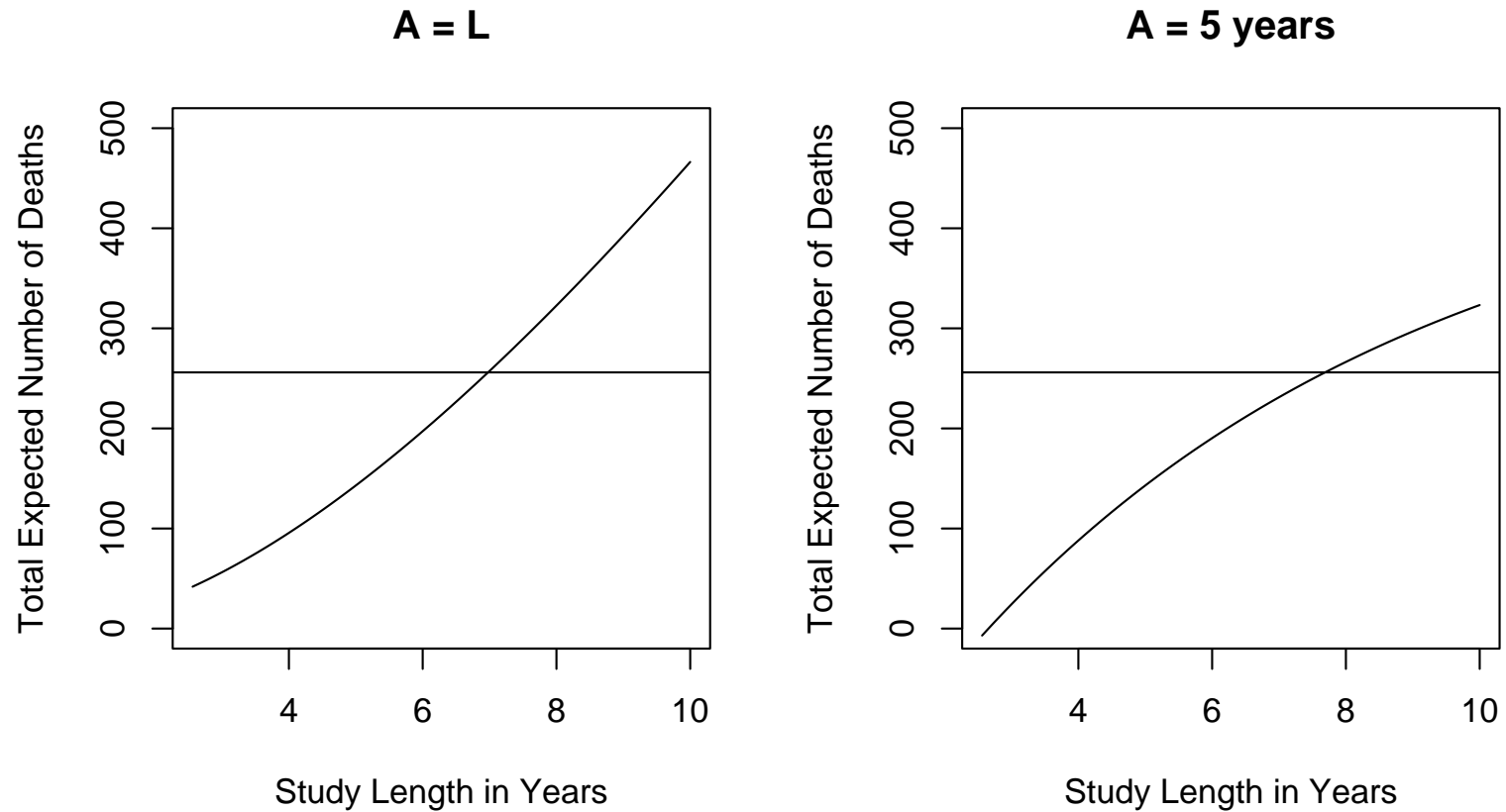
$$d_1(A, L) + d_0(A, L) = 256.$$

- Using above formula:

$$\begin{aligned} & 50 \left[ A - \frac{\exp(-.116L)}{.116} \{ \exp(.116A) - 1 \} \right] \\ & + 50 \left[ A - \frac{\exp(-.173L)}{.173} \{ \exp(.173A) - 1 \} \right] \\ & = 256. \end{aligned}$$

- Infinitely many solutions  $(A, L)$  from the above equation.
  - ★ If  $A = L$  (no additional follow-up), then  $A = L = 7$ . Total # of patients = 700.
  - ★  $A = 5$ ,  $L = ?$  Solving the equation  $\implies L = 7.65$ .
  - ★  $A$  has to be larger than 2.56 years.

Figure 7: *Total expected number of deaths as a function of the study length*



- Other factors that may affect power that we will not discuss are
  - ★ loss to follow-up (withdrawal)
  - ★ competing risks
  - ★ non-compliance

An excellent account on how to deal with these issues during the design stage is given by Lakatos (1988), *Biometrics*.

## 9.6 K-Sample Log-rank Tests

- Compare  $K$  treatments. Null hypothesis is

$$H_0 : S_1(t) = \dots = S_K(t), \quad t \geq 0,$$

or equivalently

$$H_0 : \lambda_1(t) = \dots = \lambda_K(t), \quad t \geq 0,$$

where  $S_j(t)$  and  $\lambda_j(t)$ ,  $j = 1, \dots, K$  denote the treatment-specific survival functions and hazard rates.

- Extend the two-sample log-rank test to  $K$ -sample situation.

- Denote
  - ★  $n_j(u) = \#$  of individuals at risk at  $u$  in treatment  $j$
  - ★  $d_j(u) = \#$  of observed deaths at  $u$  in treatment group  $j$ .
- We then consider

$$\begin{pmatrix} d_1(u) - \frac{n_1(u)}{n(u)} d(u) \\ \cdot \\ \cdot \\ \cdot \\ d_K(u) - \frac{n_K(u)}{n(u)} d(u) \end{pmatrix}^{K \times 1}$$

The 1st element is the numerator of the log-rank test statistic for comparing treatment 1 to other treatments combined, etc.

- One element is redundant. We can take any  $K - 1$  of them. Denote the vector by  $\mathcal{T}_n$ .

- Under  $H_0 : \lambda_1(t) = \dots = \lambda_K(t)$ ,  $t \geq 0$ ,

$$\mathcal{T}_n \stackrel{a}{\sim} \mathbf{N}(0, \mathcal{V}),$$

where

$$\mathcal{V}_{njj} = \sum_{\text{death times } u} \left[ \frac{d(u)\{n(u) - d(u)\}n_j(u)\{n(u) - n_j(u)\}}{n^2(u)\{n(u) - 1\}} \right],$$

and for  $j \neq j'$ , the off-diagonal terms are

$$\mathcal{V}_{njj'} = - \sum_{\text{death times } u} \left[ \frac{d(u)\{n(u) - d(u)\}n_j(u)n_{j'}(u)}{n^2(u)\{n(u) - 1\}} \right].$$

- Therefore, under  $H_0 : \lambda_1(t) = \dots = \lambda_K(t)$ ,  $t \geq 0$ ,

$$T_n = \mathcal{T}_n^T [\mathcal{V}_n]^{-1} \mathcal{T}_n \stackrel{a}{\sim} \chi_{K-1}^2,$$

and a level  $\alpha$  test rejects  $H_0$  if

$$T_n \geq \chi_{\alpha; K-1}^2.$$

## 9.7 Sample-size Considerations for the K-sample Log-rank Test

- Design characteristics:
  - ★ Overall type I error probability  $\alpha$
  - ★ Testing procedure: reject  $H_0$  if  $T_n \geq \chi_{\alpha; K-1}^2$
  - ★ Power  $1 - \beta$  to detect a hazard ratio (assumed constant over time) between any two treatments  $\geq \exp(\gamma_A)$ .
- The required total # of deaths has to satisfy

$$d = \frac{2K\phi^2(\alpha, \beta, K - 1)}{\gamma_A^2}$$

- **Example:**  $K = 3$ ,  $\alpha = 0.05$ , power  $1 - \beta = 0.9$  to detect a hazard ratio between any two treatments that may exceed 1.5. Then

$$d = \frac{2 \times 3 \times 12.654}{\{\log(1.5)\}^2} = 462.$$

- We can use the same strategies we used for comparing two treatments:
  1. Enter a large number of patients, say, 800 and do data analysis when we observe 462 deaths.
  2. Under some assumptions on accrual rate  $a(u)$ , underlying hazard functions and censoring patterns, calculate  $d_j$  for each treatment  $j = 1, 2, 3$ . Then solve

$$d_1(A, L) + d_2(A, L) + d_3(A, L) = 462$$

to get  $A$  and  $L$ .

## 9.8 Analyzing Data Using $K$ -sample Log-rank Test

- **CALGB 8541 Example**

- ★ treatment 1 (Intensive CAF)
- ★ treatment 2 (Low dose CAF)
- ★ treatment 3 (Standard dose CAF)

where CAF denotes the combination of Cyclophosphamide, Adriamycin and 5-Fluorouracil.

- **SAS program:**

```
title "Log-rank test comparing all three treatments";  
proc lifetest data=bcancer notable;  
  time years*censor(0);  
  strata trt;  
run;
```

Part of the output:

Log-rank test comparing all three treatments

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for years over Strata

Rank Statistics

trt	Log-Rank	Wilcoxon
1	-21.245	-27171
2	37.653	43166
3	-16.408	-15995

Covariance Matrix for the Log-Rank Statistics

trt	1	2	3
1	120.132	-57.761	-62.371
2	-57.761	114.004	-56.243
3	-62.371	-56.243	118.615

## Covariance Matrix for the Wilcoxon Statistics

trt	1	2	3
1	1.6295E8	-7.94E7	-8.355E7
2	-7.94E7	1.5675E8	-7.734E7
3	-8.355E7	-7.734E7	1.6089E8

## Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	12.4876	2	0.0019
Wilcoxon	12.1167	2	0.0023
-2Log(LR)	11.3987	2	0.0033

# 10 Early Stopping of Clinical Trials

## 10.1 General Issues in Monitoring Clinical Trials

Clinical trials are monitored periodically and may be stopped early due to

- Serious toxicity or adverse events
- Established benefit
- No trend of interest
- Design or logistical difficulties too serious to fix

Clinical trials will be monitored by an independent Data Safety Monitoring Board (DSMB).

- Primary responsibility of a DSMB: Ensure the safety and well being of the patients in the trial.
- Members of a DSMB:
  - ★ Clinical
  - ★ Laboratory
  - ★ Epidemiology
  - ★ Biostatistics
  - ★ Data Management
  - ★ Ethics
- DSMB should have no conflict of interest with the study or studies they are monitoring; All the discussions of the DSMB are confidential.

DSMB's duty:

- Protocol review
- Interim reviews
  - ★ study progress
  - ★ quality of data
  - ★ safety
  - ★ efficacy and benefit
- Manuscript review

During the early stages of a clinical trial the focus is on administrative issues regarding the conduct of the study. These include:

- Recruitment/Entry Criteria
- Baseline comparisons
- Design assumptions and modifications
  - ★ entry criteria

- ★ treatment dose
- ★ sample size adjustments
- ★ frequency of measurements
- Quality and timeliness of data collection

- Focus on possible early termination due to established treatment difference.
- **Group-sequential** methods: monitor data sequentially at some finite fixed time points
- Information-based design and monitoring of clinical trials

The typical scenario where these methods can be applied is as follows:

- Data in a study are collected over calendar time.
- Want to see if one treatment (new) is better than the other (old)
- conduct interim analysis to assess whether there is sufficient “strong evidence” to warrant early termination of the study
- At each monitoring time, a test statistic is computed and compared to a stopping boundary.
- The stopping boundaries are chosen to preserve certain operating characteristics that are desired; i.e. level and power

The methods we present are general enough to include problems where

- t-tests are used to compare the mean of continuous random variables between treatments
- proportion tests for dichotomous response variables
- logrank test for censored survival data
- tests based on likelihood methods for either discrete or continuous random variables; i.e. Score test, Likelihood ratio test, Wald test using maximum likelihood estimators

## 10.2 Information Based Design and Monitoring

Underlying structure:

- Data are generated from

$$f(y, \Delta, \theta),$$

where

★  $\Delta =$  treatment effect

★  $\theta =$  nuisance parameters

- We want to test the null hypothesis

$$H_0 : \Delta = 0$$

versus the alternative

$$H_A : \Delta \neq 0.$$

**Note** The methods discussed here can be modified to test

$$H_0 : \Delta \leq 0$$

versus

$$H_A : \Delta > 0.$$

- At any interim analysis time  $t$ , our decision making will be based on the test statistic

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}},$$

where  $\hat{\Delta}(t)$  is an estimator for  $\Delta$  and  $se\{\hat{\Delta}(t)\}$  is the estimated standard error of  $\hat{\Delta}(t)$  using all the data that have accumulated up to time  $t$ . For two-sided tests we would reject the null hypothesis if the absolute value of the test statistic  $|T(t)|$  were sufficiently large and for one-sided tests if  $T(t)$  were sufficiently large.

- **Some examples:**

1. **Example 1.** (Dichotomous response)

Let  $\pi_1, \pi_0$  denote the population response rates for treatments 1 (new treatment) and 0 (control) respectively. Let the treatment difference be given by

$$\Delta = \pi_1 - \pi_0$$

The test of the null hypothesis at time  $t$  will be based on

$$T(t) = \frac{p_1(t) - p_0(t)}{\sqrt{\bar{p}(t)\{1 - \bar{p}(t)\} \left\{ \frac{1}{n_1(t)} + \frac{1}{n_2(t)} \right\}}},$$

where using all the data available through time  $t$ ,  $p_j(t)$  denotes the sample proportion responding among the  $n_j(t)$  individuals on treatment  $j = 0, 1$ .

## 2. Example 2. (Time to event)

Suppose we assume a proportional hazards model:

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(-\Delta),$$

and we want to test the null hypothesis of no treatment difference

$$H_0 : \Delta = 0 \quad vs. \quad H_A : \Delta \neq 0$$

$$(\text{or } H_0 : \Delta \leq 0 \quad vs. \quad H_A : \Delta > 0.)$$

At time  $t$ , we would compute the test statistic

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}},$$

where  $\hat{\Delta}(t)$  is the maximum partial likelihood estimator of  $\Delta$ . For the two-sided test we would reject the null hypothesis if  $|T(t)|$  were sufficiently large and for the one-sided test if  $T(t)$  were sufficiently large.

**Remark:** The material on the use and the properties of the maximum partial likelihood estimator are taught in the classes on Survival Analysis. We note, however, that the logrank test computed using all the data up to time  $t$  is equivalent asymptotically to the test based on  $T(t)$ .

3. **Example 3.** (Any parametric models  $p(z; \Delta, \theta)$ ):  
Find MLE  $\hat{\Delta}(t)$  of  $\Delta$  and its  $se\{\hat{\Delta}(t)\}$  and conduct testing using

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}}$$

- Usually,

- ★ Under  $H_0 : \Delta = 0$ ,

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}} \stackrel{a}{\sim} N(0, 1).$$

- ★ Under  $H_A : \Delta = \Delta^* \neq 0$ ,

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}} \stackrel{\Delta=\Delta^*}{\sim} N(\Delta^* I^{1/2}(t, \Delta^*), 1),$$

where  $I(t, \Delta^*)$  is the statistical information (for  $\Delta$ ) at  $t$ , and

$$I(t, \Delta^*) \approx \{se(\hat{\Delta}(t))\}^{-2}.$$

## Group sequential test

- Determine the total number of interim analysis  $K$ .
- Determine boundary values  $b(t_j)$  for  $j = 1, 2, \dots, K$ .
- Reject  $H_0 : \Delta = 0$  at the first time  $t_j$  when

$$|T(t_j)| \geq b(t_j),$$

if we consider two-sided test.

- Question: What values  $b(t_j)$  should be used?
- Suppose we want to control the overall type I error prob at  $\alpha$ , can we use  $b(t_j) = Z_{\alpha/2}$ ?

- The actual type I error if  $b(t_j) = 1.96$  is used:

$K$	false positive rate
1	0.050
2	0.083
3	0.107
4	0.126
5	0.142
10	0.193
20	0.246
50	0.320
100	0.274
1,000	0.530
$\infty$	1.000

- $\implies$  For given overall type I error prob  $\alpha$ , need to find out  $b(t_j)$  for  $j = 1, 2, \dots, K$  such that

$$P[\text{reject } H_0 | H_0] = \alpha,$$

or equivalently

$$P[\text{accept } H_0 | H_0] = 1 - \alpha$$

- Sequential test implies that we accept  $H_0$  if

$$|T(t_j)| \leq b(t_j), \text{ for all } j = 1, 2, \dots, K.$$

$\implies$

$$P_{\Delta=0}\{|T(t_1)| < b(t_1), \dots, |T(t_K)| < b(t_K)\} = 1 - \alpha. \quad (10.1)$$

- Need to know the joint distribution of  $T(t_1), T(t_2), \dots, T(t_K)$ .

- Fundamental result:

“Any efficient based test or estimator for  $\Delta$ , properly normalized, when computed sequentially over time, has, asymptotically, a normal independent increments process whose distribution depends only on the parameter  $\Delta$  and the statistical information.”

Scharfstein, Tsiatis and Robins (1997). JASA. 1342-1350.

- Define

$$W(t) = I^{1/2}(t, \Delta^*)T(t)$$

- Since when  $\Delta = \Delta^*$ ,

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}} \sim N(\Delta^* I^{1/2}(t, \Delta^*), 1),$$

$\implies$

$$w(t) \sim N(\Delta^* I(t, \Delta^*), I(t, \Delta^*)).$$

- The joint distribution of the multivariate vector  $\{W(t_1), \dots, W(t_K)\}$  is asymptotically normal with mean vector  $\{\Delta^* I(t_1, \Delta^*), \dots, \Delta^* I(t_K, \Delta^*)\}$  and covariance matrix where

$$\text{var}\{W(t_j)\} = I(t_j, \Delta^*), \quad j = 1, \dots, K$$

and

$$\text{cov}[W(t_j), \{W(t_\ell) - W(t_j)\}] = 0, \quad j < \ell, j, \ell = 1, \dots, K.$$

- That is,
  - ★ The statistic  $W(t_j)$  has mean and variance proportional to the statistical information at time  $t_j$
  - ★ Has independent increments; that is

$$W(t_1) = W(t_1)$$

$$W(t_2) = W(t_1) + \{W(t_2) - W(t_1)\}$$

·

·

·

$$W(t_j) = W(t_1) + \{W(t_2) - W(t_1)\} + \dots + \{W(t_j) - W(t_{j-1})\}$$

has the same distribution as a partial sum of independent normal random variables

This structure implies that the covariance matrix of

$\{W(t_1), \dots, W(t_K)\}$  is given by

$$\text{var}\{W(t_j)\} = I(t_j, \Delta^*), \quad j = 1, \dots, K$$

and for  $j < \ell$

$$\begin{aligned} & \text{cov}\{W(t_j), W(t_\ell)\} \\ &= \text{cov}[W(t_j), \{W(t_\ell) - W(t_j)\} + W(t_j)] \\ &= \text{cov}[W(t_j), \{W(t_\ell) - W(t_j)\}] + \text{cov}\{W(t_j), W(t_j)\} \\ &= 0 + \text{var}\{W(t_j)\} \\ &= I(t_j, \Delta^*). \end{aligned}$$

- Since

$$T(t_j) = I^{-1/2}(t_j, \Delta^*)W(t_j), \quad j = 1, \dots, K$$

so  $\{T(t_1), \dots, T(t_K)\}$  is also multivariate normal with mean

$$E\{T(t_j)\} = \Delta^* I^{1/2}(t_j, \Delta^*), \quad j = 1, \dots, K \quad (10.2)$$

and the covariance matrix

$$\text{var}\{T(t_j)\} = 1, \quad j = 1, \dots, K \quad (10.3)$$

and for  $j < \ell$ , the covariances are

$$\begin{aligned} \text{cov}\{T(t_j), T(t_\ell)\} &= \text{cov}\{I^{-1/2}(t_j, \Delta^*)W(t_j), I^{-1/2}(t_\ell, \Delta^*)W(t_\ell)\} \\ &= I^{-1/2}(t_j, \Delta^*)I^{-1/2}(t_\ell, \Delta^*)\text{cov}\{W(t_j), W(t_\ell)\} \\ &= I^{-1/2}(t_j, \Delta^*)I^{-1/2}(t_\ell, \Delta^*)I(t_j, \Delta^*) \\ &= \frac{I^{1/2}(t_j, \Delta^*)}{I^{1/2}(t_\ell, \Delta^*)} = \sqrt{\frac{I(t_j, \Delta^*)}{I(t_\ell, \Delta^*)}}. \end{aligned} \quad (10.4)$$

- Under  $H_0 : \Delta = 0$ ,  $\{T(t_1), \dots, T(t_K)\}$  is multivariate normal with mean vector zero and covariance (correlation) matrix

$$V_T = \left[ \sqrt{\frac{I(t_j, 0)}{I(t_\ell, 0)}} \right], \quad j \leq \ell. \quad (10.5)$$

- When we conduct interim analysis with equal increment of information (usually equal # of patients)

$$I(t_1, \cdot) = I, \quad I(t_2, \cdot) = 2I, \quad \dots, \quad I(t_K, \cdot) = KI,$$

then

$$\begin{bmatrix} T(t_1) \\ T(t_2) \\ \vdots \\ T(t_K) \end{bmatrix} \stackrel{H_0}{\sim} N \left( \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sqrt{\frac{1}{2}} & \cdots & \sqrt{\frac{1}{K}} \\ \sqrt{\frac{1}{2}} & 1 & \cdots & \sqrt{\frac{2}{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{1}{K}} & \sqrt{\frac{2}{K}} & \cdots & 1 \end{bmatrix} \right)$$

- Given this structure, for given boundary values  $b_j = b(t_j), j = 1, 2, \dots, K$  we can iteratively calculate

$$P_{\Delta=0}\{|T(t_1)| < b_1, \dots, |T(t_K)| < b_K\}.$$

Please see Armitage, McPherson and Rowe (1969, JRSS-A) for more detail.

- For given type I error  $\alpha$  we want to control, too many  $b_j$ 's satisfy

$$P_{\Delta=0}\{|T(t_1)| < b_1, \dots, |T(t_K)| < b_K\} = 1 - \alpha.$$

- How do we choose  $b_j$ 's?

## 10.3 Choice of Boundaries

- Wang and Tsiatis (1987) *Biometrics* proposed a flexible class of boundaries

$$b_j = (\text{constant}) \times j^{(\Phi-.5)},$$

where  $\Phi$  determines the shape of a boundary (so  $\Phi$  is called the shape parameter).

- For given  $\alpha$ ,  $K$ , and  $\Phi$ , the constant  $c$  can be computed so that

$$P_{\Delta=0} \left\{ \bigcap_{j=1}^K |T(t_j)| < c j^{(\Phi-.5)} \right\} = 1 - \alpha.$$

- Denote this  $c$  by  $c(\alpha, K, \Phi)$ .

Table 1:  $c(\alpha, K, \Phi)$  for some selected values of  $\alpha$ ,  $K$ ,  $\Phi$ 

$\Phi$	$\alpha = 0.05$				$\alpha = 0.01$			
	$K$				$K$			
	2	3	4	5	2	3	4	5
0.0	2.7967	3.4712	4.0486	4.5618	3.6494	4.4957	5.2189	5.8672
0.1	2.6316	3.1444	3.5693	3.9374	3.4149	4.0506	4.5771	5.0308
0.2	2.4879	2.8639	3.1647	3.4175	3.2071	3.6633	4.0276	4.3372
0.3	2.3653	2.6300	2.8312	2.9945	3.0296	3.3355	3.5706	3.7634
0.4	2.2636	2.4400	2.5652	2.6628	2.8848	3.0718	3.2071	3.3137
0.5	2.1784	2.2896	2.3616	2.4135	2.7728	2.8738	2.9395	2.9869

- **Pocock boundaries:**

$$\Phi = 0.5, \implies b_j = c(\alpha, K, 0.5), \quad j = 1, 2, \dots, K$$

For example, if  $K = 5$  and  $\alpha = .05$ , then  $c(.05, 5, 0.5) = 2.41$ .

That is, we reject  $H_0 : \Delta = 0$  the first time where

$$|T(t_j)| \geq 2.41$$

Equivalently, we reject  $H_0 : \Delta = 0$  the first time where

$$P - \text{value} \leq 0.0158$$

- **O'Brien-Fleming boundaries:**

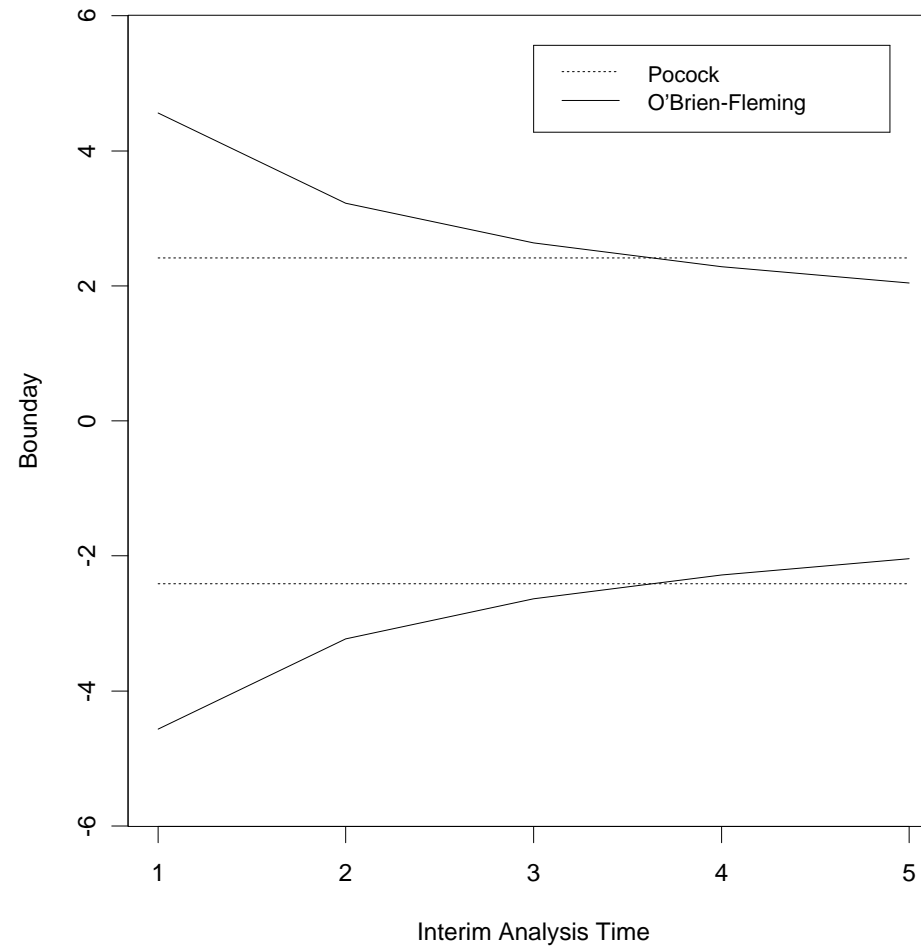
$$\Phi = 0, \implies b_j = c(\alpha, K, 0)j^{-1/2} = c(\alpha, K, 0)/\sqrt{j}, \quad j = 1, 2, \dots, K$$

For example, if  $K = 5$  and  $\alpha = .05$ , then  $c(.05, 5, 0.0) = 4.56$ . Then  $b_j = c(\alpha, K, 0)/\sqrt{j} = 4.56/\sqrt{j}$  gives 5 boundary values:

$$b_1 = 4.56, \quad b_2 = 3.22, \quad b_3 = 2.63, \quad b_4 = 2.28, \quad b_5 = 2.04$$

Table 2: *Nominal p-values for  $K = 5$  and  $\alpha = .05$* 

Nominal p-value		
j	Pocock	O'Brien-Fleming
1	0.0158	0.000005
2	0.0158	0.00125
3	0.0158	0.00843
4	0.0158	0.0225
5	0.0158	0.0413

Figure 1: *Pocock and O'Brien-Fleming Boundaries*

## 10.4 Power and Sample Size in Terms of Information

- The test statistic  $T(t)$ :

$$T(t) \stackrel{\Delta=0}{\sim} N(0, 1)$$

and under a clinically important alternative  $\Delta = \Delta_A$

$$T(t) \stackrel{\Delta=\Delta_A}{\sim} N(\Delta_A I^{1/2}(t, \Delta_A), 1),$$

where  $I(t, \Delta_A)$  denotes statistical information which can be approximated by  $[se\{\hat{\Delta}(t)\}]^{-2}$

- In order that a two-sided level- $\alpha$  test have power  $1 - \beta$  to detect  $\Delta_A$ , the noncentrality parameter  $\Delta_A I^{1/2}(t, \Delta_A)$  must satisfy

$$\Delta_A I^{1/2}(t, \Delta_A) = \mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta},$$

or

$$I(t, \Delta_A) = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2. \quad (10.6)$$

- Since  $I(t, \Delta_A) = [se\{\hat{\Delta}(t)\}]^{-2}$ , we need to collect enough data so that

$$[se\{\hat{\Delta}(t)\}]^{-2} = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2.$$

- One strategy would monitor  $se\{\hat{\Delta}(t)\}$  until it satisfies the above condition (at  $t^F$  and do data analysis). Reject  $H_0 : \Delta = 0$  if

$$|T(t^F)| \geq \mathcal{Z}_{\alpha/2}.$$

- **Example:** Compare two response rates using  $\Delta(t) = p_1(t) - p_0(t)$  at  $t$ :

$$se\{\hat{\Delta}(t)\} = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1(t)} + \frac{\pi_0(1 - \pi_0)}{n_0(t)}}.$$

Therefore, to obtain the desired power of  $1 - \beta$  to detect the alternative where the population response rates were  $\pi_1$  and  $\pi_0$ , with  $\pi_1 - \pi_0 = \Delta_A$ , we would need the sample sizes  $n_1(t^F)$  and  $n_0(t^F)$  to satisfy

$$\left\{ \frac{\pi_1(1 - \pi_1)}{n_1(t^F)} + \frac{\pi_0(1 - \pi_0)}{n_0(t^F)} \right\}^{-1} = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2.$$

**Remark:** This is essentially the same as what we did before.

- Power of group sequential test:

$$1 - P[|T(t_1)| < b_1, \dots, |T(t_K)| < b_K | \Delta = \Delta_A],$$

where  $b_j, j = 1, 2, \dots, K$  are determined by  $\alpha, K$  and  $\Phi$ .

- need to find out the maximum information (MI), similar to maximum sample size.
- If we do interim analysis after equal increment of information, then

$$\begin{bmatrix} T(t_1) \\ T(t_2) \\ \vdots \\ T(t_K) \end{bmatrix} \stackrel{H_A}{\sim} N \left( \begin{bmatrix} \Delta_A \sqrt{\frac{MI}{K}} \\ \Delta_A \sqrt{\frac{2 \times MI}{K}} \\ \vdots \\ \Delta_A \sqrt{\frac{K \times MI}{K}} \end{bmatrix}, \begin{bmatrix} 1 & \sqrt{\frac{1}{2}} & \cdots & \sqrt{\frac{1}{K}} \\ \sqrt{\frac{1}{2}} & 1 & \cdots & \sqrt{\frac{2}{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{1}{K}} & \sqrt{\frac{2}{K}} & \cdots & 1 \end{bmatrix} \right)$$

- Denote  $\delta = \Delta_A \sqrt{MI}$ . Then the distribution of  $\{T(t_1), T(t_2), \dots, T(t_K)\}$  is determined by  $\delta$  and  $K$ :

$$\begin{bmatrix} T(t_1) \\ T(t_2) \\ \vdots \\ T(t_K) \end{bmatrix} \stackrel{H_A}{\sim} \mathbf{N} \left( \begin{bmatrix} \delta \sqrt{\frac{1}{K}} \\ \delta \sqrt{\frac{2}{K}} \\ \vdots \\ \delta \sqrt{\frac{K}{K}} \end{bmatrix}, \begin{bmatrix} 1 & \sqrt{\frac{1}{2}} & \cdots & \sqrt{\frac{1}{K}} \\ \sqrt{\frac{1}{2}} & 1 & \cdots & \sqrt{\frac{2}{K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{1}{K}} & \sqrt{\frac{2}{K}} & \cdots & 1 \end{bmatrix} \right)$$

- We can find out  $\delta$  for given  $\alpha$ ,  $K$  and  $\Phi$  so that the power

$$1 - P_\delta[|T(t_1)| < b_1, \dots, |T(t_K)| < b_K] = 1 - \beta$$

- Denote this  $\delta$  by  $\delta(\alpha, K, \Phi, \beta)$ .

- In order the sequential test to have power  $1 - \beta$  to detect  $\Delta_A$ , we need:

$$\Delta_A \sqrt{MI} = \delta(\alpha, K, \Phi, \beta)$$

or

$$MI = \frac{\delta^2(\alpha, K, \Phi, \beta)}{\Delta_A^2}$$

## I. Inflation Factor

- If we do data analysis only once, then for level  $\alpha$  test to have power  $1 - \beta$  to detect  $\Delta_A$ , the information has to be

$$I^{FS} = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A} \right\}^2.$$

- The maximum information using a group sequential test is usually larger than  $I^{FS}$
- Denote the inflation factor by

$$IF(\alpha, K, \Phi, \beta) = \frac{MI}{I^{FS}} = \left\{ \frac{\delta(\alpha, K, \Phi, \beta)}{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}} \right\}^2$$

- If we know the inflation factor, we know

$$MI = IF(\alpha, K, \Phi, \beta) I^{FS}$$

- This can be used to calculate maximum sample size needed.

Table 3: *Inflation factors as a function of  $K$ ,  $\alpha$ ,  $\beta$  and  $\Phi$* 

		$\alpha=0.05$			$\alpha=0.01$		
		Power= $1-\beta$			Power= $1-\beta$		
$K$	Boundary	0.80	0.90	0.95	0.80	0.90	0.95
2	Pocock	1.11	1.10	1.09	1.09	1.08	1.08
	O-F	1.01	1.01	1.01	1.00	1.00	1.00
3	Pocock	1.17	1.15	1.14	1.14	1.12	1.12
	O-F	1.02	1.02	1.02	1.01	1.01	1.01
4	Pocock	1.20	1.18	1.17	1.17	1.15	1.14
	O-F	1.02	1.02	1.02	1.01	1.01	1.01
5	Pocock	1.23	1.21	1.19	1.19	1.17	1.16
	O-F	1.03	1.03	1.02	1.02	1.01	1.01
6	Pocock	1.25	1.22	1.21	1.20	1.19	1.17
	O-F	1.03	1.03	1.03	1.02	1.02	1.02
7	Pocock	1.26	1.24	1.22	1.22	1.20	1.18
	O-F	1.03	1.03	1.03	1.02	1.02	1.02

- **Example with dichotomous endpoint:** Let  $\pi_1$  and  $\pi_0$  be the population response rates for treatments 1 and 0.  $\Delta = \pi_1 - \pi_0$ .

Want to test  $H_0 : \Delta = 0$  vs.  $H_A : \Delta \neq 0$  at level  $\alpha = 0.05$  using a 4-look O'Brien-Fleming boundary ( $\Phi = 0$ ).

- ★ we will reject  $H_0$  the first time when

$$\begin{aligned} |T(t_j)| &= \left| \frac{p_1(t_j) - p_0(t_j)}{\sqrt{\bar{p}(t_j)\{1 - \bar{p}(t_j)\} \left\{ \frac{1}{n_1(t_j)} + \frac{1}{n_0(t_j)} \right\}}} \right| \\ &\geq 4.049/\sqrt{j}, \quad j = 1, \dots, 4. \end{aligned}$$

The boundaries are given by

Table 4: *Boundaries for a 4-look O-F test*

$j$	$b_j$	nominal p-value
1	4.05	.001
2	2.86	.004
3	2.34	.019
4	2.03	.043

- ★ Suppose  $\pi_0 = 0.3$  and we would like to have power 90% to detect  $\pi_1 = 0.45$ , how do we design the study?
- ★ The fixed sample size design requires

$$n^{FS} = \left\{ \frac{1.96 + 1.28 \sqrt{\frac{.3 \times .7 + .45 \times .55}{2 \times .375 \times .625}}}{.15} \right\}^2 \times 4 \times .375 \times .625 = 434$$

- ★ The inflation factor for  $\alpha = 0.05$ , power=0.9,  $K = 4$  and  $\Phi = 0$  is IF=1.02. The maximum sample size using a group sequential test is

$$1.02 \times 434 = 444,$$

or 222 patients for each treatment

- ★ Since  $222/4 = 56$ , we recruit 56 patients for each treatment first and then do interim analysis. If we don't reject  $H_0$ , then recruit additional 56 patients to each treatment and do interim analysis, etc.

## Information based monitoring

- $\pi_0 = 0.3$  and  $\pi_1 = 0.45$  are needed to derive the sample size. They might not be the case in practice.
- Suppose we would like a level 0.05 test to have power 0.9 to detect  $\Delta = 0.15$ , then the information needed for a fixed sample size design is

$$\left\{ \frac{Z_{\alpha/2} + Z_{\beta}}{\Delta_A} \right\}^2 = \left\{ \frac{1.96 + 1.28}{.15} \right\}^2 = 466.6$$

- So the MI for a 4-look O-F design is  $466.6 \times 1.02 = 475.9$
- So the information required at the  $j$ th interim analysis

$$\frac{j \times 475.9}{4} = 119 \times j, \quad j = 1, \dots, 4$$

- The information available at the  $j$ th interim analysis is approximately

$$[se\{\hat{\Delta}(t)\}]^{-2} = \left[ \frac{p_1(t)\{1 - p_1(t)\}}{n_1(t)} + \frac{p_0(t)\{1 - p_0(t)\}}{n_0(t)} \right]^{-1}$$

- This implies that we should do interim analysis when

$$\left[ \frac{p_1(t_j)\{1 - p_1(t_j)\}}{n_1(t_j)} + \frac{p_0(t_j)\{1 - p_0(t_j)\}}{n_0(t_j)} \right]^{-1} = 119 \times j, \quad j = 1, \dots, 4.$$

and use the test statistic and the boundary values given before.

- The information-based monitoring will maintain the overall type I error prob and desired power to detect a difference of interest even if the nuisance parameter values may be different than what we assume.

## II. Average information

- For the same design characteristics, which boundary is better? Pocock or O-F?
- Pocock design has higher IF, but it is easier to stop using Pocock design.
- Compare them using average information (similar to average sample size) needed for the alternative
- If  $H_0$  is true, the chance that  $H_0$  will be rejected will be small ( $\alpha$  is usually taken to be 0.05). So the chance the trial will be stopped is small too (at most  $\alpha$ ). So the average information under  $H_0$  will be very close to the MI

- For example, if  $K = 5$ ,  $\alpha = 0.05$ , power = 90% to detect an alternative of interest, then

Designs	Maximum information	Average information ( $H_A$ )
5-look Pocock	$I^{FS} \times 1.21$	$I^{FS} \times .68$
5-look O-F	$I^{FS} \times 1.03$	$I^{FS} \times .75$
Fixed-sample	$I^{FS}$	$I^{FS}$

where

$$I^{FS} = \left\{ \frac{Z_{\alpha/2} + Z_{\beta}}{\Delta_A} \right\}^2$$

is the information for fixed sample size design.

- Pocock designs required smaller sample size on average if  $H_A$  is true.

- **Remarks:**

- ★ If you want a design which, on average, stops the study with less information when there truly is a clinically important treatment difference, while preserving the level and power of the test, then a Pocock boundary is preferred to the O-F boundary.
- ★ By a numerical search, one can derive the “optimal” shape parameter  $\Phi$  which minimizes the average information under the clinically important alternative  $\Delta_A$  for  $\alpha$ ,  $K$ , and power  $(1 - \beta)$ . For example, when  $K = 5$ ,  $\alpha = .05$  and power of 90% the optimal shape parameter  $\Phi = .45$ , very close to the Pocock boundary (Wang and Tsiatis, 1987, *Biometrics*).
- ★ However, the designs with smaller average information under  $H_A$  requires more information if the null hypothesis were true.
- ★ Most clinical trials with a monitoring plan seem to favor more “conservative” designs such as the O-F design.

## Statistical Reasons

1. Historically, most clinical trials fail to show a significant difference; hence, from a global perspective it is more cost efficient to use conservative designs (such as O-F design)
2. Even a conservative design, such as O-F, results in a substantial reduction in average information, under the alternative  $H_A$ , before a trial is completed as compared to a fixed-sample design (in our example .75 average information) with only a modest increase in the maximum information (1.03 in our example).

## Non-statistical Reasons

3. In the early stages of a clinical trial, the data are less reliable and possibly unrepresentative for a variety of logistical reasons. It is therefore preferable to make it more difficult to stop early during these early stages.
4. Psychologically, it is preferable to have a nominal p-value at the end of the study which is close to .05. The nominal p-value at the final analysis for the 5-look O-F test is .041 as compare to .016 for the 5-look Pocock test. This minimizes the embarrassing situation where, say, a p-value of .03 at the final analysis would have to be declared not significant for those using a Pocock design.

### III. Steps in the design and analysis of group-sequential tests with equal increments of information

#### Design

1. Decide the maximum number of looks  $K$  and the boundary  $\Phi$ .  $K$  does not have to be very large.

Table 5: *O'Brien-Fleming boundaries* ( $\Phi = 0$ );  $\alpha = .05$ ,  $power = .90$

	Maximum	Average
$K$	Information	Information ( $H_A$ )
1	$I^{FS}$	$I^{FS}$
2	$I^{FS} \times 1.01$	$I^{FS} \times .85$
3	$I^{FS} \times 1.02$	$I^{FS} \times .80$
4	$I^{FS} \times 1.02$	$I^{FS} \times .77$
5	$I^{FS} \times 1.03$	$I^{FS} \times .75$

2. Compute  $I^{FS}$ , then translate it to the sample size or number of events.
3. Find the inflation factor  $IF(\alpha, K, \Phi, \beta)$  and get

$$MI = I^{FS} \times IF(\alpha, K, \Phi, \beta).$$

Also calculate the maximum sample size or maximum number of events.

## Analysis

4. Conduct data analysis after equal increment of  $MI/K$  information. This can be achieved by monitoring  $[se\{\hat{\Delta}(t)\}]^{-2}$ , although in practice, this is not generally how the analysis times are determined.
5. At the  $j$ -th interim analysis, the standardized test statistic

$$T(t_j) = \frac{\hat{\Delta}(t_j)}{se\{\hat{\Delta}(t_j)\}},$$

is computed using all the data accumulated until that time and the null hypothesis is rejected the first time the test statistic exceeds the corresponding boundary value.

**Note:** The procedure outlined above will have the correct level of significance as long as the interim analysis are conducted after equal increments of information.

However, in order for this test to have the desired power to detect  $\Delta_A$ , it must be computed after equal increments of statistical information

$MI/K$  where

$$MI = \left\{ \frac{Z_{\alpha/2} + Z_{\beta}}{\Delta_A} \right\}^2 IF(\alpha, K, \Phi, \beta).$$

If the initial guesses on the nuisance parameters were correct, then we would have the right power. Otherwise the study may be underpowered or overpowered.

We should monitor

$$[se\{\hat{\Delta}(t_j)\}]^{-2}$$

to see if it deviates significantly from the required information

$$j \times MI/K.$$

This helps detect the problem and fix it at the early stage.