ST 520
Statistical Principles of Clinical Trials

Lecture Notes

(Modified from Dr. A. Tsiatis' Lecture Notes)

Daowen Zhang

Department of Statistics

North Carolina State University

Contents

1 Introduction

1.1 Scope and objectives

The focus of this course will be on the statistical methods and principles used to study disease and its prevention or treatment in human populations. There are two broad subject areas in the study of disease; **Epidemiology** and **Clinical Trials**. This course will be devoted almost entirely to statistical methods in Clinical Trials research but we will first give a very brief introduction to Epidemiology in this Section.

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)

1.2 Brief Introduction to Epidemiology

Cross-sectional study

In a cross-sectional study the data are obtained from a random sample of the population at one point in time. This gives a snapshot of a population.
Example: Based on a single survey of a specific population or a random sample thereof, we determine the proportion of individuals with heart disease at one point in time. This is referred to as the **prevalence** of disease. We may also collect demographic and other information which will allow us to break down prevalence broken by age, race, sex, socio-economic status, geographic, etc.

Important public health information can be obtained this way which may be useful in determining how to allocate health care resources. However such data are generally not very useful in determining causation.

In an important special case where the exposure and disease are dichotomous, the 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.

In this case, all counts except n_{++}, the sample size, are random variables. The counts $(n_{11}, n_{12}, n_{21}, n_{22})$ have the following distribution:

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{multinomial}(n_{++}, P[DE], P[\bar{D}E], P[D\bar{E}], P[\bar{D}\bar{E}]).$$

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

$$\text{prevalence of disease } P[D] \text{ (estimated by } \frac{n_{+1}}{n_{++}})$$
$$\text{probability of exposure } P[E] \text{ (estimated by } \frac{n_{1+}}{n_{++}})$$

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

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

...

We can also assess the **association** between the exposure and disease using the data from a cross-sectional study. One such measure is **relative risk**, which is defined as

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

It is easy to see that the **relative risk** ψ has the following properties:

- $\psi > 1 \Rightarrow$ positive association; that is, the exposed population has higher disease probability than the unexposed population.
- $\psi = 1 \Rightarrow$ no association; that is, the exposed population has the same disease probability as the unexposed population.
- $\psi < 1 \Rightarrow$ negative association; that is, the exposed population has lower disease probability than the unexposed population.

Of course, we cannot state that the exposure E causes the disease D even if $\psi > 1$, or vice versa. In fact, the exposure E may not even occur before the event D.

Since we got good estimates of $P[D|E]$ and $P[D|\bar{E}]$

$$\widehat{P}[D|E] = \frac{n_{11}}{n_{1+}}, \quad \widehat{P}[D|\bar{E}] = \frac{n_{21}}{n_{2+}},$$

the relative risk ψ can be estimated by

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

Another measure that describes the association between the exposure and the disease is the **odds ratio**, which is defined as

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

Note that $P[D|E]/(1-P[D|E])$ is called the **odds** of $P[D|E]$. It is obvious that

- $\psi > 1 \Longleftrightarrow \theta > 1$
- $\psi = 1 \Longleftrightarrow \theta = 1$
- $\psi < 1 \Longleftrightarrow \theta < 1$

Given data from a cross-sectional study, the **odds ratio** θ can be estimated by

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

It can be shown that the variance of $\log(\widehat{\theta})$ has a very nice form given by

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

The point estimate $\widehat{\theta}$ and the above variance estimate can be used to make inference on θ. Of course, the total sample size n_{++} as well as each cell count have to be large for this variance formula to be reasonably good.

A $(1-\alpha)$ confidence interval (CI) for $\log(\theta)$ (log odds ratio) is

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

Exponentiating the two limits of the above interval will give us a CI for θ with the same confidence level $(1-\alpha)$.

Alternatively, the variance of $\widehat{\theta}$ can be estimated (by the delta method)

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

and a $(1-\alpha)$ CI for θ is obtained as

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

For example, if we want a 95% confidence interval for $\log(\theta)$ or θ, we will use $z_{0.05/2} = 1.96$ in the above formulas.

From the definition of the odds-ration, we see that 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

$$\theta \approx \psi.$$

This approximation is very useful. Since the relative risk ψ has a much better interpretation (and hence it is easier to communicate with biomedical researchers using this measure), in studies where we cannot estimate the relative risk ψ but we can estimate the odds-ratio θ (see **retrospective studies** later), if the disease under studied is a rare one, we can approximately estimate the relative risk by the odds-ratio estimate.

Longitudinal studies

In a longitudinal study, subjects are followed over time and single or multiple measurements of the variables of interest are obtained. Longitudinal epidemiological studies generally fall into two categories; **prospective** i.e. moving forward in time or **retrospective** going backward in time. We will focus on the case where a single measurement is taken.

Prospective study: In a 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.

If there is only one exposure variable which is binary, the 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+}

Since the cohorts are identified by the researcher, n_{1+} and n_{2+} are fixed sample sizes for each group. In this case, only n_{11} and n_{21} are random variables, and these random variables have the following distributions:

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

From these distributions, $P[D|E])$ and $P[D|\bar{E}]$ can be readily estimated by

$$\widehat{P}[D|E] = \frac{n_{11}}{n_{1+}}, \quad \widehat{P}[D|\bar{E}] = \frac{n_{21}}{n_{2+}}.$$

The relative risk ψ and the odds-ratio θ defined previously can be estimated in exactly the same way (have the same formula). So does the variance estimate of the odds-ratio estimate.

One problem of a prospective study is that some subjects may drop out from the study before developing the disease under study. In this case, the disease probability has to be estimated differently. This is illustrated by the following example.

Example: 40,000 British doctors were followed for 10 years. The following data were collected:

Table 1.1: *Death Rate from Lung Cancer per 1000 person years.*

# cigarettes smoked per day	death rate
0	.07
1-14	.57
15-24	1.39
35+	2.27

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

Remark: If we denote by T the time to death due to lung cancer, the death rate at time t is defined by

$$\lambda(t) = \lim_{h \to 0} \frac{P[t \le T < t+h | T \ge t]}{h}.$$

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

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

In this case, if we are interested in the event

$$D = \text{die from lung cancer within next one year} \mid \text{still alive now},$$

or statistically,

$$D = [t \leq T < t+1 | T \geq t],$$

then

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

Roughly speaking, assuming the death rate remains constant over the 10 year period for each group of doctors, we can take the rate above divided by 1000 to approximate the probability of death from lung cancer in one year. For example, the estimated probability of dying from lung cancer in one year for British doctors smoking between 15-24 cigarettes per day at the beginning of the study is $\widehat{P}[D] = 1.39/1000 = 0.00139$. Similarly, the estimated probability of dying from lung cancer in one year for the heaviest smokers is $\widehat{P}[D] = 2.27/1000 = 0.00227$.

From the table above we note that the relative risk of death from lung cancer between heavy smokers and non-smokers (in the same time window) is $2.27/0.07 = 32.43$. That is, heavy smokers are estimated to have 32 times the risk of dying from lung cancer as compared to non-smokers.

Certainly the value 32 is subject to statistical variability and moreover we must be concerned whether these results imply causality.

We can also estimate the odds-ratio of dying from lung cancer in one year between heavy smokers and non-smokers:

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

This estimate is essentially the same as the estimate of the relative risk 32.43.

Retrospective study: Case-Control

A very popular design in many epidemiological studies is the case-control design. In such a 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 relative risk as we will demonstrate.

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:

smoke	cases	controls
yes	1,289	921
no	68	436

We would like to estimate the relative risk ψ or the odds-ratio θ of getting lung cancer between smokers and non-smokers.

Before tackling this problem, let us look at a general problem. The above data can be represented by the following 2×2 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 the counts n_{11} and n_{12} are random variables having the following distributions:

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

By definition, the relative risk ψ is

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

We can estimate ψ if we can estimate these probabilities $P[D|E]$ and $P[D|\bar{E}]$. However, we cannot use the same formulas we used before for cross-sectional or prospective study to estimate them.

What is the consequence of using the same formulas we used before? The formulas would lead to the following incorrect estimates:

$$\widehat{P}[D|E] = \frac{n_{11}}{n_{1+}} = \frac{n_{11}}{n_{11}+n_{12}} \ \textbf{(incorrect!)}$$
$$\widehat{P}[D|\bar{E}] = \frac{n_{21}}{n_{2+}} = \frac{n_{21}}{n_{21}+n_{22}} \ \textbf{(incorrect!)}$$

Since we choose n_{+1} and n_{+2}, we can fix n_{+2} at some number (say, 50), and let n_{+1} grow (sample more cases). As long as $P[E|D] > 0$, n_{11} will also grow. Then $\widehat{P}[D|E] \longrightarrow 1$. Similarly $\widehat{P}[D|\bar{E}] \longrightarrow 1$. Obviously, these are NOT sensible estimates.

For example, if we used the above formulas for our example, we would get:

$$\begin{aligned}
\widehat{P}[D|E] &= \frac{1289}{1289+921} = 0.583 \text{ (\textbf{incorrect!})} \\
\widehat{P}[D|\bar{E}] &= \frac{68}{68+436} = 0.135 \text{ (\textbf{incorrect!})} \\
\widehat{\psi} &= \frac{\widehat{P}[D|E]}{\widehat{P}[D|\bar{E}]} = \frac{0.583}{0.135} = 4.32 \text{ (\textbf{incorrect!})}.
\end{aligned}$$

This incorrect estimate of the relative risk will be contrasted with the estimate from the correct method.

We introduced the odds-ratio before to assess the association between the exposure (E) and the disease (D) as follows:

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

and we stated that if the disease under study is a rare one, then

$$\theta \approx \psi.$$

Since we cannot directly estimate the relative risk ψ from a retrospective (case-control) study due to its design feature, let us try to estimate the odds-ratio θ.

For this purpose, we would like to establish the following equivalence:

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

By Bayes' theorem, we have for any two events A and B

$$P[A|B] = \frac{P[AB]}{P[B]} = \frac{P[B|A]P[A]}{P[B]}.$$

Therefore,

$$\begin{aligned}
\frac{P[D|E]}{P[D|\bar{E}]} &= \frac{P[E|D]P[D]/P[E]}{P[\bar{E}|D]P[D]/P[\bar{E}]} = \frac{P[E|D]/P[E]}{P[\bar{E}|D]/P[\bar{E}]} \\
\frac{P[\bar{D}|E]}{P[\bar{D}|\bar{E}]} &= \frac{P[E|\bar{D}]P[\bar{D}]/P[E]}{P[\bar{E}|\bar{D}]P[\bar{D}]/P[\bar{E}]} = \frac{P[E|\bar{D}]/P[E]}{P[\bar{E}|\bar{D}]/P[\bar{E}]},
\end{aligned}$$

and

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

Notice that the quantity in the right hand side is in fact the odds-ratio of being exposed between cases and controls, and the above identity says that the odds-ratio of getting disease between exposed and un-exposed is the **same** as the odds-ratio of being exposed between cases and controls. This identity is very important since by design we are able to estimate the odds-ratio of being exposed between cases and controls since we are able to estimate $P[E|D]$ and $E|\bar{D}]$ from a case-control study:

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

So θ can be estimated by

$$\widehat{\theta} = \frac{\widehat{P}[E|D]/(1-\widehat{P}[E|D])}{\widehat{P}[E|\bar{D}]/(1-\widehat{P}[E|\bar{D}])} = \frac{n_{11}/n_{+1}/(1-n_{11}/n_{+1})}{n_{12}/n_{+2}/(1-n_{12}/n_{+2})} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

which has exactly the same form as the estimate from a cross-sectional or prospective study. This means that the odds-ratio estimate is **invariant** to the study design.

Similarly, it can be shown that the variance of $\log(\widehat{\theta})$ can be estimated by the same formula we used before

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

Therefore, inference on θ or $\log(\theta)$ such as constructing a confidence interval will be exactly the same as before.

Going back to the lung cancer example, we got the following estimate of the odds ratio:

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

If lung cancer can be viewed as a rare event, we estimate the relative risk of getting lung cancer between smokers and non-smokers to be about nine fold. This estimate is much higher than the incorrect estimate (4.32) we got on page 9.

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 can give a more precise estimate of relative risk with the same number of patients than a prospective design. This is because the number of cases, which in a prospective study is small, would be over-represented by design in a case control study. This will be illustrated in a homework exercise.
- Cons
  - It may be difficult to get accurate information on the exposure status of cases and controls. The records may not be that good and depending on individuals' memory may not be very reliable. This can be a severe drawback.

1.3 Brief Introduction and History of Clinical Trials

The following are several definitions of a clinical trial that were found in different textbooks and articles.

- 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

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

"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.

Interestingly, although the trial appeared conclusive, Lind continued to propose "pure dry air" as the first priority with fruit and vegetables as a secondary recommendation. Furthermore, almost 50 years elapsed before the British navy supplied lemon juice to its ships.

Pre-20th century medical experimenters had no appreciation of the scientific method. A common medical treatment before 1800 was blood letting. It was believed that you could get rid of an ailment or infection by sucking the bad blood out of sick patients; usually this was accomplished by applying leeches to the body. There were numerous anecdotal accounts of the effectiveness of such treatment for a myriad of diseases. The notion of systematically collecting data to address specific issues was quite foreign.

1794 Rush *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 (1834): Lays a clear foundation for the use of the *numerical method* in assessing therapies.

"As to different methods of treatment, if it is possible for us to assure ourselves of the superiority of one or other among them in any disease whatever, having regard to the different circumstances

Table 1.2: 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%

of age, sex and temperament, of strength and weakness, it is doubtless to be done by enquiring if under these circumstances a greater number of individuals have been cured by one means than another. Here again it is necessary to count. And it is, in great part at least, because hitherto this method has been not at all, or rarely employed, that the science of therapeutics is still so uncertain; that when the application of the means placed in our hands is useful we do not know the bounds of this utility."

He goes on to discuss the need for

- The exact observation of patient outcome
- Knowledge of the natural progress of untreated controls
- Precise definition of disease prior to treatment
- Careful observations of deviations from intended treatment

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.

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. 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 children participated in the largest trial ever to assess the effectiveness of the Salk vaccine in preventing paralysis or death from poliomyelitis.

Such a large number was needed because the incidence rate of polio was about 1 per 2,000 and evidence of treatment effect was needed as soon as possible so that vaccine could be routinely given if found to be efficacious.

There were two components (randomized and non-randomized) to this trial. For the non-randomized component, one million children in the first through third grades participated. The second graders were offered vaccine whereas first and third graders formed the control group. There was also a randomized component where .8 million children were randomized in a **double-blind placebo-controlled** trial.

The incidence of polio in the randomized vaccinated group was less than half that in the control group and even larger differences were seen in the decline of paralytic polio.

The nonrandomized group supported these results; however non-participation by some who were offered vaccination might have cast doubt on the results. 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). This may cast doubt whether first and third graders (control group) have the same likelihood for getting polio as second graders. This is a basic assumption that needs to be satisfied in order to make unbiased treatment comparisons. Luckily, there was a randomized component to the study where the two groups (vaccinated) versus (control) were guaranteed to be similar on average by design.

Note: During the course of the semester there will be a great deal of discussion on the role of randomization and compliance and their effect on making causal statements.

Government sponsored studies

In the 1950's the National Cancer Institute (NCI) organized randomized clinical trials in acute leukemia. The successful organization of this particular clinical trial led to the formation of two collaborative groups; CALGB (Cancer and Leukemia Group B) and ECOG (Eastern Cooperative Oncology Group). More recently SWOG (Southwest Oncology Group) and POG (Pediatrics Oncology Group) have been organized. A Cooperative group is an organization with many participating hospitals throughout the country (sometimes world) that agree to conduct common clinical trials to assess treatments in different disease areas.

Government sponsored clinical trials are now routine. As well as the NCI, these include the following organizations of the National Institutes of Health.

- NHLBI- (National Heart Lung and Blood Institute) funds individual and often very large studies in heart disease. To the best of my knowledge there are no cooperative groups funded by NHLBI.
- NIAID- (National Institute of Allergic and Infectious Diseases) Much of their funding now goes to clinical trials research for patients with HIV and AIDS. The ACTG (AIDS Clinical Trials Group) is a large cooperative group funded by NIAID.
- NIDDK- (National Institute of Diabetes and Digestive and Kidney Diseases). Funds large scale clinical trials in diabetes research. Recently formed the cooperative group TRIALNET (network 18 clinical centers working in cooperation with screening sites throughout the United States, Canada, Finland, United Kingdom, Italy, Germany, Australia, and New Zealand - for type 1 diabetes)

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. Because the evaluation of drugs and the conduct, design and analysis of clinical trials depends so heavily on sound Statistical Methodology this has resulted in an explosion of statisticians working for th Pharmaceutical Industry and wonderful career opportunities.

2 Phase I and II clinical trials

2.1 Phases of Clinical Trials

The process of drug development can be broadly classified as pre-clinical and clinical. Pre-clinical refers to experimentation that occurs before it is given to human subjects; whereas, clinical refers to experimentation with humans. This course will consider only clinical research. It will be assumed that the drug has already been developed by the chemist or biologist, tested in the laboratory for biologic activity (in vitro), that preliminary tests on animals have been conducted (in vivo) and that the new drug or therapy is found to be sufficiently promising to be introduced into humans.

Within the realm of clinical research, clinical trials are classified into four phases.

- **Phase I**: To explore possible toxic effects of drugs and determine a tolerated dose for further experimentation. Also during Phase I experimentation the pharmacology of the drug may be explored.
- **Phase II**: Screening and feasibility by initial assessment for therapeutic effects; further assessment of toxicities.
- **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.

These definitions of the four phases are not hard and fast. Many clinical trials blur the lines between the phases. Loosely speaking, the logic behind the four phases is as follows:

A new promising drug is about to be assessed in humans. The effect that this drug might have on humans is unknown. We might have some experience on similar acting drugs developed in the past and we may also have some data on the effect this drug has on animals but we are not sure what the effect is on humans. To study the initial effects, a Phase I study is conducted. Using increasing doses of the drug on a small number of subjects, the possible side effects of the drug are documented. It is during this phase that the tolerated dose is determined for future experimentation. The general dogma is that the therapeutic effect of the drug will increase with dose, but also the toxic effects will increase as well. Therefore one of the goals of a Phase I study is to determine what the maximum dose should be that can be reasonably tolerated by most individuals with the disease. The determination of this dose is important as this will be used in future studies when determining the effectiveness of the drug. If we are too conservative then we may not be giving enough drug to get the full therapeutic effect. On the other hand if we give too high a dose then people will have adverse effects and not be able to tolerate the drug.

Once it is determined that a new drug can be tolerated and a dose has been established, the focus turns to whether the drug is good. Before launching into a costly large-scale comparison of the new drug to the current standard treatment, a smaller feasibility study is conducted to assess whether there is sufficient efficacy (activity of the drug on disease) to warrant further investigation. This occurs during phase II where drugs which show little promise are screened out.

If the new drug still looks promising after phase II investigation it moves to Phase III testing where a comparison is made to a current standard treatment. These studies are generally large enough so that important treatment differences can be detected with sufficiently large probability. These studies are conducted carefully using sound statistical principles of experimental design established for clinical trials to make objective and unbiased comparisons. It is on the basis of such Phase III clinical trials that new drugs are approved by regulatory agencies (such as FDA) for the general population of individuals with the disease for which this drug is targeted.

Once a drug is on the market and a large number of patients are taking it, there is always the possibility of rare but serious side effects that can only be detected when a large number are given treatment for sufficiently long periods of time. It is important that a monitoring system be in place that allows such problems, if they occur, to be identified. This is the role of Phase IV studies.

A brief discussion of phase I studies and designs and Pharmacology studies will be given based on the slides from Professor Marie Davidian, an expert in pharmacokinetics. Slides on phase I and pharmacology will be posted on the course web page.

2.2 Phase II clinical trials

After a new drug is tested in phase I for safety and tolerability, a dose finding study is sometimes conducted in phase II to identify a lowest dose level with good efficacy (close to the maximum efficacy achievable at tolerable dose level). In other situations, a phase II clinical trial uses a fixed dose chosen on the basis of a phase I clinical trial. The total dose is either fixed or may vary depending on the weight of the patient. There may also be provisions for modification of the dose if toxicity occurs. The study population are patients with a specified disease for which the treatment is targeted.

The primary objective is to determine whether the new treatment should be used in a large-scale comparative study. Phase II trials are used to assess

- feasibility of treatment
- side effects and toxicity
- logistics of administration and cost

The major issue that is addressed in a phase II clinical trial is whether there is enough evidence of efficacy to make it worth further study in a larger and more costly clinical trial. In a sense this is an initial screening tool for efficacy. During phase II experimentation the treatment efficacy is often evaluated on **surrogate markers**; i.e on an outcome that can be measured quickly and is believed to be related to the clinical outcome.

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, tumor shrinkage is a surrogate marker for overall survival time. The belief is that if the drug has no effect on tumor shrinkage it is unlikely to have an effect on the patient's overall survival and hence should be eliminated from further consideration. Unfortunately, there are many instances where a drug has a short term effect on a surrogate endpoint but ultimately may not have the long term effect on the clinical endpoint of ultimate interest. Furthermore, sometimes a drug may have beneficial effect through a biological mechanism that is not detected by the surrogate endpoint. Nonetheless, there must be some attempt at limiting the number of drugs that will be considered for further testing or else the system would be overwhelmed.

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

Most often, phase II clinical trials do not employ formal comparative designs. That is, they do not use parallel treatment groups. Often, phase II designs employ more than one stage; i.e. one group of patients are given treatment; if no (or little) evidence of efficacy is observed, then the trial is stopped and the drug is declared a failure; otherwise, more patients are entered in the next stage after which a final decision is made whether to move the drug forward or not.

2.2.1 Statistical Issues and Methods

One goal of a phase II trial is to estimate an endpoint related to treatment efficacy with sufficient precision to aid the investigators in determining whether the proposed treatment should be studied further.

Some examples of endpoints are:

- proportion of patients responding to treatment (response has to be unambiguously defined)
- proportion with side effects
- average decrease in blood pressure over a two week period

A statistical perspective is generally taken for estimation and assessment of precision. That is, the problem is often posed through a statistical model with population parameters to be estimated and confidence intervals for these parameters to assess precision.

Example: Suppose we consider patients with esophogeal cancer treated with chemotherapy prior to surgical resection. A complete response is defined as an absence of macroscopic or microscopic tumor at the time of surgery. We suspect that this may occur with 35% (guess) probability using a drug under investigation in a phase II study. The 35% is just a guess, possibly based on similar acting drugs used in the past, and the goal is to estimate the actual response rate with sufficient precision, in this case we want the 95% confidence interval to be within 15% of the truth.

As statisticians, we view the world as follows: We start by positing a statistical model; that is, let π denote the population complete response rate. We conduct an experiment: n patients with esophogeal cancer are treated with the chemotherapy prior to surgical resection and we collect data: the number of patients who have a complete response.

The result of this experiment yields a random variable X, the number of patients in a sample of size n that have a complete response. A popular model for this scenario is to assume that

$$X \sim binomial(n, \pi);$$

that is, the random variable X is distributed with a binomial distribution with sample size n and success probability π. The goal of the study is to estimate π and obtain a confidence interval.

I believe it is worth stepping back a little and discussing how the actual experiment and the statistical model used to represent this experiment relate to each other and whether the implicit assumptions underlying this relationship are reasonable.

Statistical Model

What is the population? All people now and in the future with esophogeal cancer who would be eligible for the treatment.

What is π? (the population parameter)

If all the people in the hypothetical population above were given the new chemotherapy, then π would be the proportion who would have a complete response. This is a hypothetical construct. Neither can we identify the population above or could we actually give them all the chemotherapy. Nonetheless, let us continue with this mind experiment.

We assume the random variable X follows a binomial distribution. Is this reasonable? Let us review what it means for a random variable to follow a binomial distribution.

X being distributed as a binomial $b(n, \pi)$ means that X corresponds to the number of successes (complete responses) in n independent trials where the probability of success for each trial is equal to π. This would be satisfied, for example, if we were able to identify every member of the population and then, using a random number generator, chose n individuals at random from our population to test and determine the number of complete responses.

Clearly, this is not the case. First of all, the population is a hypothetical construct. Moreover, in most clinical studies the sample that is chosen is an **opportunistic** sample. There is generally no attempt to randomly sample from a specific population as may be done in a survey sample. Nonetheless, a statistical perspective may be a useful construct for assessing variability. I sometimes resolve this in my own mind by thinking of the hypothetical population that I can make inference on as all individuals who might have been chosen to participate in the study with whatever process that was actually used to obtain the patients that were actually studied. However, this limitation must always be kept in mind when one extrapolates the results of a clinical experiment to a more general population.

Philosophical issues aside, let us continue by assuming that the posited model is a reasonable approximation to some question of relevance. Thus, we will assume that our data is a realization of the random variable X, assumed to be distributed as $b(n, \pi)$, where π is the population parameter of interest.

Reviewing properties about a binomial distribution we note the following:

- $E(X) = n\pi$, where $E(\cdot)$ denotes expectation of the random variable.
- $Var(X) = n\pi(1-\pi)$, where $Var(\cdot)$ denotes the variance of the random variable.
- $P(X = k) = \begin{pmatrix} n \\ k \end{pmatrix} \pi^k (1-\pi)^{n-k}$, where $P(\cdot)$ denotes probability of an event, and $\begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{k!(n-k)!}$
- 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 π and variance $\pi(1-\pi)/n$:

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

This approximation is useful for inference regarding the population parameter π. Because of the approximate normality, the estimator p will be within 1.96 standard deviations of π approximately 95% of the time. (Approximation gets better with increasing sample size). Therefore the population parameter π will be within the interval

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

with approximately 95% probability. Since the value π is unknown to us, we approximate using p to obtain the approximate 95% confidence interval for π, namely

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

Going back to our example, where our best guess for the response rate 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 π, then we need

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

or

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

Since the response rate of 35% is just a guess which is made before data are collected, the exercise above should be repeated for different feasible values of π before finally deciding on how large the sample size should be.

Exact Confidence Intervals

If either $n\pi$ or $n(1-\pi)$ is small, then the normal approximation given above may not be adequate for computing accurate confidence intervals. In such cases we can construct **exact** (usually conservative) confidence intervals.

We start by reviewing the definition of a confidence interval and then show how to construct an exact confidence interval for the parameter π of a binomial distribution.

Definition: The definition of a $(1-\alpha)$-th confidence region (interval) for the parameter π is as follows:

For each realization of the data $X = k$, a region of the parameter space, denoted by $\mathcal{C}(k)$ (usually an interval) is defined in such a way that the random region $C(X)$ contains the true value of the parameter with probability greater than or equal to $(1-\alpha)$ regardless of the value of the parameter. That is,

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

where $P_\pi(\cdot)$ denotes probability calculated under the assumption that $X \sim b(n,\pi)$ and $\supset$ denotes "contains". The confidence interval is the random interval $\mathcal{C}(X)$. After we collect data and obtain the realization $X = k$, then the corresponding confidence interval is defined as $\mathcal{C}(k)$.

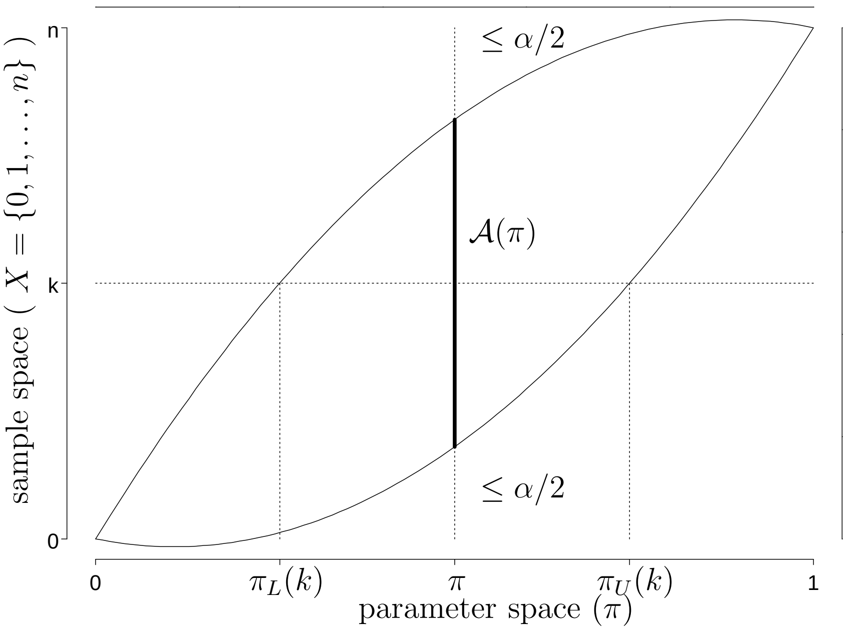
This definition is equivalent to defining an acceptance region (of the sample space) for each value π, denoted as $\mathcal{A}(\pi)$, that has probability greater than equal to $1-\alpha$, i.e.

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

in which case $\mathcal{C}(k) = \{\pi : k \in \mathcal{A}(\pi)\}$.

We find it useful to consider a graphical representation of the relationship between confidence intervals and acceptance regions.

Figure 2.1: *Exact confidence intervals*



Another way of viewing a $(1-\alpha)$-th confidence interval is to find, for each realization $X = k$, all the values π^* for which the value k would not reject the hypothesis $H_0 : \pi = \pi^*$. Therefore, a $(1-\alpha)$-th confidence interval is sometimes more appropriately called a $(1-\alpha)$-th credible region (interval).

If $X \sim b(n, \pi)$, then when $X = k$, the $(1-\alpha)$-th confidence interval is given by

$$\mathcal{C}(k) = [\pi_L(k), \pi_U(k)],$$

where $\pi_L(k)$ denotes the lower confidence limit and $\pi_U(k)$ the upper confidence limit, which are defined as

$$P_{\pi_L(k)}(X \ge k) = \sum_{j=k}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) \pi_L(k)^j \{1 - \pi_L(k)\}^{n-j} = \alpha/2,$$

and

$$P_{\pi_U(k)}(X \le k) = \sum_{j=0}^{k} \left(\begin{array}{c} n \\ j \end{array} \right) \pi_U(k)^j \{1 - \pi_U(k)\}^{n-j} = \alpha/2.$$

The values $\pi_L(k)$ and $\pi_U(k)$ need to be evaluated numerically as we will demonstrate shortly.

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

$$P_\pi\{\mathcal{C}(X) \supset \pi\} > 1-\alpha$$

(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: In a Phase II clinical trial, 3 of 19 patients respond to α-interferon treatment for multiple sclerosis. In order to find the exact confidence 95% interval for π for $X=k$, $k=3$, and $n=19$, we need to find $\pi_L(3)$ and $\pi_U(3)$ satisfying

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

Many textbooks have tables for $P(X \leq c)$, where $X \sim b(n,\pi)$ for some n's and π's. Alternatively, $P(X \leq c)$ can be obtained using statistical software such as SAS or R. Either way, we see that $\pi_U(3) \approx .40$. To find $\pi_L(3)$ we note that

$$P_{\pi_L(3)}(X \geq 3) = 1 - P_{\pi_L(3)}(X \leq 2).$$

Consequently, we must search for $\pi_L(3)$ such that

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

This yields $\pi_L(3) \approx .03$. Hence the "exact" 95% confidence interval for π is

$$[.03, .40].$$

In contrast, the normal approximation yields a confidence interval of

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

2.2.2 Gehan's Two-Stage Design

Discarding ineffective treatments early

If it is unlikely that a treatment will achieve some minimal level of response or efficacy, we may want to stop the trial as early as possible. For example, suppose that a 20% response rate is the lowest response rate that is considered acceptable for a new treatment. If we get no responses in n patients, with n sufficiently large, then we may feel confident that the treatment is ineffective. Statistically, this may be posed as follows: How large must n be so that if there are 0 responses among n patients we are relatively confident that the response rate is not 20% or better? If $X \sim b(n, \pi)$, and if $\pi \geq .2$, then

$$P_\pi(X = 0) = (1 - \pi)^n \leq (1 - .2)^n = .8^n.$$

Choose n so that $.8^n = .05$ or $n \ln(8) = \ln(.05)$. This leads to $n \approx 14$ (rounding up). Thus, with 14 patients, it is unlikely ($\leq .05$) that no one would respond if the true response rate was greater than 20%. Thus 0 patients responding among 14 might be used as evidence to stop the phase II trial and declare the treatment a failure.

This is the logic behind Gehan's two-stage design. Gehan suggested the following strategy: If the minimal acceptable response rate is π_0, then choose the first stage with n_0 patients such that

$$(1 - \pi_0)^{n_0} = .05; \; n_0 = \frac{\ln(.05)}{\ln(1 - \pi_0)};$$

if there are 0 responses among the first n_0 patients then stop and declare the treatment a failure; otherwise, continue with additional patients that will ensure a certain degree of predetermined accuracy in the 95% confidence interval.

If, for example, we wanted the 95% confidence interval for the response rate to be within ±15% when a treatment is considered minimally effective at $\pi_0 = 20\%$, then the sample size necessary for this degree of precision is

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

In this example, Gehan's design would treat 14 patients initially. If none responded, the treatment would be declared a failure and the study stopped. If there was at least one response, then another 14 patients would be treated and a 95% confidence interval for π would be computed using the data from all 28 patients.

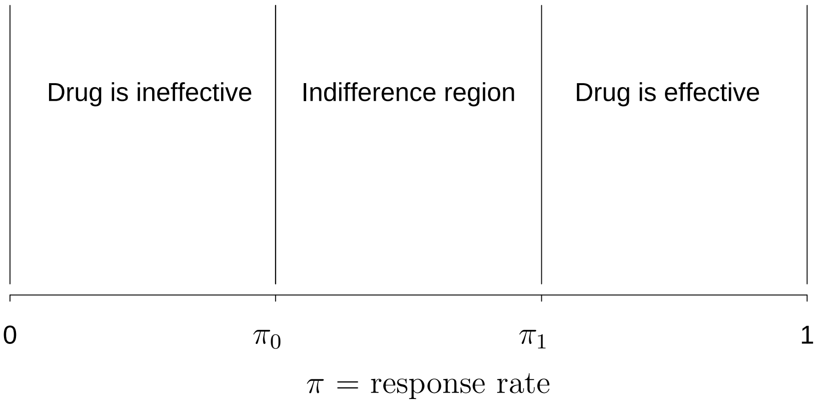
2.2.3 Simon's Two-Stage Design

Another way of using two-stage designs was proposed by Richard Simon. Here, the investigators must decide on values π_0, and π_1, with $\pi_0 < \pi_1$ for the probability of response so that

- If $\pi \leq \pi_0$, then we want to declare the drug ineffective with high probability, say $1 - \alpha$, where α 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 β is taken to be small.

The values of α and β are generally taken to be between .05 and .20.

The region of the parameter space $\pi_0 < \pi < \pi_1$ is the indifference region.



A two-stage design would proceed 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.

Statistically, this decision rule is the following: Let X_1 denote the number of responses in the first stage (among the n_1 patients) and X_2 the number of responses in the second stage (among the $n - n_1$ patients). X_1 and X_2 are assumed to be independent binomially distributed random variables, $X_1 \sim b(n_1, \pi)$ and $X_2 \sim b(n_2, \pi)$, where $n_2 = n - n_1$ and π denotes the probability of response. Declare the treatment a failure if

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

otherwise, the treatment is declared a success if

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

Note: If $n_1 > r$ and if the number of patients responding in the first stage is greater than r, then there is no need to proceed to the second stage to declare the treatment a success.

According to the constraints of the problem we want

$$P(\text{declaring treatment success}|\pi \leq \pi_0) \leq \alpha,$$

or equivalently

$$\underbrace{P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)|\pi = \pi_0\} \leq \alpha}; \tag{2.1}$$

Note: If the above inequality is true when $\pi = \pi_0$, then it is true when $\pi < \pi_0$.

Also, we want

$$P(\text{declaring treatment failure}|\pi \geq \pi_1) \leq \beta,$$

or equivalently

$$P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)|\pi = \pi_1\} \geq 1 - \beta. \tag{2.2}$$

Question: How are probabilities such as $P\{(X_1 > r_1) \text{ and } (X_1 + X_2 > r)|\pi\}$ computed?

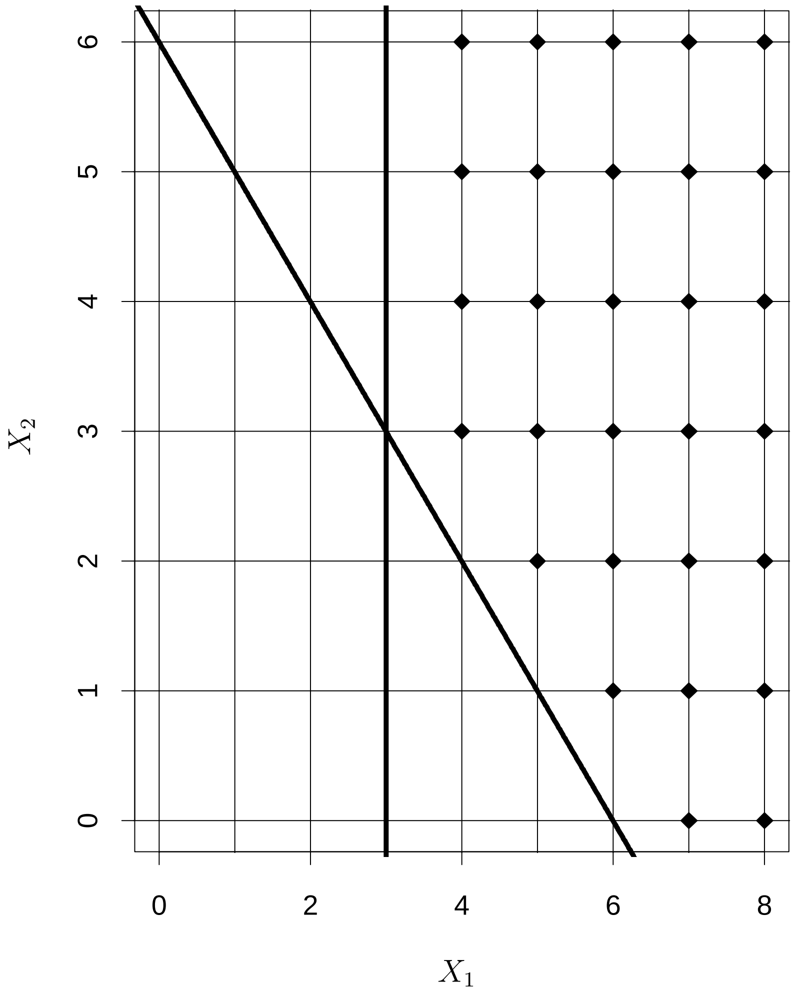
Since X_1 and X_2 are independent binomial random variables, then for any integer $0 \leq m_1 \leq n_1$ and integer $0 \leq m_2 \leq n_2$, the

$$P(X_1 = m_1, X_2 = m_2|\pi) = P(X_1 = m_1|\pi) \times P(X_2 = m_2|\pi)$$
$$= \left\{\begin{pmatrix} n_1 \\ m_1 \end{pmatrix} \pi^{m_1}(1-\pi)^{n_1-m_1}\right\} \left\{\begin{pmatrix} n_2 \\ m_2 \end{pmatrix} \pi^{m_2}(1-\pi)^{n_2-m_2}\right\}.$$

We then have to identify the pairs (m_1, m_2) where $(m_1 > r_1)$ and $(m_1 + m_2) > r$, find the probability for each such (m_1, m_2) pair using the equation above, and then add all the appropriate probabilities.

We illustrate this in the following figure:

Figure 2.2: *Example:* $n_1 = 8$, $n = 14$, $X_1 > 3$, *and* $X_1 + X_2 > 6$



As it turns out there are many combinations of (r_1, n_1, r, n) that satisfy the constraints (2.1) and (2.2) for specified $(\pi_0, \pi_1, \alpha, \beta)$. Through a computer search one can find the "**optimal design**" among these possibilities, where the optimal design is defined as the combination (r_1, n_1, r, n), satisfying the constraints (2.1) and (2.2), which gives the smallest expected sample size when

$\pi = \pi_0$.

The expected sample size for a two stage design is defined as

$$n_1 P(\text{stopping at the first stage}) + nP(\text{stopping at the second stage}).$$

For our problem, the expected sample size is given by

$$n_1\{P(X_1 \leq r_1|\pi = \pi_0) + P(X_1 > r|\pi = \pi_0)\} + nP(r_1 + 1 \leq X_1 \leq r|\pi = \pi_0).$$

Optimal two-stage designs have been tabulated for a variety of $(\pi_0, \pi_1, \alpha, \beta)$ in the article

Simon, R. (1989). Optimal two-stage designs for Phase II clinical trials. *Controlled Clinical Trials*. 10: 1-10.

The tables are given on the next two pages.

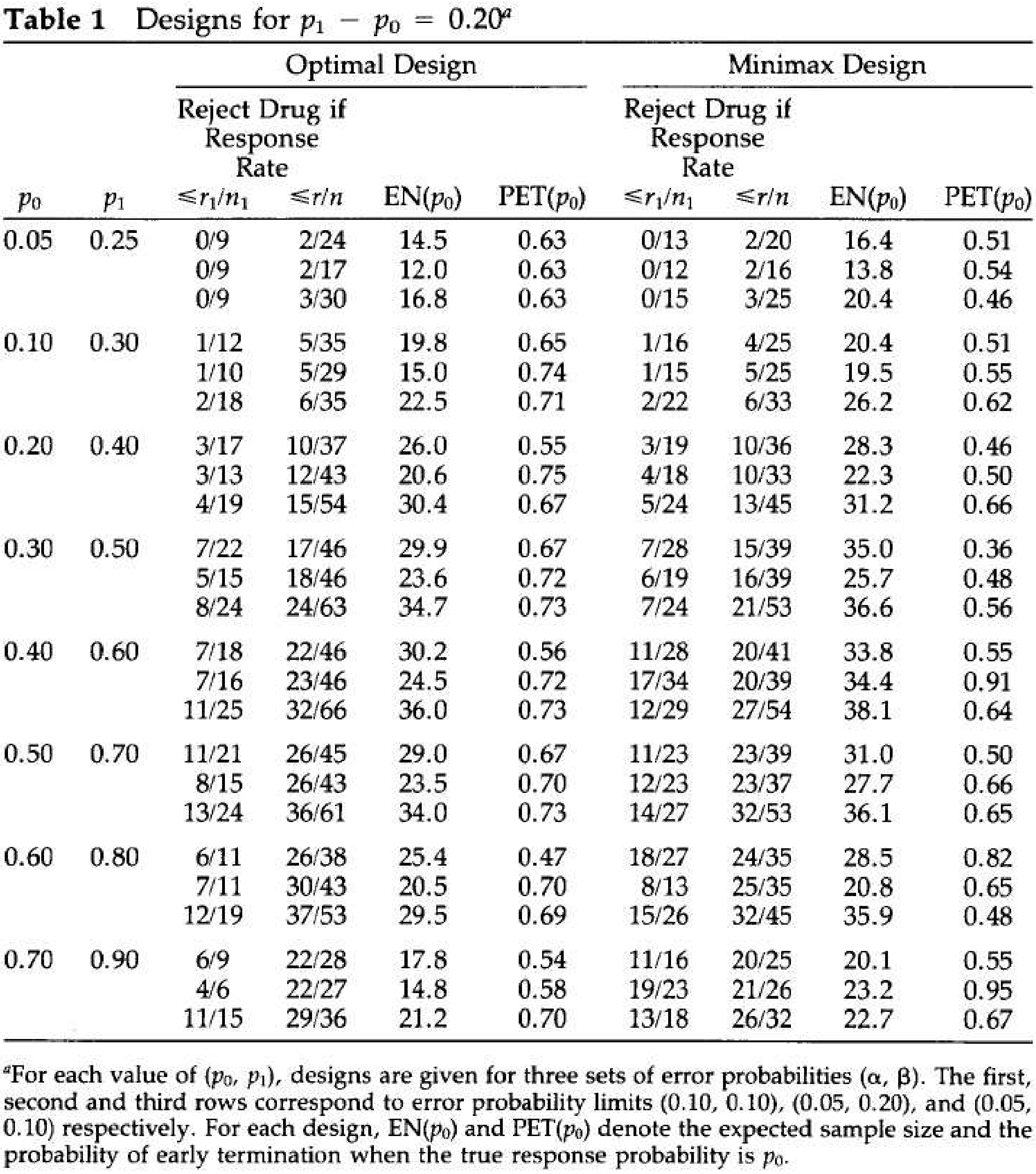


Table 1 Designs for $p_1 - p_0 = 0.20^a$

		Optimal Design				Minimax Design			
		Reject Drug if Response Rate				Reject Drug if Response Rate			
p_0	p_1	$\leq r_1/n_1$	$\leq r/n$	EN(p_0)	PET(p_0)	$\leq r_1/n_1$	$\leq r/n$	EN(p_0)	PET(p_0)
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

[a] For each value of (p_0, p_1), designs are given for three sets of error probabilities (α, β). 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.

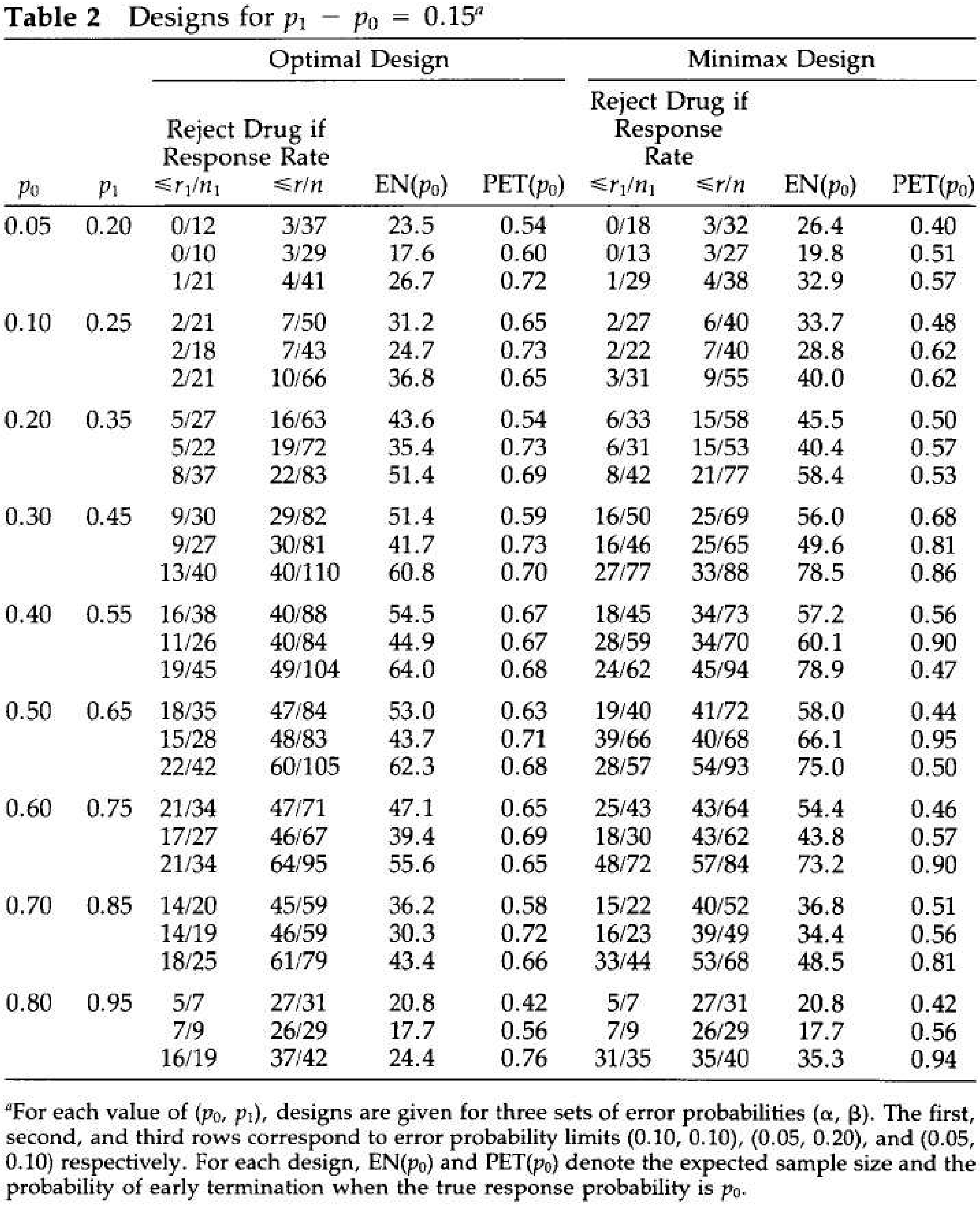


Table 2 Designs for $p_1 - p_0 = 0.15$[a]

		Optimal Design				Minimax Design			
		Reject Drug if Response Rate				Reject Drug if Response Rate			
p_0	p_1	$\leqslant r_1/n_1$	$\leqslant r/n$	EN(p_0)	PET(p_0)	$\leqslant r_1/n_1$	$\leqslant r/n$	EN(p_0)	PET(p_0)
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

[a] For each value of (p_0, p_1), designs are given for three sets of error probabilities (α, β). 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.

3 Phase III Clinical Trials

3.1 Why are clinical trials needed

A clinical trial is the clearest method of determining whether an intervention has the postulated effect. It is very easy for anecdotal information about the benefit of a therapy to be accepted and become standard of care. The consequence of not conducting appropriate clinical trials can be serious and costly. As we discussed earlier, because of anecdotal information, blood-letting was common practice for a very long time. Other examples include

- It was believed that high concentrations of oxygen was useful for therapy in premature infants until a clinical trial demonstrated its harm
- Intermittent positive pressure breathing became an established therapy for chronic obstructive pulmonary disease (COPD). Much later, a clinical trial suggested no major benefit for this very expensive procedure
- Laetrile (a drug extracted from grapefruit seeds) was rumored to be the wonder drug for Cancer patients even though there was no scientific evidence that this drug had any biological activity. People were so convinced that there was a conspiracy by the medical profession to withhold this drug that they would get it illegally from “quacks” or go to other countries such as Mexico to get treatment. The use of this drug became so prevalent that the National Institutes of Health finally conducted a clinical trial where they proved once and for all that Laetrile had no effect. You no longer hear about this issue any more.
- The Cardiac Antiarhythmia Suppression Trial (CAST) documented that commonly used antiarhythmia drugs were harmful in patients with myocardial infarction
- More recently, against common belief, it was shown that prolonged use of Hormone Replacement Therapy for women following menopause may have deleterious effects.

3.2 Issues to consider before designing a clinical trial

David Sackett gives the following six prerequisites

1. The trial needs to be done

   (i) the disease must have either high incidence and/or serious course and poor prognosis

   (ii) existing treatment must be unavailable or somehow lacking

   (iii) 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 generalizibility. 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 generalizibility

5. The trial **protocol** is feasible

   (i) The protocol must be attractive to potential investigators

   (ii) 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: Suppose we are considering a study to compare various treatments for patients with HIV disease, then what might be the appropriate primary endpoint for such a study? Let us look at some options and discuss them.

The HIV virus destroys the immune system; thus individuals infected are susceptible to various opportunistic infections which ultimately leads to death. Many of the current treatments are designed to target the virus either trying to destroy it or, at least, slow down its replication. Other treatments may target specific opportunistic infections.

Suppose we have a treatment intended to attack the virus directly, Here are some possibilities for the primary endpoint that we may consider.

1. Increase in CD4 count. Since CD4 count is a direct measure of the immune function and CD4 cells are destroyed by the virus, we might expect that a good treatment will increase 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 and 2 may be appropriate as the primary outcome in a phase II trial where we want to measure the activity of the treatment as quickly as possible.

Outcome 4 may be of ultimate interest in a phase III trial, but may not be practical for studies where patients have a long expected survival and new treatments are being introduced all the time. (Obsolescence)

Outcome 5 may be the most appropriate endpoint in a phase III trial. However, the other outcomes may be reasonable for secondary analyses.

3.3 Ethical Issues

A clinical trial involves human subjects. As such, we must be aware of ethical issues in the design and conduct of such experiments. Some ethical issues that need to be considered include the following:

- 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 patients 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

3.4 The Randomized Clinical Trial

The objective of a clinical trial is to evaluate the effects of an intervention. Evaluation implies that there must be some comparison either to

- no intervention
- placebo
- best therapy available

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

The difficulty in non-randomized clinical trials is that the control group may be different prognostically from the intervention group. Therefore, comparisons between the intervention and control groups may be biased. That is, differences between the two groups may be due to factors other than the treatment.

Attempts to correct the bias that may be induced by these confounding factors either by design (matching) or by analysis (adjustment through stratified analysis or regression analysis) may not be satisfactory.

To illustrate the difficulty with non-randomized controls, we present results from 12 different studies, all using the same treatment of 5-FU on patients with advanced carcinoma of the large bowel.

Table 3.1: *Results of Rapid Injection of 5-FU for Treatment of Advanced Carcinoma of the Large Bowel*

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

Suppose there is a new treatment for advanced carcinoma of the large bowel that we want to compare to 5-FU. We decide to conduct a new study where we treat patients only with the new drug and compare the response rate to the historical controls. At first glance, it looks as if the response rates in the above table vary tremendously from study to study even though all these used the same treatment 5-FU. If this is indeed the case, then what comparison can possibly be made if we want to evaluate the new treatment against 5-FU? It may be possible, however, that the response rates from study to study are consistent with each other and the differences we are seeing come from random sampling fluctuations. This is important because if we believe there is no study to study variation, then we may feel confident in conducting a new study using only the new treatment and comparing the response rate to the pooled response rate from the studies above. How can we assess whether these differences are random sampling fluctuations or real study to study differences?

Hierarchical Models

To address the question of whether the results from the different studies are random samples from underlying groups with a common response rate or from groups with different underlying response rates, we introduce the notion of a hierarchical model. In a hierarchical model, we assume that each of the N studies that were conducted were from possibly N different study groups each of which have possibly different underlying response rates $\pi_1, \ldots, \pi_N$. In a sense, we now think of the world as being made of many different study groups (or a population of study groups), each with its own response rate, and that the studies that were conducted correspond to choosing a small sample of these population study groups. As such, we imagine $\pi_1, \ldots, \pi_N$ to be a random sample of study-specific response rates from a larger population of study groups. Since π_i, the response rate from the i-th study group, is a random variable, it has a mean and and a variance which we will denote by μ_π and σ_π^2. Since we are imagining a super-population of study groups, each with its own response rate, that we are sampling from, we conceptualize μ_π and σ_π^2 to be the average and variance of these response rates from this super-population. Thus $\pi_1, \ldots, \pi_N$ will correspond to an iid (independent and identically distributed) sample from a population with mean μ_π and variance σ_π^2. I.e.

$$\pi_1, \ldots, \pi_N, \text{ are iid with } E(\pi_i) = \mu_\pi, var(\pi_i) = \sigma_\pi^2, i = 1, \ldots, N.$$

This is the first level of the hierarchy.

The second level of the hierarchy corresponds now to envisioning that the data collected from the i-th study (n_i, X_i), where n_i is the number of patients treated in the i-th study and X_i is the number of complete responses among the n_i treated, is itself a random sample from the i-th study group whose response rate is π_i. That is, conditional on n_i and π_i, X_i is assumed to follow a binomial distribution, which we denote as

$$X_i | n_i, \pi_i \sim b(n_i, \pi_i).$$

This hierarchical model now allows us to distinguish between random sampling fluctuation and real study to study differences. If all the different study groups were homogeneous, then there should be no study to study variation, in which case $\sigma^2_\pi = 0$. Thus we can evaluate the degree of study to study differences by estimating the parameter σ^2_π.

In order to obtain estimates for σ^2_π, we shall use some classical results of conditional expectation and conditional variance. Namely, if X and Y denote random variables for some probability experiment then the following is true

$$E(X) = E\{E(X|Y)\}$$

and

$$var(X) = E\{var(X|Y)\} + var\{E(X|Y)\}.$$

Although these results are known to many of you in the class; for completeness, I will sketch out the arguments why the two equalities above are true.

3.5 Review of Conditional Expectation and Conditional Variance

For simplicity, I will limit myself to probability experiments with a finite number of outcomes. For random variables that are continuous one needs more complicated measure theory for a rigorous treatment.

Probability Experiment

Denote the result of an experiment by one of the outcomes in the sample space $\Omega = \{\omega_1, \ldots, \omega_k\}$. For example, if the experiment is to choose one person at random from a population of size N with a particular disease, then the result of the experiment is $\Omega = \{A_1, \ldots, A_N\}$ where the different A's uniquely identify the individuals in the population, If the experiment were to sample n individuals from the population then the outcomes would be all possible n-tuple combinations of these N individuals; for example $\Omega = \{(A_{i1}, \ldots, A_{in}),$ for all $i1, \ldots, in = 1, \ldots, N$. With replacement there are $k = N^n$th combinations; without replacement there are $k = N \times (N-1) \times \ldots \times (N-n+1)$ combinations of outcomes if order of subjects in the sample is important, and $k = \begin{pmatrix} N \\ n \end{pmatrix}$ combinations of outcomes if order is not important.

Denote by $p(\omega)$ the probability of outcome ω occurring, where $\sum_{\omega \in \Omega} p(\omega) = 1$.

Random variable

A random variable, usually denoted by a capital Roman letter such as $X, Y, \ldots$ is a function that assigns a number to each outcome in the sample space. For example, in the experiment where we sample one individual from the population

$X(\omega)$= survival time for person ω
$Y(\omega)$= blood pressure for person ω
$Z(\omega)$= height of person ω

The **probability distribution** of a random variable X is just a list of all different possible values that X can take together with the corresponding probabilities.
i.e. $\{(x, P(X = x)), \text{ for all possible } x\}$, where $P(X = x) = \sum_{\omega : X(\omega)=x} p(\omega)$.

The **mean** or **expectation** of X is

$$E(X) = \sum_{\omega \in \Omega} X(\omega)p(\omega) = \sum_x xP(X = x),$$

and the **variance** of X is

$$\begin{aligned} var(X) &= \sum_{\omega \in \Omega} \{X(\omega) - E(X)\}^2 p(\omega) = \sum_x \{x - E(X)\}^2 P(X = x) \\ &= E\{X - E(X)\}^2 = E(X^2) - \{E(X)\}^2. \end{aligned}$$

Conditional Expectation

Suppose we have two random variables X and Y defined for the same probability experiment, then we denote the conditional expectation of X, conditional on knowing that $Y = y$, by $E(X|Y = y)$ and this is computed as

$$E(X|Y = y) = \sum_{\omega : Y(\omega)=y} X(\omega) \frac{p(\omega)}{P(Y = y)}.$$

The conditional expectation of X given Y, denoted by $E(X|Y)$ is itself a random variable which assigns the value $E(X|Y = y)$ to every outcome ω for which $Y(\omega) = y$. Specifically, we note that $E(X|Y)$ is a function of Y.

Since $E(X|Y)$ is itself a random variable, it also has an expectation given by $E\{E(X|Y)\}$. By the definition of expectation this equals

$$E\{E(X|Y)\} = \sum_{\omega \in \Omega} E(X|Y)(\omega)p(\omega).$$

By rearranging this sum, first within the partition $\{\omega : Y(\omega) = y\}$, and then across the partitions for different values of y, we get

$$\begin{aligned} E\{E(X|Y)\} &= \sum_y \left\{ \frac{\sum_{\omega:Y(\omega)=y} X(\omega)p(\omega)}{P(Y=y)} \right\} P(Y=y) \\ &= \sum_{\omega \in \Omega} X(\omega)p(\omega) = E(X). \end{aligned}$$

Thus we have proved the very important result that

$$E\{E(X|Y)\} = E(X).$$

Conditional Variance

There is also a very important relationship involving conditional variance. Just like conditional expectation. the conditional variance of X given Y, denoted as $var(X|Y)$, is a random variable, which assigns the value $var(X|Y=y)$ to each outcome ω, where $Y(\omega) = y$, and

$$var(X|Y=y) = E[\{X - E(X|Y=y)\}^2|Y=y] = \sum_{\omega:Y(\omega)=y} \{X(\omega) - E(X|Y=y)\}^2 \frac{p(\omega)}{p(Y=y)}.$$

Equivalently,

$$var(X|Y=y) = E(X^2|Y=y) - \{E(X|Y=y)\}^2.$$

It turns out that the variance of a random variable X equals

$$var(X) = E\{var(X|Y)\} + var\{E(X|Y)\}.$$

This follows because

$$E\{var(X|Y)\} = E[E(X^2|Y) - \{E(X|Y)\}^2] = E(X^2) - E[\{E(X|Y)\}^2] \tag{3.1}$$

and

$$var\{E(X|Y)\} = E[\{E(X|Y)\}^2] - [E\{E(X|Y)\}]^2 = E[\{E(X|Y)\}^2] - \{E(X)\}^2 \tag{3.2}$$

Adding (3.1) and (3.2) together yields

$$E\{var(X|Y)\} + var\{E(X|Y)\} = E(X^2) - \{E(X)\}^2 = var(X),$$

as desired.

If we think of partitioning the sample space into regions $\{\omega : Y(\omega) = y\}$ for different values of y, then the formula above can be interpreted in words as

"the variance of X is equal to the mean of the within partition variances of X plus the variance of the within partition means of X". This kind of partitioning of variances is often carried out in ANOVA models.

Return to Hierarchical Models

Recall

$$X_i|n_i, \pi_i \sim b(n_i, \pi_i),\ \ i = 1, \ldots, N$$

and

$$\pi_1, \ldots, \pi_N \text{ are iid } (\mu_\pi, \sigma_\pi^2).$$

Let $p_i = X_i/n_i$ denote the sample proportion that respond from the i-th study. We know from properties of a binomial distribution that

$$E(p_i|\pi_i, n_i) = \pi_i$$

and

$$var(p_i|\pi_i, n_i) = \frac{\pi_i(1-\pi_i)}{n_i}.$$

Note: In our conceptualization of this problem the probability experiment consists of

1. Conducting N studies from a population of studies

2. For each study i we sample n_i individuals at random from the i-th study group and count the number of responses X_i

3. Let us also assume that the sample sizes $n_1, \ldots, n_N$ are random variables from some distribution.

4. The results of this experiment can be summarized by the iid random vectors

$$(\pi_i, n_i, X_i),\ i = 1, \ldots, N.$$

In actuality, we don't get to see the values $\pi_i, i = 1, \ldots, N$. They are implicitly defined, yet very important in the description of the model. Often, the values π_i are referred to as random effects. Thus, the observable data we get to work with are

$$(n_i, X_i),\ i = 1, \ldots, N.$$

Using the laws of iterated conditional expectation and variance just derived, we get the following results:

$$E(p_i) = E\{E(p_i|n_i, \pi_i)\} = E(\pi_i) = \mu_\pi, \tag{3.3}$$

$$\begin{aligned} var(p_i) &= E\{var(p_i|n_i, \pi_i)\} + var\{E(p_i|n_i, \pi_i)\} \\ &= E\left\{\frac{\pi_i(1-\pi_i)}{n_i}\right\} + var(\pi_i) \\ &= E\left\{\frac{\pi_i(1-\pi_i)}{n_i}\right\} + \sigma_\pi^2. \end{aligned} \tag{3.4}$$

Since the random variables $p_i,\ i = 1, \ldots, N$ are iid, an unbiased estimator for $E(p_i) = \mu_\pi$ is given by the sample mean

$$\bar{p} = N^{-1} \sum_{i=1}^{N} p_i,$$

and an unbiased estimator of the variance $var(p_i)$ is the sample variance

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

One can also show, using properties of a binomial distribution, that a conditionally unbiased estimator for $\frac{\pi_i(1-\pi_i)}{n_i}$, conditional on n_i and π_i, is given by $\frac{p_i(1-p_i)}{n_i-1}$; that is

$$E\left\{\frac{p_i(1-p_i)}{n_i-1}|n_i, \pi_i\right\} = \frac{\pi_i(1-\pi_i)}{n_i}.$$

I will leave this as a homework exercise for you to prove.

Since $\frac{p_i(1-p_i)}{n_i-1},\ i = 1, N$ are iid random variables with mean

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

$$= E\left\{\frac{\pi_i(1-\pi_i)}{n_i}\right\},$$

this means that we can obtain an unbiased estimator for $E\left\{\frac{\pi_i(1-\pi_i)}{n_i}\right\}$ by using

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

Summarizing these results, we have shown that

- $s_p^2 = \frac{\sum_{i=1}^{N}(p_i-\bar{p})^2}{N-1}$ is an unbiased estimator for $var(p_i)$ which by (3.4) equals
$$E\left\{\frac{\pi_i(1-\pi_i)}{n_i}\right\} + \sigma_\pi^2$$
- We have also shown that $N^{-1}\sum_{i=1}^{N}\frac{p_i(1-p_i)}{n_i-1}$ is an unbiased estimator for
$$E\left\{\frac{\pi_i(1-\pi_i)}{n_i}\right\}$$

Consequently, by subtraction, we get that the estimator

$$\hat{\sigma}_\pi^2 = \left\{\frac{\sum_{i=1}^{N}(p_i-\bar{p})^2}{N-1}\right\} - \left\{N^{-1}\sum_{i=1}^{N}\frac{p_i(1-p_i)}{n_i-1}\right\}$$

is an unbiased estimator for σ_π^2.

Going back to the example given in Table 3.1, we obtain the following:

- $$\frac{\sum_{i=1}^{N}(p_i-\bar{p})^2}{N-1} = .0496$$
- $$N^{-1}\sum_{i=1}^{N}\frac{p_i(1-p_i)}{n_i-1} = .0061$$
- Hence
$$\hat{\sigma}_\pi^2 = .0496 - .0061 = .0435$$

Thus the estimate for study to study standard deviation in the probability of response is given by

$$\hat{\sigma}_\pi = \sqrt{.0435} = .21.$$

This is an enormous variation clearly indicating substantial study to study variation.

4 Randomization

In a randomized clinical trial, the allocation of treatment to patients is carried out using a chance mechanism so that neither the patient nor the physician knows in advance which treatment will be assigned. Each patient in the clinical trial has the same opportunity of receiving any of the treatments under study.

Advantages 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
- Provides a basis for standard methods of statistical analysis such as significance tests

4.1 Design-based Inference

On this last point, randomization allows us to carry out design-based inference rather than model-based inference. That is, the distribution of test statistics are induced by the randomization itself rather than assumptions about a super-population and a probability model. Let me illustrate through a simple example. Suppose we start by wanting to test the **sharp** null hypothesis. Under the sharp null hypothesis, it will be assumed that all the treatments being compared would yield exactly the same response on all the patients in a study. To test this hypothesis, patients are randomly allocated to the different treatments and their responses are observed.

To illustrate, suppose we are comparing two treatments and assign 2 of 4 total patients at random to treatment A and the other 2 to treatment B. Our interest is to see whether one treatment affects response more than the other. Let's assume the response is some continuous measurement which we denote by Y. For example, Y might be the difference in blood pressure one week after starting treatment.

We will evaluate the two treatments by computing a test statistic corresponding to the difference in the average response in patients receiving treatment A and the average response in patients receiving treatment B. If the sharp null hypothesis were true, then we would expect, on average, the test statistic to be approximately zero; that is the average response for patients receiving treatment A should be approximately the same as the average response for patients receiving treatment B. Thus, a value of the test statistic sufficiently far from zero could be used as evidence against the null hypothesis. P-values will be used to measure the strength of the evidence. The p-value is the probability that our test statistic could have taken on a value "more extreme" than that actually observed, if the experiment were repeated, under the assumption that the null hypothesis is true. If the p-value is sufficiently small, say $< .05$ or $< .025$, then we use this as evidence to reject the null hypothesis.

Main message: In a randomized experiment, the probability distribution of the test statistic under the null hypothesis is induced by the randomization itself and therefore there is no need to specify a statistical model about a hypothetical super-population.

We will illustrate this point by our simple example. Let us denote the responses for the two patients receiving treatment A as y_1 and y_2 and the responses for the two patients receiving treatment B as y_3 and y_4. Thus the value of the test statistic based on this data is given by

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

How do we decide whether this statistic gives us enough evidence to reject the sharp null hypothesis? More specifically, how do we compute the p-value?

Answer: Under the sharp null hypothesis, the permutational probability distribution of our test static, induced by the randomization, can be evaluated, see Table 4.1.

Table 4.1: *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	$\left(\frac{y_3+y_4}{2}\right)-\left(\frac{y_1+y_2}{2}\right)=t_6$

Under the sharp null hypothesis, the test statistic T (i.e. difference in the two means) can take on any of the six values $t_1, \ldots, t_6$, corresponding to the $\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$ combinations, each with probability 1/6. The value of the test statistic actually observed; in this case t_1, can be declared sufficiently large by gauging it according to the probability distribution induced above. That is, we can compute $P(T \geq t_1|\text{sharp null hypothesis})$, in this case by just counting the number of t_j for $j = 1, \ldots, 6$ that are greater than or equal to t_1 and dividing by six.

Clearly, no one would launch into a comparative trial with only four patients. We used this example for ease of illustration to enumerate the permutational distribution. Nonetheless, for a larger experiment such an enumeration is possible and the permutational distribution can be computed exactly or approximated well by computer simulation.

Note: In the above example, we were implicitly testing the null hypothesis against the one-sided alternative that treatment A was better than treatment B. In this case, larger values of T give more evidence against the null hypothesis. If we, instead, were interested in testing the null hypothesis against a two-sided alternative; that is, that one treatment is different than the other, then large values of the absolute value of T ($|T|$) would be more appropriate as evidence against the null hypothesis. For a two-sided alternative the p-value would be computed as $P(|T| \geq t_1|\text{sharp null hypothesis})$. Because of the symmetry in the permutational distribution of T about zero, this means that the p-value for a two-sided test would be double the p-value for the one-sided test (provided the p-value for the one-sided test was less than .5).

Remark: In evaluating the probability distribution above, we conditioned on the individuals chosen in the experiment. That is, we took their responses as fixed quantities. Randomness was induced by the chance assignment of treatments to individuals which in turn was used to derive the probability distribution of the test statistic.

Contrast this with the usual statistical model which may be used in such an experiment:

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

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

and we are testing the null hypothesis

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

The null hypothesis above would be tested using the t-test, where H_0 is rejected when the test statistic

$$T = \frac{\bar{Y}_A - \bar{Y}_B}{s_p(n_A^{-1} + n_B^{-1})^{1/2}} \overset{H_0}{\sim} t_{n_A+n_B-2},$$

is sufficiently large or the p-value computed using a t-distribution with $n_A + n_B - 2$ degrees of freedom.

Personal Comment: The use of the permutational distribution for inference about treatment efficacy is limiting. Ultimately, we are interested in extending our results from an experimental sample to some larger population. Therefore, in my opinion, the importance of randomization is not the ability to validly use model free statistical tests as we have just seen, but rather, it allows us to make **causal** inference. That is, the results of a randomized clinical trial can be used to infer causation of the intervention on the disease outcome.

This is in contrast to non-randomized clinical trials or epidemiological experiments where only associational inference can be made.

There will be discussion of these points later in the semester.

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 is 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?

4.2 Fixed Allocation Randomization

This scheme, which is the most widely used, assigns interventions to the participants with a prespecified probability which is not altered as the study progresses.

Suppose we are considering two treatments. We want to assign patients to one treatment with probability π and to the other treatment with probability $1 - \pi$. Often π is chosen to be .5.

In some cases, studies have been conducted with unequal allocation probabilities. Let us examine the consequences of the choice of randomization probability from a statistical perspective.

Suppose there are n individuals available for study and we allocate $n\pi \approx n_1$ patients to treatment "1" and $n(1-\pi) \approx n_2$ patients to treatment "2", with $n_1 + n_2 = n$. Say, for example, that the goal of the clinical trial is to estimate the difference in mean response between two treatments; i.e. we want to estimate

$$\mu_2 - \mu_1,$$

where μ_1 and μ_2 denote the population mean response for treatments 1 and 2 respectively.

Remark: As always, some hypothetical population is being envisioned as the population we are interested in making inference on. If every individual in this population were given treatment 1, then the mean response (unknown to us) is denoted by μ_1; whereas, if every individual in this hypothetical population were given treatment 2, then the mean response (again unknown to us) is denoted by μ_2. The n_1 and n_2 patients that are to be randomized to treatments 1 and 2 are thought of as independent random samples chosen from this hypothetical super-population.

An estimator for the treatment difference is given by

$$\bar{Y}_2 - \bar{Y}_1,$$

where $\bar{Y}_1$ is the sample average response of the n_1 patients assigned treatment 1 and $\bar{Y}_2$ is the sample average response of the n_2 patients assigned treatment 2. Let us also assume that the population variance of response is the same for the two treatments; i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$. This may be reasonable if we don't have any a-priori reason to think these variances are different. The variance of our estimator is given as

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

Question: Subject to the constraint that $n = n_1 + n_2$, how do we find the most efficient estimator for $\mu_2 - \mu_1$? That is, what treatment allocation minimizes the variance of our estimator? i.e.

$$\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = \sigma^2 \left\{ \frac{1}{n\pi} + \frac{1}{n(1-\pi)} \right\} = \frac{\sigma^2}{n} \left\{ \frac{1}{\pi(1-\pi)} \right\}.$$

The answer is the value of $0 \le \pi \le 1$ which maximizes $\pi(1-\pi) = \pi - \pi^2$. Using simple calculus, we take the derivative with respect to π and set it equal to zero to find the maximum. Namely,

$$\frac{d(\pi - \pi^2)}{d\pi} = 1 - 2\pi = 0; \ \pi = .5.$$

This is guaranteed to be a maximum since the second derivative

$$\frac{d^2(\pi - \pi^2)}{d\pi^2} = -2.$$

Thus, to get the most efficient answer we should randomize with equal probability. However, as we will now demonstrate, the loss of efficiency is not that great when we use unequal allocation.

For example, if we randomize with probability 2/3, then the variance of our estimator would be

$$\frac{\sigma^2}{n} \left\{ \frac{1}{(2/3)(1/3)} \right\} = \frac{4.5\sigma^2}{n}$$

instead of

$$\frac{\sigma^2}{n} \left\{ \frac{1}{(1/2)(1/2)} \right\} = \frac{4\sigma^2}{n},$$

with equal allocation.

Another way of looking at this relationship is to compute the ratio between the sample sizes of the equal allocation design and the unequal allocation design that yield the same accuracy. For example, if we randomize with probability 2/3, then to get the same accuracy as the equal allocation design, we would need

$$\frac{4.5\sigma^2}{n_{(\pi=2/3)}} = \frac{4\sigma^2}{n_{(\pi=1/2)}},$$

where $n_{(\pi=2/3)}$ corresponds to the sample size for the design with unequal allocation, $\pi = 2/3$ in this case, and $n_{(\pi=1/2)}$, the sample size for the equal allocation design. It is clear that to get the same accuracy we need

$$\frac{n_{(\pi=2/3)}}{n_{(\pi=1/2)}} = \frac{4.5}{4} = 1.125.$$

That is, equal allocation is 12.5% more efficient than a 2:1 allocation scheme; i.e. we need to treat 12.5% more patients with an unequal allocation (2:1) design to get the same statistical precision as with an equal allocation (1:1) design.

Some investigators have advocated putting more patients on the new treatment. Some of the reason given include:

- 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

Some disadvantages are:

- might be difficult to justify ethically; It removes equipoise for the participating clinician
- new treatment may be detrimental

4.2.1 Simple Randomization

For simplicity, let us start by assuming that patients will be assigned to one of two treatments A or B. The methods we will describe will generalize easily to more than two treatments. In a simple randomized trial, each participant that enters the trial is assigned treatment A or B with probability π or $1 - \pi$ respectively, independent of everyone else. Thus, if n patients are randomized with this scheme, the number of patients assigned treatment A is a random quantity following a binomial distribution $\sim b(n, \pi)$.

This scheme is equivalent to flipping a coin (where the probability of a head is π) to determine treatment assignment. Of course, the randomization is implemented with the aid of a computer which generates random numbers uniformly from 0 to 1. Specifically, using the computer, a sequence of random numbers are generated which are uniformly distributed between 0 and 1 and independent of each other. Let us denote these by $U_1, \ldots, U_n$ where U_i are iid $U[0,1]$. For the i-th individual entering the study we would assign treatment as follows:

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

It is easy to see that $P(U_i \leq \pi) = \pi$, which is the desired randomization probability for treatment A. As we argued earlier, most often π is chosen as .5.

Advantages of simple randomization

- easy to implement
- virtually impossible for the investigators to guess what the next treatment assignment will be. If the investigator could break the code, then he/she may be tempted to put certain patients on preferred treatments thus invalidating the unbiasedness induced by the randomization
- the properties of many statistical inferential procedures (tests and estimators) are established under the simple randomization assumption (iid)

Disadvantages

- The major disadvantage is that the number of patients assigned to the different treatments are random. Therefore, the possibility exists of severe treatment imbalance.
    - leads to less efficiency
    - appears awkward and may lead to loss of credibility in the results of the trial

For example, with $n = 20$, an imbalance of 12:8 or worse can occur by chance with 50% probability even though $\pi = .5$. The problem is not as severe with larger samples. For instance if $n = 100$, then a 60:40 split or worse will occur by chance with 5% probability.

4.2.2 Permuted block randomization

One way to address the problem of imbalance is to use **blocked randomization** or, more precisely, **permuted block randomization**.

Before continuing, we must keep in mind that patients enter sequentially over time as they become available and eligible for a study. This is referred to as staggered entry. Also, we must realize that even with the best intentions to recruit a certain fixed number of patients, the actual number that end up in a clinical trial may deviate from the intended sample size. With these constraints in mind, the permuted block design is often used to achieve balance in treatment assignment. In such a design, as patients enter the study, we define a block consisting of a certain number and proportion of treatment assignments. Within each block, the order of treatments is randomly permuted.

For illustration, suppose we have two treatments, A and B, and we choose a block size of 4, with two A's and two B's. For each block of 4 there are $\begin{pmatrix} 4 \\ 2 \end{pmatrix}$ or 6 possible combinations of treatment assignments.

These are

A A B B

A B A B

A B B A

B A A B

B A B A

B B A A

The randomization scheme, to be described shortly, chooses one of these six possibilities with equal probability (i.e. 1/6) and proceeds as follows. The first four patients entering the trial are assigned in order to treatments A and B according to the permutation chosen. For the next block of 4, another combination of treatment assignments is chosen at random from the six permutations above and the next four patients are assigned treatments A or B in accordance. This process continues until the end of the study.

It is clear that by using this scheme, the difference in the number of patients receiving A versus B can never exceed 2 regardless when the study ends. Also after every multiple of four patients, the number on treatments A and B are identical.

Choosing random permutations

This can be done by choosing a random number to associate to each of the letters "AABB" of a block and then assigning, in order, the letter ranked by the corresponding random number. For example

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

In the example above the treatment assignment from this block is "ABAB"; that is, A followed by

B followed by A followed by B. The method just described will guarantee that each combination is equally likely of being chosen.

Potential problem

If the block size *b* is known in advance, then the clinician may be able to guess what the next treatment will be. Certainly, the last treatment in a block will be known if the previous treatments have already been assigned. He/she may then be able to bias the results by putting patients that are at better or worse prognosis on the known treatment. This problem can be avoided by varying the blocking number at random. For example, the blocking number may be 2,4,6, chosen at random with, say, each with probability 1/3. Varying the block sizes at random will make it difficult (not impossible) to break the treatment code.

4.2.3 Stratified Randomization

The response of individuals often depends on many different characteristics of the patients (other than treatment) which are often called prognostic factors. Examples of prognostic factors are age, gender, race, white blood count, Karnofsky status. etc. Although randomization balances prognostic factors "on average" between the different treatments being compared in a study, imbalances may occur by chance.

If patients with better prognosis end up by luck in one of the treatment arms, then one might question the interpretability of the observed treatment differences. One strategy, to minimize this problem, is to use blocked randomization within strata formed by different combinations of prognostic factors defined a-priori. For example, suppose that age and gender are prognostic factors that we want to control for in a study. We define strata by breaking down our population into categories defined by different combinations of age and gender.

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

In the illustration above, a total of six strata were formed by the 3×2 combinations of categories of these two variables.

In a stratified blocked randomization scheme, patients are randomized using block sizes equal to b ($b/2$ on each treatment for equal allocation) within each stratum. With this scheme there could never be a treatment imbalance greater than $b/2$ within any stratum at any point in the study.

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

Suppose we have two strata, which will be denoted by the indicator variable S, where

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

There are also two treatments denoted by the indicator variable X, where

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

Let Y denote the response variable of interest. For this illustration, we take Y to be a continuous random variable; for example, the drop in log viral RNA after three months of treatment for HIV disease. Consider the following model, where for the i-th individual in our sample, we assume

$$Y_i = \mu + \alpha S_i + \beta X_i + \epsilon_i. \tag{4.1}$$

Here α denotes the magnitude of effect that strata has on the mean response, β denotes the magnitude of effect that treatment has on the mean response, and the $\epsilon_i,\ i = 1, \ldots, n$ denote random errors which are taken to be iid random variables with mean zero and variance σ^2.

Let n individuals be put into a clinical trial to compare treatments A and B and denote by n_A, the number of individuals assigned to treatment A; i.e. $n_A = \sum_{i=1}^n X_i$ and n_B the number assigned to treatment B, $n_B = n - n_A$.

Let $\bar{Y}_A$ be the average response for the sample of individuals assigned to treatment A and $\bar{Y}_B$ the similar quantity for treatment B:

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

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

The objective of the study is to estimate treatment effect given, in this case, by the parameter β. We propose to use the obvious estimator $\bar{Y}_A - \bar{Y}_B$ to estimate β.

Some of the patients from both treatments will fall into strata 1 and the others will fall into strata 0. We represent this in the following table.

Table 4.2: *Number of observations falling into the different strata by treatment*

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

Because of (4.1), we get

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

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

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

$$\bar Y_A-\bar Y_B=\beta+\alpha\left(\frac{n_{A1}}{n_A}-\frac{n_{B1}}{n_B}\right)+(\bar\epsilon_A-\bar\epsilon_B). \tag{4.2}$$

Stratified randomization

Let us first consider the statistical properties of the estimator for treatment difference if we used permuted block randomization within strata with equal allocation. Roughly speaking, the number assigned to the two treatments, by strata, would be

$$n_A=n_B=n/2$$

$$n_{A1}=n_{B1}=n_1/2$$

$$n_{A0}=n_{B0}=n_0/2.$$

Remark: The counts above might be off by $b/2$, where b denotes the block size, but when n is large this difference is inconsequential.

Substituting these counts into formula (4.2), we get

$$\bar Y_A-\bar Y_B=\beta+(\bar\epsilon_A-\bar\epsilon_B).$$

Note: The coefficient for α cancelled out.

Thus the mean of our estimator is given by

$$E(\bar Y_A-\bar Y_B)=E\{\beta+(\bar\epsilon_A-\bar\epsilon_B)\}=\beta+E(\bar\epsilon_A)-E(\bar\epsilon_B)=\beta,$$

which implies that the estimator is unbiased. The variance of the estimator is given by

$$var(\bar{Y}_A - \bar{Y}_B) = var(\bar{\epsilon}_A) + var(\bar{\epsilon}_B) = \sigma^2 \left(\frac{2}{n} + \frac{2}{n} \right)$$

$$\frac{4\sigma^2}{n}. \tag{4.3}$$

Simple randomization

With simple randomization the counts n_{A1} conditional on n_A and n_{B1} conditional on n_B follow a binomial distribution. Specifically,

$$n_{A1} | n_A, n_B \sim b(n_A, \theta) \tag{4.4}$$

and

$$n_{B1} | n_A, n_B \sim b(n_B, \theta), \tag{4.5}$$

where θ denotes the proportion of the population in stratum 1. In addition, conditional on n_A, n_B, the binomial variables n_{A1} and n_{B1} are independent of each other.

The estimator given by (4.2) has expectation equal to

$$E(\bar{Y}_A - \bar{Y}_B) = \beta + \alpha \left\{ E\left(\frac{n_{A1}}{n_A} \right) - E\left(\frac{n_{B1}}{n_B} \right) \right\} + E(\bar{\epsilon}_A - \bar{\epsilon}_B). \tag{4.6}$$

Because of (4.4)

$$E\left(\frac{n_{A1}}{n_A} \right) = E\left\{ E\left(\frac{n_{A1}}{n_A} | n_A \right) \right\} = E\left(\frac{n_A \theta}{n_A} \right) = \theta.$$

Similarly

$$E\left(\frac{n_{B1}}{n_B} \right) = \theta.$$

Hence,

$$E(\bar{Y}_A - \bar{Y}_B) = \beta;$$

that is, with simple randomization, the estimator $\bar{Y}_A - \bar{Y}_B$ is an unbiased estimator of the treatment difference β.

In computing the variance, we use the formula for iterated conditional variance; namely

$$var(\bar{Y}_A - \bar{Y}_B) = E\{var(\bar{Y}_A - \bar{Y}_B | n_A, n_B)\} + var\{E(\bar{Y}_A - \bar{Y}_B | n_A, n_B)\}.$$

As demonstrated above, $E(\bar{Y}_A - \bar{Y}_B|n_A, n_B) = \beta$, thus $var\{E(\bar{Y}_A - \bar{Y}_B|n_A, n_B)\} = var(\beta) = 0$. Thus we need to compute $E\{var(\bar{Y}_A - \bar{Y}_B|n_A, n_B)\}$. Note that

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

Therefore

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

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

We have already shown that $\left(\frac{1}{n_A} + \frac{1}{n-n_A}\right) \geq \frac{4}{n}$. Therefore, with simple randomization the variance of the estimator for treatment difference; namely

$$var(\bar{Y}_A - \bar{Y}_B) = \{\sigma^2 + \alpha^2\theta(1-\theta)\}E\left(\frac{1}{n_A} + \frac{1}{n - n_A}\right)$$

is greater than the variance of the estimator for treatment difference with stratified randomization; namely

$$var(\bar{Y}_A - \bar{Y}_B) = \frac{4\sigma^2}{n}.$$

Remark: In order to take advantage of the greater efficiency of the stratified design, one has to recognize that the variance for $(\bar{Y}_A - \bar{Y}_B)$ is different when using a stratified design versus simple randomization. Since many of the statistical tests and software are based on assumptions that the data are iid, this point is sometimes missed.

For example, suppose we used a permuted block design within strata but analyzed using a t-test (the test ordinarily used in conjunction with simple randomization). The t-test is given by

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

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. Some of these issues will be discussed later in the course.

Disadvantage of blocking within strata

One can inadvertently counteract the balancing effects of blocking by having too many strata. As we consider more prognostic factors to use for balance, we find that the number of strata grow exponentially. For example, with 10 prognostic factors, each dichotomized, we would have $2^{10} = 1024$ strata. In the most extreme case, suppose we have so many strata that there is no more than one patient per stratum. The result is then equivalent to having simple randomization and blocking would have no impact at all. Generally, we should use few strata relative to the size of the clinical trial. That is, most blocks should be filled because unfilled blocks permit imbalances.

4.3 Adaptive Randomization Procedures

This is where the rule for allocation to different treatments may vary according to the results from prior patients already in the study. Baseline adaptive procedures attempt to balance the allocation of patients to treatment overall and/or by prognostic factors. Some examples are

4.3.1 Efron biased coin design

Choose an integer D, referred to as the discrepancy, and a probability ϕ less than .5. For example, we can take $D = 3$ and $\phi = .25$. The allocation scheme is as follows. Suppose at any point in time in the study the number of patients on treatment A and treatment B are n_A and n_B respectively. The next patient to enter the study will be randomized to treatment A or B with probability π_A or $1 - \pi_A$, where

$$\begin{aligned}
\pi_A &= .5 \quad \text{if } \ |n_A - n_B| \leq D \\
\pi_A &= \phi \quad \text{if } \ n_A - n_B > D \\
\pi_A &= 1 - \phi \quad \text{if } \ n_B - n_A > D
\end{aligned}$$

The basic idea is that as soon as the treatments become sufficiently imbalanced favoring one treatment, then the randomization is chosen to favor the other treatment in an attempt to balance more quickly while still incorporating randomization so that the physician can never be certain of the next treatment assignment,

A criticism of this method is that design-based inference is difficult to implement. **Personally**, I don't think this issue is of great concern because model-based inference is generally the accepted practice. However, the complexity of implementing this method may not be worthwhile.

4.3.2 Urn Model (L.J. Wei)

In this scheme we start with $2m$ balls in an urn; m labeled A and m labeled B. When the first patient enters the study you choose a ball at random from the urn and assign that treatment. If you chose an A ball, you replace that ball in the urn and add an additional B ball. If you chose a B ball then you replace it and add an additional A ball. Continue in this fashion. Clearly, the reference to an urn is only for conceptualization, such a scheme would be implemented by a computer.

Again, as soon as imbalance occurs in favor of one treatment, the chances become greater to get the other treatment in an attempt to balance more quickly. This scheme makes it even more difficult than the biased coin design for the physician to guess what treatment the next patient will be assigned to. Again, design-based inference is difficult to implement, but as before, this may not be of concern for most clinical trial statisticians.

Both the biased coin design and the urn model can be implemented within strata.

4.3.3 Minimization Method of Pocock and Simon

This is almost a deterministic scheme for allocating treatment with the goal of balancing many prognostic factors (marginally) by treatment. Suppose there are K prognostic factors, indexed by $i = 1, \ldots, K$ and each prognostic factor is broken down into k_i levels, then the total number of strata is equal to $k_1 \times \ldots \times k_K$. This can be a very large number, in which case, permuted block randomization within strata may not be very useful in achieving balance. Suppose, instead, we wanted to achieve some degree of balance for the prognostic factors marginally rather than within each stratum (combination of categories from all prognostic factors). Pocock and Simon suggested using a measure of marginal discrepancy where the next patient would be assigned to whichever treatment that made this measure of marginal discrepancy smallest. Only in the case where the measure of marginal discrepancy was the same for both treatments would the next patient be randomized to one of the treatments with equal probability.

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

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

Similarly,

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

The measure of marginal discrepancy is given by

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

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

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 factor, 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 at this point 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.

4.4 Response Adaptive Randomization

In response adaptive schemes, the responses of the past participants in the study are used to determine the treatment allocation for the next patient. Some examples are

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			

Urn Model (L.J. Wei)

The first patient is assigned to either treatment by equal probability. Then 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

To illustrate the last point, we consider the results of the ECMO trial which used a play-the-winner rule.

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. Because the mortality rate was very high for this population and because of the 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.

4.5 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
  - (a) 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
  - (b) 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.
  - (c) For a multi-center trial, randomization is carried out by the central office by phone or by computer.
  - (d) For a double-blind multi-center trial, the randomization may need to be decentralized to each center according to (b). 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

5.1 Blinding and Placebos

Even in a randomized trial the comparison of treatments may be distorted if the patients and those responsible for administering the treatment and evaluation know which treatment is being used. These problems can be avoided in some cases by making the trial **double blind**, whereby, neither patient, physician nor evaluator are aware which treatment the patient is receiving.

- The patient— If the patient knows he/she is receiving a new treatment then this may result in psychological benefit. The degree of psychological effect depends on the type of disease and the nature of treatments. One should not underestimate the importance of psychology for patients with disease. Whether it is asthma, cancer, heart disease, etc, the manner in which patients are informed of therapy has a profound effect on subsequent performance.
- The treatment team—(anyone who participates in the treatment or management of the patient). Patients known to be receiving a new therapy may be treated differently than those on standard treatment. Such difference in ancillary care may affect the response.
- The evaluator— It is especially important that the individual or individuals evaluating response be objective. A physician who has pre-conceived ideas how a new treatment might work may introduce bias in his/her evaluation of the patient response if they know the treatment that the patient received.

The biases above may be avoided with proper blinding. However, blinding treatments takes a great deal of care and planning. If the treatment is in the form of pills, then the pills for the different treatments should be indistinguishable; i.e the same size, color, taste, texture. If no treatment is to be used as the control group then we may consider using a placebo for patients randomized to the control group. A placebo is a pill or other form of treatment which is indistinguishable from the active treatment but contains no active substance. (sugar pill, saline, etc.) If you are comparing two active treatments each, say, with pills that cannot be made to be similar, then we may have to give each patient two pills; one active pill for one treatment and a placebo pill for the other treatment. (This can become overwhelming if we are comparing different combinations of drugs).

It has been well documented that there is a placebo effect. That is, there have been randomized studies conducted that gave some patients placebo and the other patients nothing with the placebo group responding significantly better. Consequently, in a randomized clinical trial which compares a new drug to a placebo control, we are actually testing whether the active drug has effect equal to or greater than a placebo effect.

One must realize that although the principles of blinding are good, they are not feasible in some trials. For example, if we are comparing surgery versus chemotherapy in a cancer clinical trial, there is no way to blind these treatments. In such cases we must be as careful as possible to choose endpoints that are as objective as possible. For example, time to death from any cause.

5.2 Ethics

Clinical trials are ethical in the setting of uncertainty.

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:

To consider dear to me as my parents him who taught me this art; to live in common with him and if necessary to share my goods with him; to look upon his children as my own brothers, to teach them this art if they so desire without fee or written promise; to impart to my sons and the sons of the master who taught me and the disciples who have enrolled themselves and have agreed to the rules of the profession, but to these alone, the precepts and the instruction. I will prescribe regimen for the good of my patients according to my ability and my judgment and never do harm to anyone. To please no one will I prescribe a deadly drug, nor give advice which may cause his death. Nor will I give a woman a pessary to procure abortion. But I will preserve the purity of my life and my art. I will not cut for stone, even for patients in whom disease is manifest; I will leave this operation to be performed by practitioners (specialists in this art). In every house where I come I will enter only for the good of my patients, keeping myself far from all intentional ill-doing and all seduction, and especially from the pleasures of love with women or men, be they free or slaves. All that may come to my knowledge in the exercise of my profession or outside of my profession or in daily commerce with men, which ought not to be spread abroad, I will keep secret and I will never reveal. If I keep this oath faithfully, may I enjoy life and practice my art, respected by all men and in all times; but if I swerve from it or violate it, may the reverse be my lot.

Even today physicians may take the Hippocratic oath although it is not repeated in full. Clearly many of the issues, such as abortion, surgery for kidney stones, use of deadly drugs, no longer apply. Nor does the pledge for "free instruction" still apply.

Ethical considerations have been addressed by the Nuremberg Code and Helsinki Declaration (see the class website for more details)

In the United States, the Congress established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research as part of the National Research Act. This act required the establishment of an Institutional Review Board (IRB) for all research funded in whole or in part by the federal government. These were later modified to require IRB approval for all drugs or products regulated by the Food and Drug Administration (FDA).

IRB's must have at least five members with expertise relevant to safeguarding the rights and welfare of patients participating in biomedical research. At least one should be a scientist, and at least one must not be affiliated with the institution. The IRB should be made up of individuals with diverse racial, gender and cultural backgrounds. The scope of the IRB includes, but is not limited to consent procedures and research design.

IRB's approve human research studies that meet specific prerequisites.

(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

5.3 The Protocol Document

Definition: The protocol is a scientific planning document for a medical study on human subjects. It contains the study 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

I will hand out two protocols in class. One is from the Womens Health Initiative (WHI) and one from the Cancer and Leukemia Group B (CALGB) study 8541.

In the WHI study, the clinical trial will evaluate the benefits and risks of Hormone Replacement Therapy (HRT), Dietary Modification, DM, and supplementation with calcium/vitamin D (CaD) on the overall health of postmenopausal women. Health will be assessed on the basis of quality of life measurements, cause-specific morbidity and mortality, and total mortality.

CALGB 8541 is a study of different regimen of adjuvant CAF (combination of Cyclophosphamide, Adriamycin and 5 Fluorouracil (5-FU)) as treatment for women with pathological stage II, node positive breast cancer. Specifically, intensive CAF for four cycles versus low dose CAF for four cycles versus standard dose CAF for six cycles will be compared in a three arm randomized clinical trial.

Protocols generally have the following elements:

1. **Schema**. Depicts the essentials of a study design.
   WHI: page 18
   CALGB 8541: page 1

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

   CALGB 8541: page 3

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

   WHI: pages 2-13

   CALGB 8541: pages 1-3

   This generally includes

   - standard therapy
   - predecessor studies (phase I and II if appropriate)
   - previous or concurrent studies of a similar nature
   - 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

   CALGB 8541: pages 4-5

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

   WHI: pages 29-38

   CALGB 8541: pages 5-7

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

   CALGB 8541: pages 7-20

7. **Study parameters** This section gives the schedule of the required and optional investigations/tests.

   WHI: pages 38-39

   CALGB 8541: page 20

8. **Statistical Considerations**

   WHI: pages 52-55 and an extensive appendix (not provided)

   CALGB 8541: pages 22-23

   - 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.

   For both WHI and CALGB 8541 these are in an appendix (not included)

   The informed consent should include

   - an explanation of the procedures to be followed and their purposes
   - a description of the benefits that might reasonably be expected
   - a description of the discomforts and risks that could reasonably be expected
   - a disclosure of any appropriate alternative procedures that might be advantageous
   - 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

    CALGB 8541: Was not included

6 Sample Size Calculations

One of the major responsibilities of a clinical trial statistician is to aid the investigators in determining the sample size required to conduct a study. The most common procedure for determining the sample size involves computing the minimum sample size necessary in order that important treatment differences be determined with sufficient accuracy. We will focus primarily on hypothesis testing.

6.1 Hypothesis Testing

In a hypothesis testing framework, the question is generally posed as a decision problem regarding a parameter in a statistical model:

Suppose a population parameter corresponding to a measure of treatment difference using the primary endpoint is defined for the study. This parameter will be denoted by Δ. For example, if we are considering the mean response of some continuous variable between two treatments, we can denote by μ_1, the population mean response for treatment 1 and μ_2, the mean response on treatment 2. We then denote by

$$\Delta = \mu_1 - \mu_2$$

the measure of treatment difference. A clinical trial will be conducted in order to make inference on this population parameter. If we take a hypothesis testing point of view, then we would consider the following decision problem: Suppose treatment 2 is currently the standard of care and treatment 1 is a new treatment that has shown promise in preliminary testing. What we want to decide is whether we should recommend the new treatment or stay with the standard treatment. As a starting point we might say that if, in truth, $\Delta \leq 0$ then we would want to stay with the standard treatment; whereas, if, in truth, $\Delta > 0$, then we would recommend that future patients go on the new treatment. We refer to $\Delta \leq 0$ as the null hypothesis "H_0" and $\Delta > 0$ as the alternative hypothesis "H_A". The above is an example of a one-sided hypothesis test. In some cases, we may be interested in a two-sided hypothesis test where we test the null hypothesis $H_0 : \Delta = 0$ versus the alternative $H_A : \Delta \neq 0$.

In order to make a decision on whether to choose the null hypothesis or the alternative hypothesis, we conduct a clinical trial and collect data from a sample of individuals. The data from n individuals in our sample will be denoted generically as $(z_1, \ldots, z_n)$ and represent realizations of random vectors $(Z_1, \ldots, Z_n)$. The Z_i may represent a vector of random variables for individual i; e.g. response, treatment assignment, other covariate information.

As statisticians, we posit a probability model that describes the distribution of $(Z_1, \ldots, Z_n)$ in terms of population parameters which includes Δ (treatment difference) as well as other parameters necessary to describe the probability distribution. These other parameters are referred to as nuisance parameters. We will denote the nuisance parameters by the vector θ. As a simple example, let the data for the i-th individual in our sample be denoted by $Z_i = (Y_i, A_i)$, where Y_i denotes the response (taken to be some continuous measurement) and A_i denotes treatment assignment, where A_i can take on the value of 1 or 2 depending on the treatment that patient i was assigned. We assume the following statistical model: let

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

and

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

i.e. since $\Delta = \mu_1 - \mu_2$, then $\mu_1 = \mu_2 + \Delta$. The parameter Δ is the test parameter (treatment difference of primary interest) and $\theta = (\mu_2, \sigma^2)$ are the nuisance parameters.

Suppose we are interested in testing $H_0 : \Delta \leq 0$ versus $H_A : \Delta > 0$. The way we generally proceed is to combine the data into a summary test statistic that is used to discriminate between the null and alternative hypotheses based on the magnitude of its value. We refer to this test statistic by

$$T_n(Z_1, \ldots, Z_n).$$

Note: We write $T_n(Z_1, \ldots, Z_n)$ to emphasize the fact that this statistic is a function of all the data $Z_1, \ldots, Z_n$ and hence is itself a random variable. However, for simplicity, we will most often refer to this test statistic as T_n or possibly even T.

The statistic T_n should be constructed in such a way that

(a) Larger values of T_n are evidence against the null hypothesis in favor of the alternative

(b) The probability distribution of T_n can be evaluated (or at least approximated) at the **border** between the null and alternative hypotheses; i.e. at $\Delta = 0$.

After we conduct the clinical trial and obtain the data, i.e. the realization $(z_1, \ldots, z_n)$ of $(Z_1, \ldots, Z_n)$, we can compute $t_n = T_n(z_1, \ldots, z_n)$ and then gauge this observed value against the distribution of possible values that T_n can take under $\Delta = 0$ to assess the strength of evidence for or against the null hypothesis. This is done by computing the p-value

$$P_{\Delta=0}(T_n \geq t_n).$$

If the p-value is small, say, less than .05 or .025, then we use this as evidence against the null hypothesis.

Note:

1. Most test statistics used in practice have the property that $P_\Delta(T_n \geq x)$ increases as Δ increases, for all x. In particular, this would mean that if the p-value $P_{\Delta=0}(T_n \geq t_n)$ were less than α at $\Delta = 0$, then the probability $P_\Delta(T_n \geq t_n)$ would also be less than α for all Δ corresponding to the null hypothesis $H_0 : \Delta \leq 0$.

2. Also, most test statistics are computed in such a way that the distribution of the test statistic, when $\Delta = 0$, is approximately a standard normal; i.e.

$$T_n \overset{(\Delta=0)}{\sim} N(0, 1),$$

regardless of the values of the nuisance parameters θ.

For the problem where we were comparing the mean response between two treatments, where response was assumed normally distributed with equal variance by treatment, but possibly difference means, we would use the two-sample t-test; namely,

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

where $\bar{Y}_1$ denotes the sample average response among the n_1 individuals assigned to treatment 1, $\bar{Y}_2$ denotes the sample average response among the n_2 individuals assigned to treatment 2,

$n = n_1 + n_2$ and the sample variance is

$$s_Y^2 = \left\{ \frac{\sum_{j=1}^{n_1}(Y_{1j} - \bar{Y}_1)^2 + \sum_{j=1}^{n_2}(Y_{2j} - \bar{Y}_2)^2}{(n_1 + n_2 - 2)} \right\}.$$

Under $\Delta = 0$, the statistic T_n follows a central t distribution with $n_1 + n_2 - 2$ degrees of freedom. If n is large (as it generally is for phase III clinical trials), this distribution is well approximated by the standard normal distribution.

If the decision to reject the null hypothesis is based on the p-value being less that α (.05 or .025 generally), then this is equivalent to rejecting H_0 whenever

$$T_n \geq \mathcal{Z}_\alpha,$$

where $\mathcal{Z}_\alpha$ denotes the $1-\alpha$-th quantile of a standard normal distribution; e.g. $\mathcal{Z}_{.05} = 1.64$ and $\mathcal{Z}_{.025} = 1.96$. We say that such a test has level α.

Remark on two-sided tests: If we are testing the null hypothesis $H_0 : \Delta = 0$ versus the alternative hypothesis $H_A : \Delta \neq 0$ then we would reject H_0 when the absolute value of the test statistic $|T_n|$ is sufficiently large. The p-value for a two-sided test is defined as $P_{\Delta=0}(|T_n| \geq t_n)$, which equals $P_{\Delta=0}(T_n \geq t_n) + P_{\Delta=0}(T_n \leq -t_n)$. If the test statistic T_n is distributed as a standard normal when $\Delta = 0$, then a level α two-sided test would reject the null hypothesis whenever the p-value is less than α; or equivalently

$$|T_n| \geq \mathcal{Z}_{\alpha/2}.$$

The **rejection region** of a test is defined as the set of data points in the sample space that would lead to rejecting the null hypothesis. For one sided level α tests, the rejection region is

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

and for two-sided level α tests, the rejection region is

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

Power

In hypothesis testing, the sensitivity of a decision (i.e. level-α test) is evaluated by the probability of rejecting the null hypothesis when, in truth, there is a clinically important difference. This is referred to as the **power** of the test. We want power to be large; generally power is chosen to be .80, .90, .95. Let us denote by Δ_A the clinically important difference. This is the minimum value of the population parameter Δ that is deemed important to detect. If we are considering a one-sided hypothesis test, $H_0 : \Delta \leq 0$ versus $H_A : \Delta > 0$, then by defining the clinically important difference Δ_A, we are essentially saying that the region in the parameter space $\Delta = (0, \Delta_A)$ is an indifference region. That is, if, in truth, $\Delta \leq 0$, then we would want to conclude that the null hypothesis is true with high probability (this is guaranteed to be greater than or equal to $(1 - \alpha)$ by the definition of a level-α test). However, if, in truth, $\Delta \geq \Delta_A$, where $\Delta_A > 0$ is the clinically important difference, then we want to reject the null hypothesis in favor of the alternative hypothesis with high probability (probability greater than or equal to the power). These set of constraints imply that if, in truth, $0 < \Delta < \Delta_A$, then either the decision to reject or not reject the null hypothesis may plausibly occur and for such values of Δ in this indifference region we would be satisfied by either decision.

Thus the level of a one-sided test is

$$P_{\Delta=0}(\text{falling into the rejection region}) = P_{\Delta=0}(T_n \geq \mathcal{Z}_\alpha),$$

and the power of the test is

$$P_{\Delta=\Delta_A}(\text{falling into the rejection region}) = P_{\Delta=\Delta_A}(T_n \geq \mathcal{Z}_\alpha).$$

In order to evaluate the power of the test we need to know the distribution of the test statistic under the alternative hypothesis. Again, in most problems, the distribution of the test statistic T_n can be well approximated by a normal distribution. Under the alternative hypothesis

$$T_n \overset{H_A=(\Delta_A,\theta)}{\sim} N(\phi(n, \Delta_A, \theta), \sigma_*^2(\Delta_A, \theta)).$$

In other words, the distribution of T_n under the alternative hypothesis H_A follows a normal distribution with non zero mean which depends on the sample size n, the alternative Δ_A and the nuisance parameters θ. We denote this mean by $\phi(n, \Delta_A, \theta)$. The standard deviation $\sigma_*(\Delta_A, \theta)$ may also depend on the parameters Δ_A and θ.

Remarks

1. Unlike the null hypothesis, the distribution of the test statistic under the alternative hypothesis also depends on the nuisance parameters. Thus during the design stage, in order to determine the

power of a test and to compute sample sizes, we need to not only specify the clinically important difference Δ_A, but also plausible values of the nuisance parameters.

2. It is often the case that under the alternative hypothesis the standard deviation $\sigma_*(\Delta_A, \theta)$ will be equal to (or approximately equal) to one. If this is the case, then the mean under the alternative $\phi(n, \Delta_A, \theta)$ is referred to as the **non-centrality parameter**.

For example, when testing the equality in mean response between two treatments with normally distributed continuous data, we often use the t-test

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

which is approximately distributed as a standard normal under the null hypothesis. Under the alternative hypothesis $H_A : \mu_1 - \mu_2 = \Delta = \Delta_A$, the distribution of T_n will also be approximately normally distributed with mean

$$E_{H_A}(T_n) \approx E\left\{\frac{\bar{Y}_1 - \bar{Y}_2}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}\right\} = \frac{\mu_1 - \mu_2}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}} = \frac{\Delta_A}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}},$$

and variance

$$var_{H_A}(T_n) = \frac{\{var(\bar{Y}_1) + var(\bar{Y}_2)\}}{\sigma_Y^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{\sigma_Y^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sigma_Y^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 1.$$

Hence

$$T_n \overset{H_A}{\sim} N\left(\frac{\Delta_A}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}, 1\right).$$

Thus

$$\phi(n, \Delta_A, \theta) = \frac{\Delta_A}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}, \tag{6.1}$$

and

$$\sigma_*(\Delta_A, \theta) = 1. \tag{6.2}$$

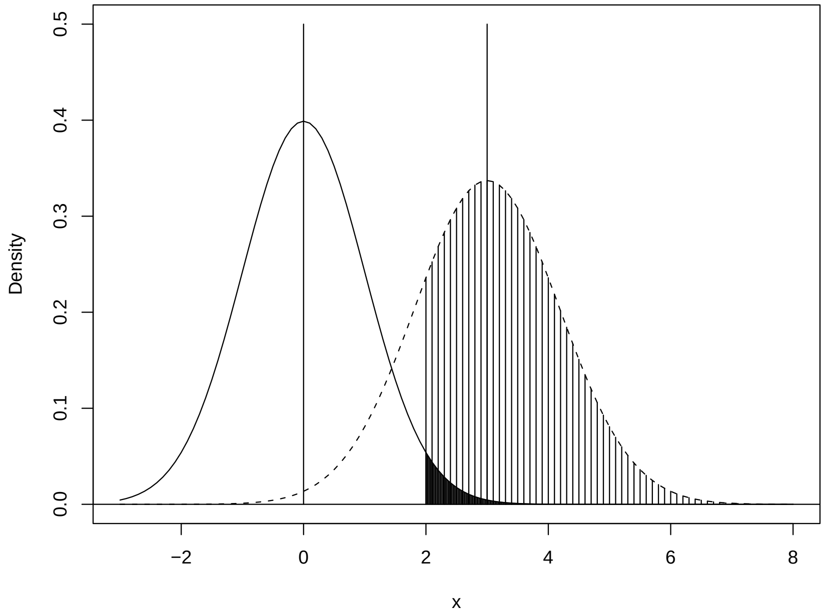
Hence $\left\{\frac{\Delta_A}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}\right\}$ is the non-centrality parameter.

Note: In actuality, the distribution of T_n is a non-central t distribution with $n_1 + n_2 - 2$ degrees of freedom and non-centrality parameter $\left\{\frac{\Delta_A}{\sigma_Y \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}\right\}$. However, with large n this is well approximated by the normal distribution given above.

6.2 Deriving sample size to achieve desired power

We are now in a position to derive the sample size necessary to detect a clinically important difference with some desired power. Suppose we want a level-α test (one or two-sided) to have power at least equal to $1-\beta$ to detect a clinically important difference $\Delta = \Delta_A$. Then how large a sample size is necessary? For a one-sided test consider the figure below.

Figure 6.1: *Distributions of T under H_0 and H_A*



It is clear from this figure that

$$\phi(n, \Delta_A, \theta) = \{\mathcal{Z}_\alpha + \mathcal{Z}_\beta \sigma_*(\Delta_A, \theta)\}. \tag{6.3}$$

Therefore, if we specify

- the level of significance (type I error) "α"
- the power (1 - type II error) "$1-\beta$"
- the clinically important difference "Δ_A"
- the nuisance parameters "θ"

then we can find the value n which satisfies (6.3).

Consider the previous example of normally distributed response data where we use the t-test to test for treatment differences in the mean response. If we randomize patients with equal probability to the two treatments so that $n_1 = n_2 \approx n/2$, then substituting (6.1) and (6.2) into (6.3), we get

$$\frac{\Delta_A}{\sigma_Y \left(\frac{4}{n}\right)^{1/2}} = (\mathcal{Z}_\alpha + \mathcal{Z}_\beta),$$

or

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

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

Note: For two-sided tests we use $\mathcal{Z}_{\alpha/2}$ instead of $\mathcal{Z}_\alpha$.

Example

Suppose we wanted to 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 expect the population standard deviation of response σ_Y to be about 60 units.

In this example $\alpha = .05$, $\beta = .10$, $\Delta_A = 20$ and $\sigma_Y = 60$. Also, $\mathcal{Z}_{\alpha/2} = \mathcal{Z}_{.025} = 1.96$, and $\mathcal{Z}_\beta = \mathcal{Z}_{.10} = 1.28$. Therefore,

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

or about 189 patients per treatment group.

6.3 Comparing two response rates

We will now consider the case where the primary outcome is a dichotomous response; i.e. each patient either responds or doesn't respond to treatment. Let π_1 and π_2 denote the population response rates for treatments 1 and 2 respectively. Treatment difference is denoted by $\Delta = \pi_1 - \pi_2$. We wish to test the null hypothesis $H_0 : \Delta \leq 0$ $(\pi_1 \leq \pi_2)$ versus $H_A : \Delta > 0$ $(\pi_1 > \pi_2)$. In some cases we may want to test the null hypothesis $H_0 : \Delta = 0$ against the two-sided alternative $H_A : \Delta \neq 0$.

A clinical trial is conducted where n_1 patients are assigned treatment 1 and n_2 patients are assigned treatment 2 and the number of patients who respond to treatments 1 and 2 are denoted by X_1 and X_2 respectively. As usual, we assume

$$X_1 \sim b(n_1, \pi_1)$$

and

$$X_2 \sim b(n_2, \pi_2),$$

and that X_1 and X_2 are statistically independent. If we let $\pi_1 = \pi_2 + \Delta$, then the distribution of X_1 and X_2 is characterized by the test parameter Δ and the nuisance parameter π_2. If we denote the sample proportions by $p_1 = X_1/n_1$ and $p_2 = X_2/n_2$, then we know from the properties of a binomial distribution that

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

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

This motivates the test statistic

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

where $\bar{p}$ is the combined sample proportion for both treatments; i.e. $\bar{p} = (X_1 + X_2)/(n_1 + n_2)$.

Note: The statistic T_n^2 is the usual chi-square test used to test equality of proportions.

We can also write

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

As such, $\bar{p}$ is an approximation (consistent estimator) for

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

where $\bar{\pi}$ is a weighted average of π_1 and π_2. Thus

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

The mean and variance of this test statistic under the null hypothesis $\Delta = 0$ (border of the null and alternative hypotheses for a one-sided test) are

$$E_{\Delta=0}(T_n) \approx E_{\Delta=0}\left\{\frac{p_1 - p_2}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}^{1/2}}\right\} = \frac{E_{\Delta=0}(p_1 - p_2)}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}^{1/2}} = 0,$$

$$var_{\Delta=0}(T_n) \approx \frac{\{var_{\Delta=0}(p_1) + var_{\Delta=0}(p_2)\}}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}} = \frac{\left\{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right\}}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}}.$$

But since $\pi_1 = \pi_2 = \bar{\pi}$, we get $var_{\Delta=0}(T_n) = 1$.

Because the distribution of sample proportions are approximately normally distributed, this will imply that the distribution of the test statistic, which is roughly a linear combination of independent sample proportions, will also be normally distributed. Since the normal distribution is determined by its mean and variance, this implies that,

$$T_n \overset{(\Delta=0)}{\sim} N(0,1).$$

For the alternative hypothesis $H_A : \Delta = \pi_1 - \pi_2 = \Delta_A$,

$$E_{H_A}(T_n) \approx \frac{(\pi_1 - \pi_2)}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}^{1/2}} = \frac{\Delta_A}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}^{1/2}},$$

and using the same calculations for the variance as we did above for the null hypothesis we get

$$var_{H_A}(T_n) \approx \frac{\left\{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right\}}{\left\{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}}.$$

When $n_1 = n_2 = n/2$, we get some simplification; namely

$$\bar{\pi} = (\pi_1 + \pi_2)/2 = (\pi_2 + \Delta_A/2)$$

and

$$var_{H_A}(T_n) = \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{2\bar{\pi}(1-\bar{\pi})}.$$

Note: The variance under the alternative is not exactly equal to one, although, if π_1 and π_2 are not very different, then it is close to one.

Consequently, with equal treatment allocation,

$$T_n \overset{H_A}{\sim} N\left(\frac{\Delta_A}{\left\{\bar{\pi}(1-\bar{\pi})\frac{4}{n}\right\}^{1/2}}, \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{2\bar{\pi}(1-\bar{\pi})}\right).$$

Therefore,

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

and

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

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

Using formula (6.3), the sample size necessary to have power at least $(1-\beta)$ to detect an increase of Δ_A, or greater, in the population response rate of treatment 1 above the population response rate for treatment 2, using a one-sided test at the α level of significance is

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

Hence

$$n = \frac{\left\{\mathcal{Z}_\alpha + \mathcal{Z}_\beta\left\{\frac{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{2\bar{\pi}(1-\bar{\pi})}\right\}^{1/2}\right\}^2 4\bar{\pi}(1-\bar{\pi})}{\Delta_A^2}. \tag{6.4}$$

Note: For two-sided tests we replace $\mathcal{Z}_\alpha$ by $\mathcal{Z}_{\alpha/2}$.

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

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

Substituting these values into (6.4) 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.

6.3.1 Arcsin square root transformation

Since the binomial distribution may not be well approximated by a normal distribution, especially when n is small (not a problem in most phase III clinical trials) or π is near zero or one, other approximations have been suggested for deriving test statistics that are closer to a normal distribution. We will consider the arcsin square root transformation which is a variance stabilizing transformation. Before describing this transformation, I first will give a quick review of the delta method for deriving distributions of transformations of estimators that are approximately normally distributed.

Delta Method

Consider an estimator $\hat{\gamma}_n$ of a population parameter γ such that

$$\hat{\gamma}_n \sim N(\gamma, \frac{\sigma^2_\gamma}{n}).$$

Roughly speaking, this means that

$$E(\hat{\gamma}_n) \approx \gamma$$

and

$$var(\hat{\gamma}_n) \approx \frac{\sigma^2_\gamma}{n}.$$

Consider the variable $f(\hat{\gamma}_n)$, where $f(\cdot)$ is a smooth monotonic function, as an estimator for $f(\gamma)$.

Using a simple Taylor series expansion of $f(\hat{\gamma}_n)$ about $f(\gamma)$, we get

$$f(\hat{\gamma}_n) = f(\gamma) + f'(\gamma)(\hat{\gamma}_n - \gamma) + \text{(small remainder term)},$$

where $f'(\gamma)$ denotes the derivative $\frac{df(\gamma)}{d\gamma}$. Then

$$E\{f(\hat{\gamma}_n)\} \approx E\{f(\gamma) + f'(\gamma)(\hat{\gamma}_n - \gamma)\}$$

$$= f(\gamma) + f'(\gamma)E(\hat{\gamma}_n - \gamma) = f(\gamma).$$

and

$$\begin{aligned} var\{f(\hat{\gamma}_n)\} &\approx var\{f(\gamma) + f'(\gamma)(\hat{\gamma}_n - \gamma)\} \\ &= \{f'(\gamma)\}^2 var(\hat{\gamma}_n) = \{f'(\gamma)\}^2 \left(\frac{\sigma_\gamma^2}{n}\right). \end{aligned}$$

Thus

$$f(\hat{\gamma}_n) \sim N\left(f(\gamma), \{f'(\gamma)\}^2 \left(\frac{\sigma_\gamma^2}{n}\right)\right).$$

Take the function $f(\cdot)$ to be the arcsin square root transformation; i.e

$$f(x) = sin^{-1}(x)^{1/2}.$$

If $y = sin^{-1}(x)^{1/2}$, then $sin(y) = x^{1/2}$. The derivative $\frac{dy}{dx}$ is found using straightforward calculus. That is,

$$\begin{aligned} \frac{dsin(y)}{dx} &= \frac{dx^{1/2}}{dx}, \\ cos(y)\frac{dy}{dx} &= \frac{1}{2}x^{-1/2}. \end{aligned}$$

Since $cos^2(y) + sin^2(y) = 1$, this implies that $cos(y) = \{1 - sin^2(y)\}^{1/2} = (1 - x)^{1/2}$. Therefore.

$$(1-x)^{1/2}\frac{dy}{dx} = \frac{1}{2}x^{-1/2},$$

or

$$\frac{dy}{dx} = \frac{1}{2}\{x(1-x)\}^{-1/2} = f'(x).$$

If $p = X/n$ is the sample proportion, where $X \sim b(n, \pi)$, then $var(p) = \frac{\pi(1-\pi)}{n}$. Using the delta method, we get that

$$\begin{aligned} var\{sin^{-1}(p)^{1/2}\} &\approx \{f'(\pi)\}^2 \left\{\frac{\pi(1-\pi)}{n}\right\}, \\ &= \left[\frac{1}{2}\{\pi(1-\pi)\}^{-1/2}\right]^2 \left\{\frac{\pi(1-\pi)}{n}\right\}, \\ &= \left\{\frac{1}{4\pi(1-\pi)}\right\}\left\{\frac{\pi(1-\pi)}{n}\right\} = \frac{1}{4n}. \end{aligned}$$

Consequently,

$$sin^{-1}(p)^{1/2} \sim N\left(sin^{-1}(\pi)^{1/2}, \frac{1}{4n}\right).$$

Note: The variance of $sin^{-1}(p)^{1/2}$ does not involve the parameter π, thus the term “variance stabilizing”.

The null hypothesis $H_0 : \pi_1 = \pi_2$ is equivalent to $H_0 : sin^{-1}(\pi_1)^{1/2} = sin^{-1}(\pi_2)^{1/2}$. This suggests that another test statistic which could be used to test H_0 is given by

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

The expected value of T_n is

$$E(T_n) = \frac{E\{sin^{-1}(p_1)^{1/2}\} - E\{sin^{-1}(p_2)^{1/2}\}}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)^{1/2}} \approx \frac{sin^{-1}(\pi_1)^{1/2} - sin^{-1}(\pi_2)^{1/2}}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)^{1/2}}.$$

and the variance of T_n is

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

$$\approx \frac{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)} = 1.$$

In addition to the variance stabilizing property of the arcsin square root transformation for the sample proportion of a binomial distribution, this transformed sample proportion also has distribution which is closer to a normal distribution. Since the test statistic T_n is a linear combination of independent arcsin square root transformations of sample proportions, the distribution of T_n will also be well approximated by a normal distribution. Specifically,

$$T_n \overset{H_0}{\sim} N(0, 1)$$

$$T_n \overset{H_A}{\sim} N\left(\frac{sin^{-1}(\pi_1)^{1/2} - sin^{-1}(\pi_2)^{1/2}}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)^{1/2}}, 1\right).$$

If we take $n_1 = n_2 = n/2$, then the non-centrality parameter equals

$$\phi(n, \Delta_A, \theta) = n^{1/2}\Delta_A,$$

where Δ_A, the clinically important treatment difference, is parameterized as

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

Consequently, if we parameterize the problem by considering the arcsin square root transformation, and use the test statistic above, then with equal treatment allocation, the sample size necessary to detect a clinically important treatment difference of Δ_A in the arcsin square root of the population proportions with power $(1-\beta)$ using a test (one-sided) at the α level of significance, is derived by using (6.3); yielding

$$n^{1/2}\Delta_A = (\mathcal{Z}_\alpha + \mathcal{Z}_\beta),$$

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

Remark: Remember to use radians rather than degrees in computing the arcsin or inverse of the sine function. Some calculators give the result in degrees where $\pi = 3.14159$ radians is equal to 180 degrees; i.e. radians$=\frac{\text{degrees}}{180} \times 3.14159$.

If we return to the previous example where we computed sample size based on the proportions test, but now instead use the formula for the arc sin square root transformation we would get

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

or 502 patients per treatment arm. This answer is virtually identical to that obtained using the proportions test. This is probably due to the relatively large sample sizes together with probabilities being bounded away from zero or one where the normal approximation of either the sample proportion of the arcsin square root of the sample proportion is good.

7 Comparing More Than Two Treatments

Suppose we randomize patients to $K > 2$ treatments and are interested in testing whether there are treatment differences in response. We proceed by positing the null hypothesis that there are no treatment differences and then test this null hypothesis against the alternative hypothesis that there exist some difference in response among the treatments.

Let us first consider the case where the response is dichotomous; i.e. the patient either responds or doesn't respond to treatment. Let $\pi_1, \ldots, \pi_K$ denote the population response rates for the K treatments under study. The null hypothesis of no treatment differences is given by

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

The alternative hypothesis H_A : states that the K population response rates "the π's" are not all equal.

To decide between H_0 and H_A, we conduct an experiment where we allocate $n_1, \ldots, n_K$ individuals on each of the K treatments respectively and count the number that respond to each of the treatments. Therefore, the data from such a clinical trial can be viewed as realizations of the independent random variables $X_1, \ldots, X_K$, where

$$X_i \sim b(n_i, \pi_i),\ i = 1, \ldots, K.$$

Each of the population response rates π_i are estimated using the sample proportions $p_i = X_i/n_i,\ i = 1, \ldots, K$.

The strategy for decision making is to combine the data from such an experiment into a test statistic for which larger values would provide increasing evidence against the null hypothesis. In addition, the distribution of this test statistic, under the null hypothesis, is necessary in order to gauge how extreme (evidence against H_0) the observed data are compared to what may have occurred by chance if the null hypothesis were true.

We begin by by first giving some general results regarding estimators that are normally distributed which will be used later for constructing test statistics.

7.1 Testing equality using independent normally distributed estimators

For $i = 1, \ldots, K$, let the estimator $\hat{\theta}_i$ be an unbiased estimator for the parameter θ_i which has a normal distribution; i.e.

$$\hat{\theta}_i \sim N(\theta_i, \sigma_i^2).$$

Assume σ_i^2 is either known or can be estimated well (consistently) using the data.

Suppose we have K independent estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$ of $\theta_1, \ldots, \theta_K$ respectively such that

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

If we assume that $\theta_1 = \ldots = \theta_K$ (i.e. similar to a K-sample null hypothesis), then the weighted estimator

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

is the best unbiased estimator for the common value θ.

Remark: By best, we mean that among all functions of $\hat{\theta}_1, \ldots, \hat{\theta}_K$ that are unbiased estimators of θ, $\hat{\bar{\theta}}$ (defined above) has the smallest variance.

In addition, when $\theta_1 = \ldots = \theta_K$,, then the random variable

$$\sum_{i=1}^K w_i (\hat{\theta}_i - \hat{\bar{\theta}})^2 \tag{7.1}$$

is distributed as a central chi-square distribution with $K - 1$ degrees of freedom. If the $\theta_i, i = 1, \ldots, K$ are not all equal, then

$$\sum_{i=1}^K w_i (\hat{\theta}_i - \hat{\bar{\theta}})^2 \tag{7.2}$$

is distributed as a non-central chi-square distribution with $K - 1$ degrees of freedom and non-centrality parameter equal to

$$\sum_{i=1}^K w_i (\theta_i - \bar{\theta})^2, \tag{7.3}$$

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 on the true population parameters.

7.2 Testing equality of dichotomous response rates

How do the results above help us in formulating test statistics? Returning to the problem of dichotomous response, the null hypothesis is

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

This is equivalent to

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

We showed in chapter 6 that if $p_i = X_i/n_i$ is the sample proportion responding to treatment i, then

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

Letting

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

then by (7.1), the test statistic

$$T_n = \sum_{i=1}^{K} 4n_i(sin^{-1}\sqrt{p_i} - \bar{A}_p)^2, \tag{7.4}$$

where

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

is distributed as a central chi-square distribution with $K-1$ degrees of freedom under the null hypothesis H_0.

This test statistic is a reasonable measure for assessing the strength of evidence of the alternative hypothesis. If the population response rates $\pi_i, i = 1, \ldots, K$ are all the same, then we would expect the sample proportions p_i to also be close to each other; in which case the test statistic T_n would be near zero. The more different the p_i's are from each other the greater T_n would be. Consequently, a reasonable strategy would be to reject the null hypothesis when the statistic T_n

is sufficiently large. The degree of evidence is obtained by gauging the observed value of the test statistic to the chi-square distribution with $K-1$ degrees of freedom. If we denote by $\chi^2_{\alpha;K-1}$, the $(1-\alpha)$-th quantile of the central chi-square distribution with $K-1$ degrees of freedom, then a test at the α-th level of significance can be obtained by rejecting H_0 when

$$T_n \geq \chi^2_{\alpha;K-1}.$$

Power and sample size calculations

In order to assess the power of the test, we need to know the distribution of the test statistic under the alternative hypothesis. Suppose we entertain a specific alternative hypothesis that we believe is clinically important to detect (more about this later). Say,

$$H_A : \pi_1 = \pi_{1A}, \ldots, \pi_K = \pi_{KA},$$

where the π_{iA}'s are not all equal. Using (7.2), the distribution of T_n is a non-central chi-square with $K-1$ degrees of freedom and non-centrality parameter given by (7.3), which in this case equals

$$\phi^2 = \sum_{i=1}^{K} 4n_i(sin^{-1}\sqrt{\pi_{iA}} - \bar{A}_{\pi A})^2, \tag{7.5}$$

where

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

The level-α test rejects H_0 when $T_n \geq \chi^2_{\alpha;K-1}$. Therefore, the power of the test to detect H_A is the probability that a non-central chi-square distribution, with $K-1$ degrees of freedom and non-centrality parameter ϕ^2, given by (7.5), exceed the value $\chi^2_{\alpha;K-1}$. See illustration below.

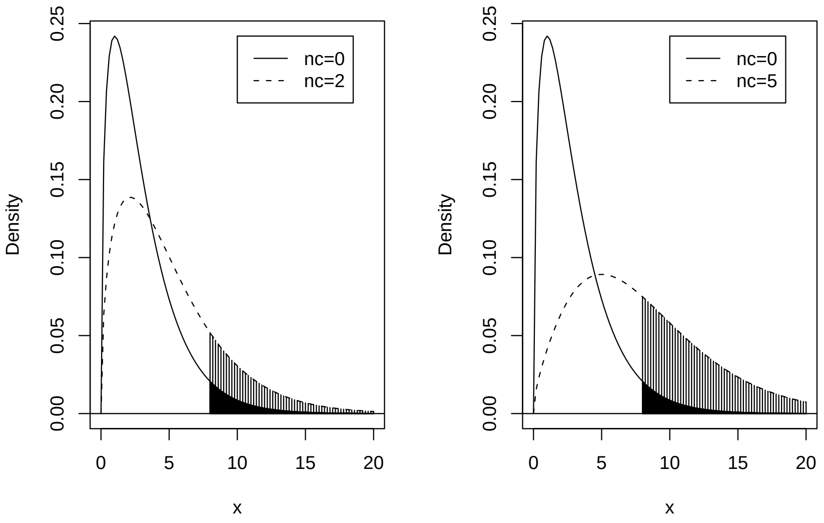
There are tables and/or computer packages available to carry out these probability calculations.

Sample size calculations

Suppose we allocate patients to each of the K treatments equally. Thus, on average, $n_1 = \ldots = n_K = n/K$, where n denotes the total sample size. In that case, the non-centrality parameter (7.5) equals

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

Figure 7.1: *Distributions of* T_n *under* H_0 *and* H_A



where

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

Note: $\bar{A}_{\pi A}$ is just a simple averages of the $sin^{-1}\sqrt{\pi_{iA}}$s for the alternative of interest.

Let us define by

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

the value of the non-centrality parameter necessary so that a non-central chi-square distributed random variable with $K-1$ degrees of freedom and non-centrality parameter $\phi^2(\alpha, \beta, K-1)$ will exceed the value $\chi^2_{\alpha;K-1}$ with probability $(1-\beta)$. Values for $\phi^2(\alpha, \beta, K-1)$ are given in tables or derived from software packages for different values of α, β and degrees of freedom $K-1$. For illustration, we have copied a subset of such tables that we will use later in examples.

Thus, if we want a level-α test of the null hypothesis $H_0 : \pi_1 = \ldots = \pi_K$ to have power $(1-\beta)$ to detect the alternative hypothesis $H_A : \pi_1 = \pi_{1A}, \ldots, \pi_K = \pi_{KA}$ (not all equal), then we need the non-centrality parameter

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

or

$$n = \frac{K\phi^2(\alpha, \beta, K-1)}{4\left\{\sum_{i=1}^{K}(sin^{-1}\sqrt{\pi_{iA}} - \bar{A}_{\pi A})^2\right\}}. \tag{7.7}$$

Choosing clinically important alternatives

Specifying alternatives that may be of clinical importance for comparing K treatments may not be that straightforward. (It is not that easy even for two-treatment comparisons). One conservative strategy is to find the sample size necessary to have sufficient power if any of the K treatments differ from each other by Δ_A or more.

Remark: Keep in mind that by using the arcsin square-root transformation, all treatment differences are measured on this scale.

To do this we must find the least favorable configuration of population response probabilities subject to the constraint that at least two treatment response rates on the arcsin square-root scale differ by Δ_A or more. By the least favorable configuration, we mean the configuration which would lead to the smallest non-centrality parameter. If we can find such a least favorable configuration, then this would imply that any other configuration has a larger non-centrality parameter, thus, any other configuration would have power at least as large as the least favorable one.

For a fixed sample size we showed in (7.6) that the non-centrality parameter is proportional to $\{\sum_{i=1}^{K}(sin^{-1}\sqrt{\pi_{iA}} - \bar{A}_{\pi A})^2\}$, although we will not prove this formally, it is intuitively clear that the least favorable configuration is obtained when two treatments, the one with the largest and smallest response, differ by Δ_A (in this case on the arcsin square root scale) and the other treatments have response half way in between.

It is clear from the picture above that

$$\{\sum_{i=1}^{K}(sin^{-1}\sqrt{\pi_{iA}} - \bar{A}_{\pi A})^2\} = \left(\frac{\Delta_A}{2}\right)^2 + 0 + \ldots + 0 + \left(\frac{\Delta_A}{2}\right)^2 = \frac{\Delta_A^2}{2}.$$

Substituting the result for the least favorable configuration in equation (7.7) yields

$$n = \frac{K\phi^2(\alpha, \beta, K-1)}{2\Delta_A^2}. \tag{7.8}$$

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?

For this example:

- $\alpha = .05$
- $1 - \beta = .90$, 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 (7.8), we get

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

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

7.3 Multiple comparisons

If we had done a two-treatment comparison for the previous example, and wanted 90% power to detect a difference from .30 to .40 in response rate at the .05 (two-sided) level of significance, then the sample size necessary is

$$n = \frac{2 \times 10.507}{2(.1051)^2} = 952,$$

or 476 patients per treatment arm.

For a four-treatment comparison, we needed 642 patients per treatment arm under similar parameter and sensitivity specifications. Why is there a difference and what are the implications of this difference?

We first note that when we test four treatments simultaneously, we can make six different pairwise comparisons; i.e.

$$1 \text{ vs } 2,\ 1 \text{ vs } 3,\ 1 \text{ vs } 4,\ 2 \text{ vs } 3,\ 2 \text{ vs } 4,\ 3 \text{ vs } 4.$$

If each of these comparisons were made using separate studies, each at the same level and power to detect the same alternative, then six studies would be conducted, each with 952 patients, or $952 \times 6 = 5,712$ total patients. This is in contrast to the 2,567 patients used in the four-treatment comparison.

This brings up the controversial issue regarding multiple comparisons. When conducting a clinical trial with K treatments, $K > 2$, there are $\begin{pmatrix} K \\ 2 \end{pmatrix}$ possible pairwise comparisons that can be made. If each comparison was made at the α level of significance, then even if the global null hypothesis were true; (i.e. in truth, all treatments had exactly the same response rate), the probability of finding at least one significant difference will be greater than α. That is, there is an inflated study-wide type I error that is incurred due to the multiple comparisons.

Let's be more specific. Denote by T_{nij} the test statistic used to compare treatment i to treatment j, $i < j,\ i, j = 1, \ldots, K$. There are $\begin{pmatrix} K \\ 2 \end{pmatrix}$ such pairwise treatment comparisons, each using the test statistic

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

For a two-sample comparison, the level α test (two-sided) would reject the hypothesis $H_{0ij} : \pi_i = \pi_j$ if

$$|T_{nij}| \geq \mathcal{Z}_{\alpha/2},$$

That is,

$$P_{H_{0ij}}(|T_{nij}| \geq \mathcal{Z}_{\alpha/2}) = \alpha.$$

However, if we made all possible pairwise treatment comparisons then the event of finding at least one significant result is

$$\bigcup_{i<j,\ i,j=1,\ldots,K} (|T_{nij}| \geq \mathcal{Z}_{\alpha/2}).$$

It is clear that

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

If we are concerned about this issue and want to control the overall study-wide type I error rate, then we would require greater evidence of treatment difference before declaring a pairwise-treatment comparison to be significant. One simple, albeit conservative, strategy is to use Bonferroni correction. This follows by noting that

$$P_{H_0}\{\bigcup_{i<j}(|T_{nij}| \geq c\} \leq \sum_{i<j} P_{H_{0ij}}(|T_{nij}| \geq c).$$

There are $\begin{pmatrix} K \\ 2 \end{pmatrix} = K(K-1)/2$ elements in the sum on the right hand side of the formula above. Therefore, if we choose the constant "c" so that each of the probabilities in the sum is equal to

$$\alpha/\begin{pmatrix} K \\ 2 \end{pmatrix},$$

then we are ensured that the probability on the left will be less than or equal to α. Since under the null hypothesis T_{nij} follows a standard normal distribution, we choose

$$c = \mathcal{Z}_{\alpha/\{K(K-1)\}}.$$

Hence, with a Bonferroni correction, we would declare a pairwise treatment comparison (say treatment i versus treatment j) significant if

$$|T_{nij}| \geq \mathcal{Z}_{\alpha/\{K(K-1)\}}.$$

Using this as the convention, the probability of declaring any treatment pairwise comparison significant under the null hypothesis is

$$P_{H_0}\{\bigcup_{i<j}(|T_{nij}| \geq \mathcal{Z}_{\alpha/\{K(K-1)\}}\} \leq \sum_{i<j} P_{H_{0ij}}(|T_{nij}| \geq \mathcal{Z}_{\alpha/\{K(K-1)\}}) = \begin{pmatrix} K \\ 2 \end{pmatrix} \times 2\alpha/\{K(K-1)\} = \alpha.$$

Example: If we were comparing four treatments, there would be six possible treatment comparisons. Using the Bonferroni method we would declare any such pairwise comparison significant at a study-wide global significance level of α if the nominal pairwise p-value was less than $\alpha/6$. For two sided tests, this would imply that the test statistic

$$|T_{nij}| \geq \mathcal{Z}_{\alpha/12}.$$

If we choose $\alpha = .05$, then a simple two-sample comparison requires the two-sample test to exceed 1.96 in order to declare significance. However, as part of a four-treatment comparison, if we use a Bonferroni correction, we need the two-sample test to exceed $\mathcal{Z}_{.05/12} = 2.635$ to declare significance.

Typically, the way one proceeds when testing for treatment differences with $K > 2$ treatments is to first conduct a global test of the null hypothesis at level α using (7.4). If this test fails to reject the null hypothesis, then we conclude that there are no treatment differences. If the null hypothesis is rejected, then we may want to consider the different pairwise treatment comparisons to determine where the treatment differences occur. A pairwise comparison would be declared significant based on the Bonferroni correction described above. Such a strategy conservatively protects all type I errors that can be committed to be less than α.

There are other more sophisticated methods for multiple comparisons which will not be discussed here. There are some clinical trials statisticians that object to the whole notion of correcting for multiple comparisons and believe that each pairwise comparison should be considered as a separate experiment.

Quote: Why should we be penalized for having the insight to test more than two treatments simultaneously which allows for $\begin{pmatrix} K \\ 2 \end{pmatrix}$ treatment comparisons as opposed to conducting $\begin{pmatrix} K \\ 2 \end{pmatrix}$ separate studies each of which individually would not be penalized. (By penalize we mean increased evidence of treatment difference before significance can be declared.)

The FDA's position is that they will consider such arguments but it must be agreed to up front with good rationale.

Chi-square tests for comparing response rates in K treatments

It was convenient, in developing the theory and deriving sample size calculations, to use the arcsin square-root transformation of sample proportions to test the equality of K treatment response rates. In general, however, the standard K-sample test used to test this null hypothesis is the chi-square test; namely,

$$\sum_{\text{over } 2\times K \text{ cells}} \frac{(O_j - E_j)^2}{E_j},$$

where O_j denotes the observed count in the j-th cell of a $2\times K$ contingency table of response-non-response by the K treatments, and E_j denotes the expected count under the null hypothesis. For comparison, we will construct both the chi-square test and the test based on the arcsin square-root transformation given by (7.4) on the same set of data.

Table 7.1: Observed Counts

	Treatment				
Response	1	2	3	4	Total
yes	206	273	224	275	978
no	437	377	416	364	1594
Total	643	650	640	639	2572

For each cell in the $2 \times K$ table, the expected count is obtained by multiplying the corresponding marginal totals and dividing by the grand total. For example, the expected number responding for treatment 1, under H_0, is $\frac{978\times 643}{2572} = 244.5$.

Table 7.2: Expected Counts

	Treatment				
Response	1	2	3	4	Total
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

$$\frac{(206-244.5)^2}{244.5} + \frac{(437-398.5)^2}{398.5} + \ldots + \frac{(364-396)^2}{396} = 23.43.$$

Gauging this value against a chi-square distribution with 3 degrees of freedom we get a p-value $< .005$.

Using the same set of data we now construct the K-sample test (7.4) using the arcsin square-root transformation.

Treatment i	p_i	$sin^{-1}\sqrt{p_i}$
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

$$T_n = 4\{643(.601 - .664)^2 + 650(.705 - .664)^2 + 640(.633 - .664)^2 + 639(.715 - .664)^2\} = 23.65.$$

Note: This is in good agreement with the chi-square test which, for the same set of data, gave a value of 23.43. This agreement will occur when sample sizes are large as often is the case in phase III clinical trials.

Example of pairwise comparisons

Suppose the data above were the results of a clinical trial. Now that we've established that it is unlikely that the global null hypothesis H_0 is true (p-value $< .005$), i.e. we reject H_0, we may want to look more carefully at the individual pairwise treatment differences.

Additionally, you are told that treatment 1 was the standard control and that treatments 2, 3 and 4 are new promising therapies. Say, that, up front in the protocol, it was stated that the major objectives of this clinical trial were to compare each of the new treatments individually to the standard control. That is, to compare treatments 1 vs 2, 1 vs 3, and 1 vs 4. If we want to control the overall experimental-wide error rate to be .05 or less, then one strategy is to declare any of the above three pairwise comparisons significant if the two-sided p-value were less than .05/3=.0167. With a two-sided test, we would declare treatment j significantly different than treatment 1, for $j = 2, 3, 4$ if the two-sample test

$$|T_{n1j}| \geq \mathcal{Z}_{.0167/2} = 2.385, \; j = 2, 3, 4.$$

Remark: The exact same considerations would be made using a one-sided test at the .025 level of significance. The only difference is that we would reject when $T_{nij} \geq 2.385$. We must be careful in defining T_{n1j} that the sign is such that large values are evidence against the one-sided null hypothesis of interest.

With that in mind, we consider the test based on the arcsin square-root transformation

$$T_{n1j} = 2\left(\frac{n_1 n_j}{n_1 + n_j}\right)^{1/2} (sin^{-1}\sqrt{p_j} - sin^{-1}\sqrt{p_1}).$$

Substituting the data above we get

$$T_{n12} = 2\left(\frac{643 \times 650}{643 + 650}\right)^{1/2} (.705 - .601) = 3.73*$$

$$T_{n13} = 2\left(\frac{643 \times 640}{643 + 640}\right)^{1/2} (.633 - .601) = 1.14$$

$$T_{n14} = 2\left(\frac{643 \times 639}{643 + 639}\right)^{1/2} (.715 - .601) = 4.08*$$

Thus we conclude that treatments 2 and 4 are significant better than treatment 1 (the standard control).

Suppose all four treatments were experimental, in which case, all six pairwise comparisons are of interest. To account for multiple comparisons, we would declare a pairwise comparison significant if the two-sided p-value were less than .05/6=.0083. Thus we would reject the null hypothesis $H_{0ij} : \pi_i = \pi_j$ when

$$|T_{nij}| \geq \mathcal{Z}_{.0083/2} = 2.635.$$

Note: the comparisons of treatments 1 vs 2 and 1 vs 4 would still be significant with this more stringent criterion.

Rounding out the remaining pairwise comparisons we get

$$|T_{n23}| = |2\left(\frac{650 \times 640}{650 + 640}\right)^{1/2} (.705 - .633)| = 2.59$$

$$|T_{n24}| = |2\left(\frac{650 \times 639}{650 + 639}\right)^{1/2} (.705 - .715)| = 0.36$$

$$|T_{n34}| = |2\left(\frac{640 \times 639}{640 + 639}\right)^{1/2} (.633 - .715)| = 2.94 * .$$

Clearly, treatments 2 and 4 are the better treatments, certainly better than control. The only controversial comparison is treatment 2 versus treatment 3, where, we may not be able to conclude that treatment 2 is significantly better than treatment 3 because of the conservativeness of the Bonferroni correction

7.4 K-sample tests for continuous response

For a clinical trial where we randomize patients to one of $K > 2$ treatments and the primary outcome is a continuous measurement, then our primary interest may be to test for differences in the mean response among the K treatments. Data from such a clinical trial may be summarized as realizations of the iid random vectors

$$(Y_i, A_i), i = 1, \ldots, n,$$

where Y_i denotes the response (continuously distributed) for the i-th individual and A_i denotes the treatment $(1, 2, \ldots, K)$ that the i-th individual was assigned. Let us denote the treatment-specific mean and variance of response by

$$E(Y_i|A_i = j) = \mu_j, j = 1, \ldots, K$$

and

$$var(Y_i|A_i = j) = \sigma^2_{Yj}, j = 1, \ldots, K.$$

Note:

1. Often, we make the assumption that the treatment-specific variances are equal; i.e. $\sigma^2_{Y1} = \ldots = \sigma^2_{YK} = \sigma^2_Y$, but this assumption is not necessary for the subsequent development.

2. Moreover, it is also often assumed that the treatment-specific distribution of response is normally distributed with equal variances; i.e.

$$(Y_i|A_i = j) \sim N(\mu_j, \sigma^2_Y), j = 1, \ldots, K$$

Again, this assumption is not necessary for the subsequent development.

Our primary focus will be on testing the null hypothesis

$$H_0 : \mu_1 = \ldots = \mu_K.$$

Let us redefine our data so that $(Y_{ij}, i = 1, \ldots, n_j, j = 1, \ldots, K)$ denotes the response for the i-th individual within treatment j, and n_j denotes the number of individuals in our sample assigned to treatment j ($n = \sum_{j=1}^{K} n_j$). From standard theory 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 μ_j and that asymptotically

$$\bar{Y}_j \sim N(\mu_j, \frac{\sigma^2_{Yj}}{n_j}), j = 1, \ldots, K.$$

Remark: If the Y's are normally distributed, then the above result is exact. However, with the large sample sizes that are usually realized in phase III clinical trials, the asymptotic approximation is generally very good.

Also, we know that the treatment-specific sample variance

$$s^2_{Yj} = \frac{\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2}{n_j - 1}$$

is an unbiased estimator for σ^2_{Yj}, and that asymptotically

$$\bar{Y}_j \sim N(\mu_j, \frac{s^2_{Yj}}{n_j}), j = 1, \ldots, K.$$

Remark: If the treatment specific variances are all equal, then the common variance is often estimated using the pooled estimator

$$s^2_Y = \frac{\sum_{j=1}^{K}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2}{n - K}.$$

Returning to the general results of section 7.1 of the notes, we let

- $\bar{Y}_j$ take the role of $\hat{\theta}_j$
- μ_j take the role of θ_j

- $\frac{s^2_{Yj}}{n_j}$ take the role of σ^2_j; hence $w_j = \frac{n_j}{s^2_{Yj}}$

Using (7.1), we construct the test statistic

$$T_n = \sum_{j=1}^{K} w_j(\bar{Y}_j - \bar{\bar{Y}})^2, \tag{7.9}$$

where

$$\bar{\bar{Y}} = \frac{\sum_{j=1}^{K} w_j \bar{Y}_j}{\sum_{j=1}^{K} w_j},$$

and $w_j = \frac{n_j}{s^2_{Yj}}$.

For the case where we are willing to assume equal variances, we get that

$$T_n = \frac{\sum_{j=1}^{K} n_j(\bar{Y}_j - \bar{\bar{Y}})^2}{s^2_Y}, \tag{7.10}$$

where

$$\bar{\bar{Y}} = \frac{\sum_{j=1}^{K} n_j \bar{Y}_j}{n}.$$

Under the null hypothesis H_0, T_n is approximately distributed as a chi-square distribution with $K-1$ degrees of freedom. Under the alternative hypothesis

$$H_A : \mu_1 = \mu_{1A}, \ldots, \mu_K = \mu_{KA}$$

where the μ_{jA}'s are not all equal, T_n is approximately distributed as a non-central chi-square distribution with $K-1$ degrees of freedom and non-centrality parameter

$$\phi^2 = \sum_{j=1}^{K} w_j(\mu_{jA} - \bar{\mu}_A)^2, \tag{7.11}$$

where

$$\bar{\mu}_A = \frac{\sum_{j=1}^{K} w_j \mu_{jA}}{\sum_{j=1}^{K} w_j},$$

and $w_j = \frac{n_j}{\sigma^2_{Yj}}$.

Assuming equal variances, we get the simplification

$$\phi^2 = \frac{\sum_{j=1}^{K} n_j(\mu_{jA} - \bar{\mu}_A)^2}{\sigma^2_Y}, \tag{7.12}$$

where

$$\bar{\mu}_A = \frac{\sum_{j=1}^{K} n_j \mu_{jA}}{n}.$$

Remark: In the special case where the response data are exactly normally distributed with equal treatment-specific variances, the test statistic T_n given by (7.10), under the null hypothesis, has a distribution which is exactly equal to $K-1$ times a central F distribution with $K-1$ numerator degrees of freedom and $n-K$ denominator degrees of freedom. That is

$$T_n/(K-1) = \frac{\sum_{j=1}^{K} n_j(\bar{Y}_j - \bar{\bar{Y}})^2/(K-1)}{s_Y^2}$$

has an F distribution under the null hypothesis. This is exactly the test statistic that is used for testing the equality of means in a one-way ANOVA model.

However, when the sample size n is large, we can use the test statistic given by (7.9), based on an asymptotic chi-square distribution, to test H_0 without making either the normality assumption or the assumption of equal variances.

7.5 Sample size computations for continuous response

Let us consider the case where patients are allocated equally to the K treatments so that

$$n_1 = \ldots = n_K = n/K,$$

and, for design purposes, we assume that the treatment specific variances are all equal which we posit to be the value σ_Y^2. The question is how do we compute the sample size that is necessary to have power $(1-\beta)$ to detect an alternative where any two treatments population mean responses may differ by Δ_A or more? Using considerations almost identical to those used for testing equality of response rates for dichotomous outcomes, we can find the least favorable configuration which after substituting into (7.12), yields the non-centrality parameter

$$\frac{n\Delta_A^2}{2K\sigma_Y^2}. \tag{7.13}$$

Hence, to obtain the desired power of $(1-\beta)$, we need

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

or

$$n = \frac{2K\sigma_Y^2\phi^2(\alpha, \beta, K-1)}{\Delta_A^2}. \tag{7.14}$$

Example:

We expand on the example used for two-sample comparisons given on page 84 of the notes, 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 (7.14), 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.

Remark: The 255 patients per arm represents an increase of 35% over the 189 patients per arm necessary in a two-sample comparison (see page 84 of notes). This percentage increase is the same as when we compare response rates for a dichotomous outcome with 4 treatments versus 2 treatments. This is not a coincidence, but rather, has to do with the relative ratio of the non-centrality parameters for a test with 3 degrees of freedom versus a test with 1 degree of freedom.

7.6 Equivalency Trials

The point of view we have taken thus far in the course is that of proving the superiority of one treatment over another. It may also be the case that there already exists treatments that have been shown to have benefit and work well. For example, a treatment may have been proven to be significantly better than placebo in a clinical trial and has been approved by the FDA and is currently on the market. However, there still may be room for other treatments to be developed that may be equally effective. This may be the case because the current treatment or treatments may have some undesirable side-effects, at least for some segment of the population, who would like to have an alternative. Or perhaps, the cost of the current treatments are high and some new treatments may be cheaper. In such cases, the company developing such a drug would like to demonstrate that their new product is equally effective to those already on the market or, at least, has beneficial effect compared to a placebo. The best way to prove that the new product has biological effect is to conduct a placebo-controlled trial and demonstrate superiority over the placebo using methods we have discussed. However, in the presence of established treatments that have already been proven effective, such a clinical trial would be un-ethical. Consequently, the new treatment has to be compared to one that is already known to be effective. The comparison treatment is referred to as an active or positive control.

The purpose of such a clinical trial would not necessarily be to prove that the new drug is better than the positive control but, rather, that it is equivalent in some sense. Because treatment comparisons are based on estimates obtained from a sample of data and thus subject to variation, we can never be certain that two products are identically equivalent in their efficacy. Consequently, a new drug is deemed equivalent to a positive control if it can be proved with high probability that it has response at least within some tolerable limit of the positive control. Of course the tricky issue is to determine what might be considered a tolerable limit for purposes of equivalency. If the positive control was shown to have some increase in mean response compared to placebo, say Δ^*, then one might declare a new drug equivalent to the positive control if it can be proved that the mean response of the new drug is within $\Delta^*/2$ of the mean response of the positive control or better with high probability. Conservatively, Δ^* may be chosen as the lower confidence limit derived from the clinical trial data that compared the positive control to placebo. Let us assume that the tolerable limit has been defined, usually, by some convention, or in negotiations of a company with the regulatory agency. Let us denote the tolerable limit by Δ_A.

Remark: In superiority trials we denoted by Δ_A, the clinically important difference that we wanted to detect with desired power. For equivalency trials, Δ_A refers to the tolerable limit.

Let us consider the problem where the primary response is a dichotomous outcome. (Identical arguments for continuous response outcomes can be derived analogously). Let π_2 denote the population response rate for the positive control, and π_1 be the population response rate for the new treatment.

Evaluating equivalency is generally stated as a one-sided hypothesis testing problem; namely,

$$H_0 : \pi_1 \leq \pi_2 - \Delta_A \text{ versus } H_A : \pi_1 > \pi_2 - \Delta_A.$$

If we denote by the parameter Δ the treatment difference $\pi_1 - \pi_2$, then the null and alternative hypotheses are

$$H_0 : \Delta \leq -\Delta_A \text{ versus } H_A : \Delta > -\Delta_A.$$

The null hypothesis corresponds to the new treatment being inferior to the positive control. This is tested against the alternative hypothesis that the new treatment is at least equivalent to the positive control. As always, we need to construct a test statistic, T_n, which, when large, would provide evidence against the null hypothesis and whose distribution at the border between the null and alternative hypotheses (i.e. when $\pi_1 = \pi_2 - \Delta_A$) is known. Letting p_1 and p_2 denote the sample proportion that respond on treatments 1 and 2 respectively, an obvious test statistic to test H_0 versus H_A is

$$T_n = \frac{p_1 - p_2 + \Delta_A}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}},$$

where n_1 and n_2 denote the number of patients allocated to treatments 1 and 2 respectively.

This test statistic was constructed so that at the border of the null and alternative hypotheses; i.e. when $\pi_1 = \pi_2 - \Delta_A$, the distribution of T_n will be approximately a standard normal; that is

$$T_n \overset{(\pi_1=\pi_2-\Delta_A)}{\sim} N(0,1).$$

Clearly, larger values of T_n give increasing evidence that the null hypothesis is not true in favor of the alternative hypothesis. Thus, for a level α test, we reject when

$$T_n \geq \mathcal{Z}_\alpha.$$

With this strategy, one is guaranteed with high probability ($\geq 1-\alpha$) that the drug will not be approved if, in truth, it is not at least equivalent to the positive control.

Remark: Notice that we didn't use the arcsin square-root transformation for this problem. This is because the arcsin square-root is a non-linear transformation; thus, a fixed difference of Δ_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 equivalency trials

In computing sample sizes for equivalency trials, one usually considers the power, i.e the probability of declaring equivalency, if, in truth, $\pi_1 = \pi_2$. That is, if, in truth, the new treatment has a response rate that is as good or better than the positive control, then we want to declare equivalency with high probability, say $(1-\beta)$. To evaluate the power of this test to detect the alternative $(\pi_1 = \pi_2)$, we need to know the distribution of T_n when $\pi_1 = \pi_2$.

Because

$$T_n = \frac{(p_1 - p_2 + \Delta_A)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx \frac{p_1 - p_2 + \Delta_A}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}},$$

straightforward calculations can be used to show that

$$E(T_n) \overset{(\pi_1=\pi_2=\pi)}{\approx} \frac{\Delta_A}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

and

$$var(T_n) \overset{(\pi_1=\pi_2=\pi)}{\approx} 1.$$

Hence

$$T_n \overset{(\pi_1=\pi_2=\pi)}{\sim} N\left(\frac{\Delta_A}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, 1\right),$$

and the non-centrality parameter equals

$$\phi(\cdot) = \frac{\Delta_A}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

If $n_1 = n_2 = n/2$, then

$$\phi(\cdot) = \frac{\Delta_A}{\sqrt{\pi(1-\pi)\left(\frac{4}{n}\right)}}.$$

To get the desired power, we solve

$$\frac{\Delta_A}{\sqrt{\pi(1-\pi)\left(\frac{4}{n}\right)}} = \mathcal{Z}_\alpha + \mathcal{Z}_\beta$$

or

$$n = \frac{(\mathcal{Z}_\alpha + \mathcal{Z}_\beta)^2 \times 4\pi(1-\pi)}{\Delta_A^2}. \qquad (7.15)$$

Generally, it requires larger sample sizes to establish equivalency because the tolerable limit Δ_A that the regulatory agency will agree to is small. For example, a pharmaceutical company has developed a new drug that they believe has similar effects to drugs already approved and decides to conduct an equivalency trial to get approval from the FDA to market the new drug. Suppose the clinical trial that was used to demonstrate that the positive control was significantly better than placebo had a 95% confidence interval for Δ (treatment difference) that ranged from .10-.25. Conservatively, one can only be relatively confident that this new treatment has a response rate that exceeds the response rate of placebo by .10. Therefore the FDA will only allow a new treatment to be declared equivalent to the positive control if the company can show that their new drug has a response rate that is no worse than the response rate of the positive control minus .05. Thus they require a randomized two arm equivalency trial to compare the new drug to the positive control with a type I error of $\alpha = .05$. The response rate of the positive control is about .30. (This estimate will be used for planning purposes). The company believes their drug is similar but probably not much better than the positive control. Thus, they want to have good power, say 90%, that they will be successful (i.e. be able to declare equivalency by rejecting H_0) if, indeed, their drug was equally efficacious. Thus they use formula (7.15) to derive the sample size

$$n = \frac{(1.64 + 1.28)^2 \times 4 \times .3 \times .7}{(.05)^2} = 2864,$$

or 1432 patients per treatment arm.

8 Causality, Non-compliance and Intent-to-treat

8.1 Causality and Counterfactual Random Variables

Throughout the course, we've repeatedly said that randomization allows us to make causal statements regarding the effect of treatment on the primary response of interest. In this section, we will be more formal. The actual definition of cause and effect has been hotly debated over the years especially by philosophers. We will take a particular point of view using what are called counterfactual random variables. As usual, we consider a super-population of individuals that we are interested in and assume that the participants in a clinical trial represent a random sample from this population. For concreteness, let us consider a clinical trial which will compare a new treatment (treatment 1) to no treatment or placebo (treatment 0).

We define the counterfactual random variable Y_1^* to denote the response (may be a binary or continuous outcome) that a randomly selected individual from our population if, possibly contrary to fact, that individual received treatment 1 (new treatment). Similarly, we define the counterfactual random variable Y_0^* to denote the response that a randomly selected individual from our population if, possibly contrary to fact, that individual received treatment 0 (placebo). We imagine the existence of both random variables (Y_0^*, Y_1^*) even though in actuality it would be impossible to observe both responses on any given individual. Thus the term counterfactual or "contrary to fact". At the individual level, we say that treatment causes the effect $Y_1^* - Y_0^*$. Clearly, if an individual knew their response to both treatment and placebo, then he/she would choose whichever gave the better response. Of course, this is not possible at the individual level but perhaps we can look at this question at the population level. That is, we will consider the population mean causal effect; namely $\Delta = E(Y_1^* - Y_0^*) = E(Y_1^*) - E(Y_0^*)$. If Δ is positive, then on average the response on treatment 1 will be better than on treatment 0. At the individual level this will not necessarily imply that any specific individual will be guaranteed to benefit, but on average, the population as a whole will benefit.

Knowledge of the average causal effect at the population level, if it can be obtained or estimated, is still useful at the individual level. Say, for example, it has been proved that Δ is positive; i.e. that on average treatment 1 is better than treatment 0, then in the absence of any additional a-priori knowledge that would distinguish one individual from the other, the best treatment choice for any individual in the population is treatment 1. If there are additional pre-treatment characteristics that can distinguish individuals from each other; say, covariates X (e.g. age, gender, race, etc.), then we would be interested in knowing or estimating the conditional expectation $\Delta(x) = E(Y_1^* - Y_0^*|X = x)$. If such a relationship were known, then the best choice for an individual whose X characteristics were equal to x, without any other additional knowledge of the individual, would be to choose treatment 1 if $\Delta(x) > 0$ and to choose treatment 0 if $\Delta(x) < 0$. The question then becomes whether we can estimate the average causal treatment effect Δ or the average causal treatment effect conditional on $X = x$, $\Delta(x)$, from a sample of individuals in a clinical trial.

The data that we actually get to observe from a clinical trial can be summarized by $(Y_i, A_i, X_i), i = 1, \ldots, n$, where, for a randomly selected individual i from our population, $A_i = (0, 1)$ denotes the treatment assignment, Y_i denotes the response and X_i denotes any additional characteristics, (prognostic factors) that are collected on the individual prior to treatment assignment (baseline characteristics). We will refer to these as the observable random variables. We distinguish here between the observed response Y_i for the i-th individual and the counterfactual responses Y_{1i}^*, Y_{0i}^*. We will, however, make the reasonable assumption that

$$Y_i = Y_{1i}^* I(A_i = 1) + Y_{0i}^* I(A_i = 0), \tag{8.1}$$

where $I(\cdot)$ denotes the indicator function of an event; that is, if the event is true this function equals 1, otherwise, it equals 0. In words, assumption (8.1) means that the observed response Y_i equals the counterfactual response Y_{1i}^* if the i-th individual were assigned treatment 1; whereas, the observed response would equal the counterfactual response Y_{0i}^* if the i-th individual were assigned treatment 0.

Traditional statistical methods and models allow us to make associational relationships regarding the probability distribution of the observable random variables. For example, we can posit regression models that allow us to estimate relationships such as $E(Y_i|A_i, X_i)$. However, these associational relationships are not the causal relationships that we argue are the important parameters of interest. Thus, the question is under what conditions or assumptions can we estimate causal parameters such as Δ or $\Delta(x)$, from the sample of observable data. This is where randomization plays a key role. Since the assignment of treatment to the patient in a randomized study is made using a random number generator, it is completely independent of any pre-treatment characteristics of the individual. An individual's counterfactual responses can be thought of as pre-destined inherent characteristics of that individual and, as such, can be reasonably assumed to be independent of treatment in a randomized clinical trial. That is, how an individual would have responded if given treatment 1 and how he/she would have responded if given treatment 0 would not have an effect on which treatment he/she was randomized to.

Thus, we make the assumption that

$$A_i \text{ is independent of } (Y^*_{1i}, Y^*_{0i}, X_i). \tag{8.2}$$

Remark: It is important to note that assumption (8.2) is not the same as saying that A_i is independent of Y_i. Since by assumption (8.1) $Y_i = Y^*_{1i}I(A_i = 1) + Y^*_{0i}I(A_i = 0)$ (i.e. Y_i is a function both of counterfactuals and treatment assignment), A_i being independent of (Y^*_{1i}, Y^*_{0i}) will not imply that A_i is independent of Y_i. In fact, if treatment is effective, as one hopes, then we would expect (and want) the distribution of Y_i to depend on A_i.

We will now use assumptions (8.1) and (8.2) to show that the distribution of the counterfactual random variable Y^*_{1i}, i.e. $P(Y^*_{1i} \le u)$ is the same as the conditional distribution $P(Y_i \le u|A_i = 1)$. This follows because

$$\begin{aligned} P(Y_i \le u|A_i = 1) &= P(Y^*_{1i} \le u|A_i = 1) &(8.3)\\ &= P(Y^*_{1i} \le u). &(8.4)\end{aligned}$$

Equation (8.3) is a consequence of assumption (8.1) and equation (8.4) is a consequence of assumption (8.2). Similarly, we can show that the distribution of the counterfactual random variable Y^*_{0i}, i.e. $P(Y^*_{0i} \le u)$ is the same as the conditional distribution $P(Y_i \le u|A_i = 0)$.

Consequently, the average causal treatment effect

$$\Delta = E(Y^*_1) - E(Y^*_0) = E(Y|A = 1) - E(Y|A = 0).$$

This is an important relationship as we now have an expression for the causal parameter Δ in terms of quantities that are functions of the distribution of the observable random variables. Thus, to estimate Δ it suffices to estimate $E(Y|A = 1)$ and $E(Y|A = 0)$. But these can be estimated easily using the treatment-specific sample averages. If we let $n_1 = \sum_{i=1}^n I(A_i = 1)$ and $n_0 = \sum_{i=1}^n I(A_i = 0)$, denote the treatment-specific sample sizes, then the treatment specific sample averages

$$\bar{Y}_1 = \sum_{i=1}^{n} Y_i I(A_i = 1)/n_1; \quad \bar{Y}_0 = \sum_{i=1}^{n} Y_i I(A_i = 0)/n_0,$$

are unbiased estimators for $E(Y|A = 1)$ and $E(Y|A = 0)$ respectively. Thus, an unbiased estimator for the causal treatment effect Δ can be derived from a randomized study using

$$\hat{\Delta} = \bar{Y}_1 - \bar{Y}_0.$$

Remark: The above arguments make formal what is intuitively obvious; that is, we can use the difference in the treatment-specific sample averages in a randomized clinical trial to estimate the true population treatment effect.

Similar arguments can also be used to show that in a randomized trial

$$\Delta(x) = E(Y_1^* - Y_0^*|X = x) = E(Y|A = 1, X = x) - E(Y|A = 0, X = x).$$

Thus, to estimate $\Delta(x)$, it suffices to estimate the conditional mean of the observable random variables $E(Y|A, X)$. This can be accomplished, say, by positing a model (either linear or nonlinear with or without interaction terms)

$$E(Y|A, X) = \mu(A, X, \beta),$$

in terms of unknown parameters β. For example, if X is a single covariate, we might consider a linear model with treatment-covariate interactions such as

$$\mu(A, X, \beta) = \beta_0 + \beta_1 A + \beta_2 X + \beta_3 AX. \tag{8.5}$$

From a sample of data $(Y_i, A_i, X_i), i = 1, \ldots, n$ we can estimate the parameters β in our model using standard statistical techniques such as least squares or weighted least squares. If we denote the estimator by $\hat{\beta}$, then the estimate for $\Delta(x)$ is given by

$$\hat{\Delta}(x) = \mu(1, x, \hat{\beta}) - \mu(0, x, \hat{\beta}).$$

For example, if we used model (8.5), then

$$\hat{\Delta}(x) = (\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 x) - (\hat{\beta}_0 + \hat{\beta}_2 x) = \hat{\beta}_1 + \hat{\beta}_3 x.$$

8.2 Noncompliance and Intent-to-treat analysis

The arguments outlined above assume that patients receive and take the treatment (or control) to which they are randomized. In many clinical trials, if not most, this is rarely the case. There is almost always some form of noncompliance from the intended treatment regimen.

Some reasons for a departure from the ideal treatment schedule are:

- 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 do
- 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.

How should we analyze the data when there is noncompliance? Some strategies that have been proposed and is a source of much debate between clinicians and statisticians include

- **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.

To illustrate some of the difficulties that can result from noncompliance, we consider some results from a study conducted by the Coronary Drug Project which was published in the New England Journal of Medicine, October, 1980, 303: 1038-1041. This study was a double-blind placebo-controlled clinical trial comparing Clofibrate to Placebo.

Table 8.1: *Intent-to-Treat Analysis*

	Clofibrate	Placebo
5 year mortality rate	.18	.19
number of patients	1065	2695

Table 8.2: *Clofibrate Patients Only*

Adherence	5 year	Number
(% of capsules taken)	mortality rate	patients
Poor (< 80%)	.25	357
Good (> 80%)	.15	708

p-value=.001

Table 8.3: *Clofibrate and Placebo Patients*

	Clofibrate		Placebo	
	5 year	number	5 year	number
Adherence	mortality	patients	mortality	patients
Poor (< 80%)	.25	357	.28	882
Good (> 80%)	.15	708	.15	1813

It is clear from this data that patients who comply are prognostically different from those who do not comply. Therefore, analyzing the data according to as-treated may lead to severe biases because we cannot separate out the prognostic effect of noncompliance from the prognostic effect of treatment. An intent-to-treat analysis does not suffer from this type of potentially biased exclusions; nonetheless, we intuitively realize that when some patients do not comply with the intended treatment then an intent-to-treat analysis would diminish the effect of a treatment. Let us look at this issue a bit more carefully with the use of counterfactual modeling.

8.3 A Causal Model with Noncompliance

We consider the following simple example for illustrative purposes.

- A randomized study is conducted where 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 do not have access to the study drug
- Compliance cannot be determined for patients randomized to placebo

Counterfactual and observable random variables

The problem above can be conceptualized as follows:

Let the counterfactual random variables (Y_1^*, Y_0^*) denote the response (1 if they respond, 0 if they don't respond) of a randomly selected individual in our population if they received treatment 1 or treatment 0 respectively. Also let C denote the counterfactual random variable corresponding to whether or not a randomly selected individual in our population would comply or not $C = (1, 0)$ if offered the new treatment. We refer to this as a counterfactual random variable because we will not know the compliance status if offered new treatment for patients randomized to placebo. Some of the population parameters associated with these counterfactuals are

- A complier (COM) and a noncomplier (NC) refer only to patients complying or not if offered active drug
- $\theta = P(C = 1)$ 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 8.4: *Hypothetical Population*

	COMPLIERS	NONCOMPLIERS
	θ	$(1-\theta)$
Treatment	π_1^{COM}	π_1^{NC}
Placebo	π_0^{COM}	π_0^{NC}
Difference	Δ^{COM}	Δ^{NC}

As we argued previously, it is not reasonable to assume that (Y_1^*, Y_0^*) are independent of C. Thus, we would **not** expect $\pi_1^{COM} = \pi_1^{NC}$ or $\pi_0^{COM} = \pi_0^{NC}$.

Using this notation and some simple probability calculations we get that

$$E(Y_1^*) = P(Y_1^* = 1) = \pi_1^{COM}\theta + \pi_1^{NC}(1-\theta) = \pi_1$$

and

$$E(Y_0^*) = P(Y_0^* = 1) = \pi_0^{COM}\theta + \pi_0^{NC}(1-\theta) = \pi_0.$$

Therefore, the average causal treatment effect equals

$$\Delta = E(Y_1^*) - E(Y_0^*) = \pi_1 - \pi_0 = \Delta^{COM}\theta + \Delta^{NC}(1-\theta).$$

The question is whether we can estimate these counterfactual parameters using the data we get to observe from a randomized clinical trial when there is noncompliance. We denote the observable data as (Y, A, AC), where $A = (1, 0)$ denotes the treatment that a patient is randomized to. C will denote the indicator of compliance which is only observed for patients randomized to active treatment. Thus the possible values that (A, AC) can take are

- $(0, 0)$ corresponding to a patient randomized to placebo
- $(1, 1)$ corresponding to a patient randomized to active treatment who complies
- $(1, 0)$ corresponding to a patient randomized to active treatment who does not comply

Finally, the observable random variable $Y = (0, 1)$ denotes the observed response.

Remark: In our scenario, if a patient is randomized to placebo, then that patient will not receive active treatment; whereas, if a patient is randomized to active treatment, then he/she will receive active treatment only if he/she complies; otherwise if the patient doesn't comply, then he/she will not receive active treatment.

The above considerations lead to the following reasonable assumptions:

$$Y = Y_0^* I(A=0) + Y_1^* I(A=1, C=1) + Y_0^* I(A=1, C=0)$$

and because of randomization

$$A \text{ is independent of } (Y_1^*, Y_0^*, C).$$

Because of these assumptions, we can equate some of the parameters regarding the distribution of the observable random variables to the parameters of the distribution of the counterfactual random variables.

Namely,

$$P(C=1|A=1)=P(C=1)=\theta$$

$$P(Y=1|A=0)=P(Y_0^*=1|A=0)=P(Y_0^*=1)=\pi_0$$

$$P(Y=1|A=1,C=1)=P(Y_1^*=1|A=1,C=1)=P(Y_1^*=1|C=1)=\pi_1^{COM}$$

$$P(Y=1|A=1,C=0)=P(Y_0^*=1|A=1,C=0)=P(Y_0^*=1|C=0)=\pi_0^{NC}$$

Note: All the probabilities above can be estimated using the corresponding sample proportions in a clinical trial.

Interestingly, since we can get estimates of π_0, π_0^{NC} and θ, then we can use the relationship that

$$\pi_0=\pi_0^{COM}\theta+\pi_0^{NC}(1-\theta)$$

to get that

$$\pi_0^{COM}=\frac{\pi_0-\pi_0^{NC}(1-\theta)}{\theta}.$$

In fact, the only counterfactual probability that cannot be estimated is π_1^{NC} as it is impossible to deduce the proportion of noncompliers who would have responded if forced to take treatment.

Intent-to-treat analysis

In an intent-to-treat analysis, where we compare the response rate among patients **randomized** to active drug to the response rate among patients **randomized** to placebo, we are estimating

$$\Delta_{ITT}=P(Y=1|A=1)-p(Y=1|A=0).$$

Again, by the assumptions made and some probability calculations we get

$$
\begin{aligned}
P(Y=1|A=1) &= P(Y=1|A=1,C=1)P(C=1|A=1) \\
&\quad + P(Y=1|A=1,C=0)P(C=0|A=1) \\
&= P(Y_1^*=1|A=1,C=1)P(C=1|A=1) + P(Y_0^*=1|A=1,C=0)P(C=0|A=1) \\
&= P(Y_1^*=1|C=1)P(C=1) + P(Y_0^*=1|C=0)P(C=0) \\
&= \pi_1^{COM}\theta + \pi_0^{NC}(1-\theta) \qquad (8.6)
\end{aligned}
$$

Also

$$P(Y=1|A=0) = P(Y_0^*=1|A=0) = P(Y_0^*=1) = \pi_0 = \pi_0^{COM}\theta + \pi_0^{NC}(1-\theta). \qquad (8.7)$$

Subtracting (8.7) from (8.6) we get that

$$\Delta_{ITT} = P(Y=1|A=1) - P(Y=1|A=0) = (\pi_1^{COM} - \pi_0^{COM})\theta,$$

or

$$\Delta_{ITT} = \Delta^{COM}\theta. \qquad (8.8)$$

Recall that

$$\Delta^{COM} = P(Y_1^*=1|C=1) - P(Y_0^*=1|C=1) = E(Y_1^* - Y_0^*|C=1)$$

is the difference in the mean counterfactual responses between treatment and placebo among patients that would comply with treatment. As such, Δ^{COM} is a causal parameter and some argue that it is the causal parameter of most interest since the patients that will benefit from a new treatment are those who will comply with the new treatment. Equation (8.8) makes it clear that the intention-to-treat analysis will yield an estimator which diminishes a causal treatment effect. In fact, since we are able to estimate the parameter θ, the probability of complying if offered the new treatment, then the causal parameter Δ^{COM} can be identified using parameters from the observable random variables; namely

$$\Delta^{COM} = \frac{P(Y=1|A=1) - P(Y=1|A=0)}{P(C=1|A=1)}. \qquad (8.9)$$

Since all the quantities on the right hand side of (8.9) are easily estimated from the data of a clinical trial, this means we can estimate the causal parameter Δ^{COM}.

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 Δ^{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 π_1^{NC}, we are not able to estimate Δ^{NC} or Δ

As-treated analysis

In one version of an as treated analysis we compare the response rate of patients randomized to receive active drug **who comply** to **all patients** randomized to receive placebo. That is, we compute

$$\Delta_{AT} = \pi_1^{COM} - \pi_0.$$

Since $\pi_0 = \pi_0^{COM}\theta + \pi_0^{NC}(1-\theta)$, after some algebra we get that

$$\Delta_{AT} = \Delta + (\pi_1^{COM} - \pi_1^{NC})(1-\theta),$$

where Δ denotes the average causal treatment effect. This makes clear that when there is noncompliance, $(\theta < 1)$, the as-treated analysis will yield an unbiased estimate of the average causal treatment effect only if $\pi_1^{COM} = \pi_1^{NC}$. As we've argued previously, this assumption is not generally true and hence the as-treated analysis can result in biased estimation even under the null hypothesis.

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 alway 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 Δ^{COM} is larger than the intent-to-treat estimator Δ_{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

In chronic disease clinical trials; e.g. Cancer, AIDS, Cardiovascular disease, Diabetes, etc., the primary endpoint is often time to an event, such as time to death, time to relapse of disease, etc. For such clinical trials the major focus is to compare the distribution of time to event among competing treatments.

Typically, the clinical trials occur over a finite period of time and consequently the time to event is not ascertained on all the patients in the study. This results in censored data. In addition, since patients enter the clinical trial at different calendar times (staggered entry), the length of follow-up varies by the individual. The combination of censoring and differential follow-up creates some unusual difficulties in the analysis of such data that do not allow standard statistical techniques to be used. Because of this, a whole new research area in Statistics has emerged to study such problems. This is called **Survival Analysis** or **Censored Survival Analysis**. A brief introduction of Survival Analysis will be given in this chapter, but for a more thorough study of this area I recommend the course in Applied Survival Analysis offered every Spring semester and for individuals interested in a more rigorous treatment of the subject there is a course on Advanced Survival Analysis offered every other year.

In survival analysis, the endpoint of interest is time to an event which we denote by the positive random variable T. Some examples include

- 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

In order to be unambiguous, the start and end of the event must be clearly identified.

9.1 Describing the Distribution of Time to Event

We will describe some different, but equivalent, ways to define the distribution of the random variable, T, "time to event."

- The distribution function:
$$F(t) = P(T \leq t);$$
- The survival function:
$$S(t) = P(T \geq t);$$

The right-continuous version of the survival function will be denoted by

$$S(t^-) = P(T > t) = 1 - F(t).$$

Remark: For the most part, we will assume that T is a continuous random variable in which case $S(t^-) = S(t) = 1 - F(t)$. We will also assume that T has a density function

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}.$$

Clearly:

$$F(t) = \int_0^t f(u)du,$$

and

$$S(t) = \int_t^\infty f(u)du.$$

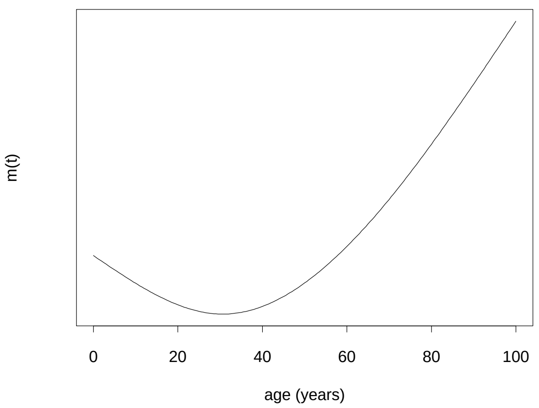
Hazard rate

The hazard rate is a useful way of defining the distribution of a survival time which can also be used to describe the aging of a population. We motivate the definition of a hazard rate by first introducing "mortality rate" or discrete hazard rate.

The mortality rate at time t (where t is usually taken to be an integer of some unit of time; i.e. day, week, month, year, etc.) is the proportion of the population who fail between times t and $(t + 1)$ among individuals alive (who have not failed) at time t.

$$m(t) = P(t \leq T < t + 1 | T \geq t).$$

Figure 9.1: *A typical mortality pattern for human*



In a human population, the mortality rate has a pattern like

The hazard rate $\lambda(t)$ is the limit of the mortality rate or the instantaneous rate of failure at time t given the individual is alive at time t. That is,

$$\lambda(t) = \lim_{h \to 0} \left\{ \frac{P(t \leq T < t+h | T \geq t)}{h} \right\}.$$

This can be expressed as

$$\begin{aligned} \lambda(t) &= \lim_{h \to 0} \left\{ \frac{P(t \leq T < t+h)/h}{P(T \geq t)} \right\} = \frac{f(t)}{S(t)} \\ &= \frac{-\frac{dS(t)}{dt}}{S(t)} = \frac{-d\log\{S(t)\}}{dt}. \end{aligned}$$

Integrating both sides of the equation above, we get

$$-\log\{S(t)\} = \int_0^t \lambda(u)du = \Lambda(t),$$

where $\Lambda(t)$ is defined as the **cumulative hazard function**. Consequently,

$$\begin{aligned} S(t) &= \exp\left\{-\int_0^t \lambda(u)du\right\} \\ &= \exp\left\{-\Lambda(t)\right\}. \end{aligned}$$

Note: Although the mortality rate is a probability, the hazard rate is NOT a probability; thus it can take on any positive value unlike a mortality rate which must be bounded by 1.

The mortality rate

$$\begin{aligned} m(t) &= \frac{P(T \geq t) - P(T \geq t+1)}{P(T \geq t)} \\ &= 1 - \frac{P(T \geq t+1)}{P(T \geq t)} \\ &= 1 - \frac{\exp\{-\Lambda(t+1)\}}{\exp\{-\Lambda(t)\}} \\ &= 1 - \exp\left\{ - \int_t^{t+1} \lambda(u) du \right\}. \end{aligned}$$

Notice that if the probability of an event occurring in a single time unit is small and the hazard rate doesn't change quickly within that time unit, then the hazard rate is approximately the same as the mortality rate. To see this, note that

$$\begin{aligned} m(t) &= 1 - \exp\left\{ - \int_t^{t+1} \lambda(u) du \right\} \approx 1 - \left\{ 1 - \int_t^{t+1} \lambda(u) du \right\} \\ &= \int_t^{t+1} \lambda(u) du \approx \lambda(t). \end{aligned}$$

Also, by definition, the hazard rate depends on the time scale being used. Therefore, at the same point in time the hazard rate in days is 1/365 times the hazard rate in years.

Because of the one-to-one relationships that were previously derived, the distribution of a continuous survival time T can be defined by any of the following:

$$S(t), F(t), f(t), \lambda(t).$$

Exponential distribution

If the hazard rate is constant over time

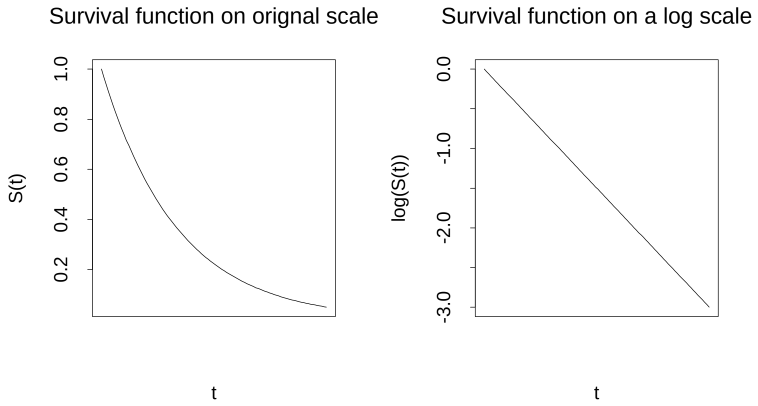
$$\lambda(t) = \lambda, \text{ then}$$

$$S(t) = \exp\left\{ - \int_0^t \lambda(u) du \right\} = \exp(-\lambda t).$$

This is an exponential distribution with hazard equal to λ. Sometimes this is referred to as the negative exponential.

It is sometimes useful to plot the log survival probability over time. This is because $-\log\{S(t)\} = \Lambda(t)$.

Figure 9.2: *The survival function of an exponential distribution on two scales*



If T follows an exponential distribution with hazard rate λ, then the the median survival time

$$\{m : P(T \geq m) = .5\}; \exp(-\lambda m) = .5; m = -\log(.5)/\lambda, = \log(2)/\lambda = .6931/\lambda,$$

and the mean survival time

$$\mathrm{E}(T) = \int_0^\infty t\lambda \exp(-\lambda t) dt = \lambda^{-1}.$$

Other parametric models commonly used

Weibull distribution

The hazard function of the Weibull distribution is given by

$$\lambda(t) = \lambda t^{\gamma-1};\ \lambda, \gamma > 0.$$

This is a two-parameter model with the scale parameter λ and the shape parameter γ. This model allows us to consider hazard functions which increase or decrease over time according to the choice of γ.

$$S(t) = \exp\frac{(-\lambda t^\gamma)}{\gamma}.$$

Gompertz-Makeham distribution

This distribution is useful for modeling the hazard function of human populations especially later in life. The hazard function is given by

$$\lambda(t) = \theta + \beta \mathrm{e}^{\gamma t}.$$

Must be careful to choose θ, β, γ so that $\lambda(t) \geq 0$ and if we also want a proper distribution, i.e. $S(t) \rightarrow 0$ as $t \rightarrow \infty$ then

$$\Lambda(\infty) = \infty.$$

Other popular distributions include the log normal distribution where

$$\log(T) \sim N(\mu, \sigma^2),$$

and the gamma distribution whose density

$$f(t) \text{ is proportional to } t^{\rho} \mathrm{e}^{-\lambda t}.$$

Remark: In most clinical trials applications and research in survival analysis it has become common practice to use non-parametric and semi-parametric models where the shape of the distribution function is left unspecified.

9.2 Censoring and Life-Table Methods

Two important issues in clinical trials where the primary endpoint of interest is time to an event which are different than most studies:

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.

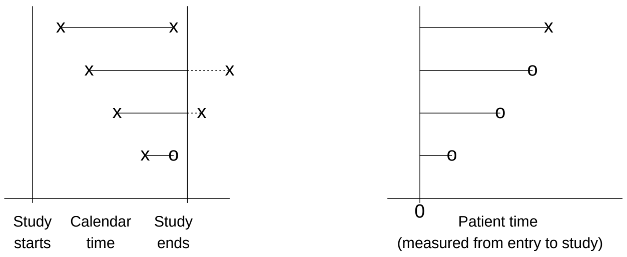
This is illustrated in the schematic shown in the next page.

The time to event of interest in most clinical trials is the time from entry into the study until death (right-hand panel).

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)

Figure 9.3: *Illustration of censored data*



- death from other causes

  (competing risks; e.g. gets run over by a bus)

The above are examples of what is called random right censoring. That is, we conceptualize a random variable (possibly unobserved) corresponding to the potential time that an individual may be censored. This censoring time is a random variable which varies by individual. Random right censoring creates unique difficulties which does not allow the use of standard inferential techniques. This is illustrated in the following example from a study of 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 into study	Number at risk at beginning of interval	Number dying in interval	Number censored 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

Question: Estimate the five-year mortality rate? Two naive estimators are as follows:

1. $\frac{\text{76 deaths in 5 years}}{\text{146 individuals}} = .521, \ \hat{S}(5) = .479$
2. $\frac{\text{76 deaths in 5 years}}{\text{146 - 29 (withdrawn)}} = .650, \ \hat{S}(5) = .350.$

Estimator 1. corresponds to censoring on the right; that is, if everyone that was withdrawn in the first 5 years was withdrawn exactly at 5 years, then this approach would give an unbiased estimator. Since this isn't what happened, this estimator is too **optimistic**.

In contrast, estimator 2. would be appropriate if everyone that was withdrawn in the first 5 years was withdrawn immediately at time "0". If this were the case then this approach would yield an unbiased estimator. Since this isn't what happened, this estimator is too **pessimistic**.

The more appropriate method uses life-table estimates, illustrated as follows:

Assume censoring occurs at the right of each yearly interval

Time	n_r	d	w	$m^R = d/n_r$	$1 - m^R$	$\hat{S}^R = \Pi(1 - m^R)$
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 estimate = .432

5 year mortality rate estimate = .568

Assume censoring occurs at the left of each interval

time	n_r	d	w	$m^L = d/(n_r - w)$	$1 - m^L$	$\hat{S} = \Pi(1 - m^L)$
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 estimate = .400

5 year mortality rate = .600

We note that the naive estimator for the five year survival probability ranged from .35 to .479, whereas the life-table estimates ranged from .40 to .432 depending on whether we assumed censoring occurred on the left or right of each interval.

More than likely, censoring occurred during the interval. Thus $\hat{S}^L$ and $\hat{S}^R$ are under and over estimates respectively. A compromise would be to use

$$m = d/(n_r - w/2) \quad \text{in the tables above.}$$

This is what is referred to as the life-table estimate and for this example leads to the estimate of the 5 year survival probability $\hat{S}(5) = .417$.

Since the life-table estimator is an estimator for the underlying population survival probability based on a sample of data, it is subject to variability. To assess the variability of this estimator, the standard error can be computed using Greenwood's formulas as follows:

$$\text{se}\{\hat{S}(t)\} = \hat{S}(t)\left\{\sum_{j=1}^{t} \frac{d_j}{(n_{rj} - w_j/2)(n_{rj} - d_j - w_j/2)}\right\}^{1/2}.$$

With sufficiently large sample sizes this estimator is approximately normally distributed; in which case, the $(1-\alpha)^{\text{th}}$ 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	$\hat{S}$	$\frac{d}{(n_r-w/2)(n_r-d-w/2)}$	$\sum \frac{d}{(n-w/2)(n-d-w/2)}$	se
0-1	146	27	3	.813	.00159	.00159	.032
1-2	118	18	10	.681	.00168	.00327	.039
2-3	88	21	10	.509	.00408	.00735	.044
3-4	57	9	3	.426	.00345	.01084	.044
4-5	45	1	3	.417	.00054	.01138	.044

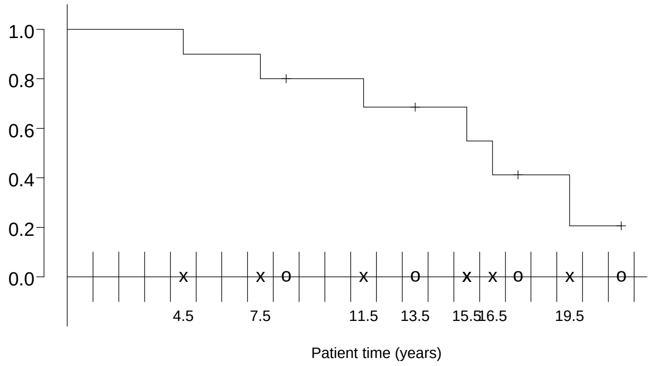
The 95% confidence interval for $S(5)$ is given as $.417 \pm 1.96(.044) = (.331 - .503)$.

9.3 Kaplan-Meier or Product-Limit Estimator

We notice in our previous example that the bias that occurs in estimating the survival distribution by incorrectly assuming that censoring occurs at the left or right of each interval was decreased when the interval was taken to be smaller (i.e. 1 year intervals as opposed to 5 year intervals). If the data were not grouped (i.e. we know the exact times to death or censoring), then this suggests that we may want to apply the life-table estimator using many intervals with small interval widths. The limit of the life-table estimator when the intervals are taken so small that at most only one observation occurs within any interval is called the product-limit estimator which is also the same as the Kaplan-Meier estimator. Kaplan and Meier (1958) derived the estimator based on likelihood principles. We believe it is more instructive and intuitive to consider this estimator as the limit of the life-table estimator.

To illustrate how this estimator is constructed, consider the following example:

Figure 9.4: *An illustrative example of Kaplan-Meier estimator*



$$\begin{array}{rcccccccccccccccccccccc}
1-\hat{m}(x): & 1 & 1 & 1 & 1 & \frac{9}{10} & 1 & 1 & \frac{8}{9} & 1 & 1 & 1 & \frac{6}{7} & 1 & 1 & 1 & \frac{4}{5} & \frac{3}{4} & 1 & 1 & \frac{1}{2} & 1 & 1 \\
\hat{S}(t): & 1 & 1 & 1 & 1 & \frac{9}{10} & . & . & \frac{8}{10} & . & . & . & \frac{48}{70} & . & . & . & \frac{192}{350} & \frac{144}{350} & . & . & \frac{144}{700} & . & .
\end{array}$$

$$\begin{aligned}
m = d/n_r &= \frac{\text{number of deaths in an interval}}{\text{number at risk at beginning of interval}} \\
&= (1/n_r \text{ or } 0 \text{ depending on whether or not a death occurred in interval}) \\
(1-m) &= (1-d/n_r) = ((1-1/n_r) \text{ or } 1)\,.
\end{aligned}$$

In the limit, the Kaplan-Meier (product-limit) estimator will be a step function taking jumps at times where a failure occurs. Therefore at any time t, the product-limit estimator of the survival distribution is computed as the product

$$\prod_{\text{all deaths}} \left(1 - \frac{1}{\text{number at risk}}\right)$$

over all death times occurring up to and including time t.

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.

Some notation and software

In describing censored survival data, it is useful to conceptualize two latent random variables (possibly unobserved) corresponding to the failure time and censoring time. For the i-th individual we will denote these by T_i and C_i respectively. Specifically, T_i represents the survival time if that individual was followed until death; whereas, C_i corresponds to the time measured from their entry into the study until they were censored in the hypothetical situation that they could not die. For example, C_i may be the time from entry into the study until the time the final analysis was conducted. However, if the individual could be lost to follow-up, then the variable C_i would have to account for that possibility as well. In any case, C_i corresponds to the time that an individual would have been censored in a study if their death could be prevented.

In contrast to these latent variables, the variables we actually get to observe for the i-th individual are denoted by (U_i, Δ_i), where U_i denotes the observed time on study (i.e. the time to death or censoring, and Δ_i denotes the failure indicator taking on the value 1 if the patient is observed to die and the value 0 if the patient is censored. In terms of the latent variables we assume $U_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \le C_i)$.

The main objective of a clinical trial is to make inference about the probability distribution of the latent survival time T_i even though this variable is not always observed due to censoring. In the one-sample problem we are interested in estimating the survival distribution $S(t) = P(T_i \ge t)$ using a sample of observable data

$$(U_i, \Delta_i),\ i = 1, \ldots, n.$$

If we define the number of individuals at risk at any time t by

$$n(t) = \sum_{i=1}^{n} I(U_i \ge t),$$

that is, the number of individuals in our sample who neither died or were censored by time t, then the Kaplan-Meier estimator is given by

$$KM(t) = \prod_{i:U_i \le t} \left\{ \frac{n(U_i) - 1}{n(U_i)} \right\}^{\Delta_i}.$$

This is the Kaplan-Meier estimator when there are no tied survival times in our sample. More generally, if we denote by

$$d(t) = \sum_{i=1}^{n} I(U_i = t, \Delta_i = 1),$$

the number of observed deaths in our sample at time t, thus allowing the possibility that $d(t) \geq 2$ in cases where survival times are tied, then we can write the Kaplan-Meier estimator as

$$KM(t) = \prod_{\text{death times } u \le t} \left\{ 1 - \frac{d(u)}{n(u)} \right\}.$$

The standard error of the Kaplan-Meier estimator is also taken as the limit in Greenwood's formula, Namely,

$$se\{KM(t)\} = KM(t) \left\{ \sum_{\text{death times } u \le t} \frac{d(u)}{n(u)\{n(u) - d(u)\}} \right\}^{1/2}.$$

Proc lifetest in SAS

Many statistical packages, including SAS, have software available for censored survival analysis. For example, the above Kaplan-Meier estimator can be obtained using the following SAS program:

```
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;
```

And part of the output from the above program is

```
                        The LIFETEST Procedure

                   Product-Limit Survival Estimates

                                           Survival
                                           Standard     Number      Number
 SURVTIME      Survival      Failure        Error       Failed       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 Tests

The major objective of many Phase III clinical trials is to compare two or more treatments with respect to some primary endpoint. If the primary endpoint is time to an event (e.g. survival time), then interest will focus on whether one treatment will increase or decrease the distribution of this time as compared to some standard or control treatment. Let us begin by considering the comparison of two treatments. Let the variable A denote treatment group, where we take $A = 0$ to denote the control group or standard treatment and $A = 1$ the new treatment.

The problem of comparing two treatments is often cast as a hypothesis testing question. The null hypothesis being that the distribution of time to death (event) is the same for both treatments. Letting T denote a patient's underlying survival time, we define the treatment specific survival distributions by $S_1(t) = P(T \geq t|A = 1)$ and $S_0(t) = P(T \geq t|A = 0)$. The null hypothesis is given as

$$H_0 : S_1(t) = S_0(t) = S(t),\ t > 0,$$

or equivalently

$$H_0 : \lambda_1(t) = \lambda_0(t) = \lambda(t),$$

where $\lambda_j(t), j = 0, 1$ denote the treatment-specific hazard rates.

The alternative hypothesis of most interest in such trials is that the survival time for one treatment is stochastically larger than the survival time for the other treatment. Specifically, we say the survival time for treatment 1 is stochastically larger than the survival time for treatment 0 if $S_1(t) \geq S_0(t)$ for all $t > 0$ with strict inequality for at least one value of t.

It has become standard practice in clinical trials to use nonparametric tests; that is, tests based on statistics whose distribution under the null hypothesis does not depend on the underlying survival distribution $S(t)$ (At least asymptotically). The most widely used test with censored survival data is the **logrank test** which we now describe.

Data from a clinical trial comparing the survival distribution between two treatments can be viewed as realizations of the random triplets

$$(U_i, \Delta_i, A_i), i = 1, \ldots, n,$$

where

- $U_i = \min(T_i, C_i)$
  - 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

We also define the following notation:

- $n_j = \sum_{i=1}^n I(A_i = j)$ 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 notation above allows the possibility of more than one death occurring at the same time (tied survival times).

The logrank test is based on the statistic

$$\sum_{\text{all death times } u} \left\{ d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right\}.$$

This statistic can be viewed as the sum over the distinct death times of the observed number of deaths from treatment 1 minus the expected number of deaths from treatment 1 if the null hypothesis were true.

Thus at any point in time u corresponding to a time where a death was observed, i.e. $d(u) \geq 1$, the data at that point in time can be viewed as a 2 × 2 table; namely,

	treatment		
	1	0	total
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 $\left\{ d_1(u) - \frac{d(u)}{n(u)} n_1(u) \right\}$

From this point of view, the survival data from a clinical trial can be summarized as k 2 × 2 tables, where k denotes the number of distinct death times. If the null hypothesis is true, then we would expect $\left\{ d_1(u) - \frac{d(u)}{n(u)} n_1(u) \right\}$ to be about zero on average for all $\{u : d(u) \geq 1\}$. However, if the hazard rate for treatment 0 is greater than the hazard rate for treatment 1 consistently over all u, then, on average, we would expect $\left\{ d_1(u) - \frac{d(u)}{n(u)} n_1(u) \right\}$ to be negative. The opposite would be expected if the hazard rate for treatment 1 was greater than the hazard rate for treatment 0 for all u.

This suggests that the null hypothesis of treatment equality should be rejected if the test statistic is sufficiently large or small depending on the alternative of interest for one-sided tests or if the absolute value of the test statistic is sufficiently large for two-sided tests. In order to gauge the strength of evidence against the null hypothesis we must be able to evaluate the distribution of the test statistic (at least approximately) under the null hypothesis. Therefore, the test statistic has to be standardized appropriately. Specifically, the logrank test is given by

$$T_n = \frac{\sum \left\{ d_1(u) - \frac{d(u)}{n(u)} n_1(u) \right\}}{\left[\sum \frac{n_1(u) n_0(u) d(u) \{n(u) - d(u)\}}{n^2(u)\{n(u)-1\}} \right]^{1/2}}. \tag{9.1}$$

Remark: In a 2×2 contingency table

$d_1(u)$	$\cdot$	$d(u)$
$\cdot$	$\cdot$	$n(u) - d(u)$
$n_1(u)$	$n_0(u)$	$n(u)$

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

$$\left[\sum \frac{n_1(u) n_0(u) d(u) \{n(u) - d(u)\}}{n^2(u)\{n(u)-1\}} \right].$$

The sum of the hypergeometric variances of these 2×2 tables, summed over the distinct death times, is the estimator of the variance of the test statistic under the null hypothesis. Therefore, the logrank test T_n given by (9.1) is distributed as a standard normal under H_0; i.e.

$$T_n \overset{H_0}{\sim} N(0,1).$$

Consequently, a level α test (two-sided) would reject the null hypothesis when $|T_n| \geq \mathcal{Z}_{\alpha/2}$. One sided level α tests would reject whenever $T_n \geq \mathcal{Z}_\alpha$ or $-T_n \geq \mathcal{Z}_\alpha$ depending on the question. For example, if we were interested in showing that treatment 1 is better (longer survival times) than treatment 0, then we would reject H_0 when $-T_n \geq \mathcal{Z}_\alpha$ because under the alternative hypothesis we would expect the observed number of deaths from treatment 1 to be less than that expected under the null hypothesis.

Note: All the arguments made above were based on summarizing the data as 2 × 2 tables at distinct death times. Nowhere did we have to make any assumptions (other than the null hypothesis) about the actual shape of the underlying survival distribution in deriving the numerator of the logrank test statistic or its variance. This, intuitively, explains why this test is nonparametric.

If censored survival data are organized as $(U_i, \Delta_i, A_i), i = 1, \ldots, n$, where U_i denotes time to failure or censoring, Δ_i denotes failure indicator, and A_i denotes treatment indicator, then the logrank test can be computed using SAS. To illustrate, we again use the data from CALGB 8541 (clinical trial on beast cancer).

Recall that CALGB 8541 was a randomized three arm clinical trial for patients with stage II node positive breast cancer. Although there were three treatments, the major focus was comparing treatment 1 (Intensive CAF) to treatment 2 (Low dose CAF), where CAF is the combination of the drugs Cyclophosphamide, Adriamycin an 5 Fluorouracil. For the purpose of this illustration we will restrict attention to the comparison of these two treatments. Later we will discuss the comparison of all three treatments.

```
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

                                                      Pr >
                 Test      Chi-Square      DF    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

Thus far, we have only considered the properties of the logrank test under the null hypothesis. In order to assess the statistical sensitivity of this test, we must also consider the power to detect clinically meaningful alternatives. A useful way to define alternative hypotheses is through the proportional hazards assumption. That is, letting $\lambda_1(t)$ and $\lambda_0(t)$ denote the hazard functions at time t, for treatments 1 and 0 respectively, the proportional hazards assumption assumes that

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\gamma), \text{ for all } t \geq 0. \tag{9.2}$$

We use $\exp(\gamma)$ here because a hazard ratio must be positive and because $\gamma = 0$ will correspond to a hazard ratio of one which would imply that both treatments have the same hazard function (i.e. the null hypothesis). The proportional hazards assumption, if true, also has a nice interpretation. The hazard ratio $\exp(\gamma)$ can be viewed as a relative risk and 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)

If the proportional hazards assumption were true; that is,

$$\lambda_1(t) = \lambda_0(t)\exp(\gamma),$$

then this would imply that

$$-\frac{d\log S_1(t)}{dt} = -\frac{d\log S_0(t)}{dt}\exp(\gamma),$$

or

$$-\log S_1(t) = -\log S_0(t)\exp(\gamma).$$

Consequently,

$$S_1(t) = \{S_0(t)\}^{\exp(\gamma)},$$

and

$$\log\{-\log S_1(t)\} = \log\{-\log S_0(t)\} + \gamma.$$

This last relationship can be useful if we want to assess whether a proportional hazards assumption is a reasonable representation of the data. By plotting the two treatment-specific Kaplan-Meier curves on a $\log\{-\log\}$ scale we can visually inspect whether these two curves differ from each other by a constant over time.

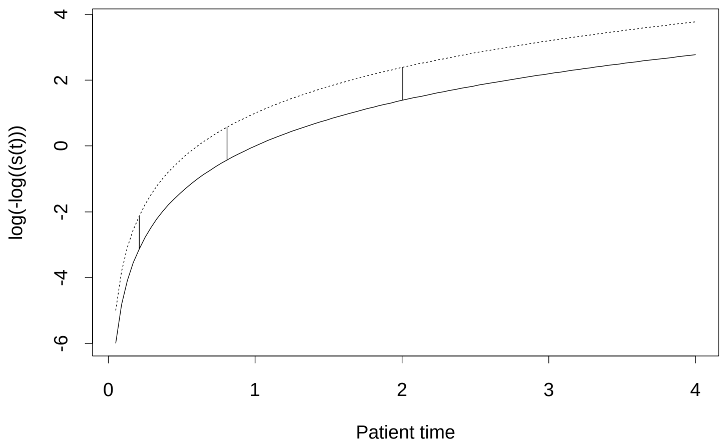
Also, in the special case where we feel comfortable in assuming that the survival distributions follow an exponential distribution; i.e. constant hazards, the proportional hazards assumption is guaranteed to hold. That is,

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \frac{\lambda_1}{\lambda_0}.$$

In section 9.1 we showed that the median survival time for an exponential distribution with hazard λ is equal to $m = \log(2)/\lambda$. Therefore, the ratio of the median survival times for two treatments whose survival distributions are exponentially distributed with hazard rates λ_1 and λ_0 is

$$\frac{m_1}{m_0} = \frac{\{\log(2)/\lambda_1\}}{\{\log(2)/\lambda_0\}} = \frac{\lambda_0}{\lambda_1}.$$

Figure 9.5: *Two survival functions with proportional hazards on log[-log] scale*



That is, the ratio of the medians of two exponentially distributed random variables is inversely proportional to the ratio of the hazards. This relationship may be useful when one is trying to illicit clinically important differences from medical collaborators during the design stage of an experiment. Clinical investigators generally have a good sense of the median survival for various treatments and can more easily relate to the question of determining an important increase in median survival. However, as we just illustrated, if the survival distributions are well approximated by exponential distributions then the differences in median survival can be easily translated to a hazard ratio through the inverse relationship derived above.

The reason we focus on proportional hazards alternatives is that, in addition to having “nice” interpretation, theory has been developed that shows that the logrank test is the most powerful nonparametric test to detect proportional hazards alternatives. Moreover, it has also been shown that the distribution of the logrank test under the alternative

$$H_A : \frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\gamma_A),\ t \geq 0$$

is approximately distributed as a normal distribution

$$T_n \overset{H_A}{\sim} N(\{d\theta(1-\theta)\}^{1/2}\gamma_A, 1),$$

where d denotes the total number of deaths (events), and θ 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 α 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

$$d = \frac{4(\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta)^2}{\gamma_A^2}. \tag{9.3}$$

This means that the power of the logrank test to detect a proportional hazards alternative is directly related to the number of events (deaths) d during the course of the clinical trial.

Remark: This very nice and simple relationship would not apply if we didn't use the logrank test in conjunction with a proportional hazards alternative.

If we take $\alpha = .05$ (two-sided), power $(1-\beta) = .90$, and $\theta = .5$, then

$$d = \frac{4(1.96 + 1.28)^2}{\gamma_A^2}. \tag{9.4}$$

Some examples of the number of deaths necessary to detect an alternative where the hazard ratio equals $\exp(\gamma_A)$ with 90% power using the logrank test at the .05 (two-sided) level of significance is given in the following table.

Hazard Ratio $\exp(\gamma_A)$	Number of deaths d
2.00	88
1.50	256
1.25	844
1.10	4623

During the design stage we must ensure that a sufficient number of patients are entered into the trial and followed long enough so that the requisite number of events are attained.

Sample Size Considerations

One straightforward approach to ensure the desired power of a trial to detect a clinically important alternative is to continue a clinical trial until we obtain the required number of failures.

Example: Suppose patients with advanced lung cancer historically have a median survival of six months and the survival distribution is approximately exponentially distributed. An increase in the median survival from six months to nine months is considered clinically important and we would like to have at least 90% power to detect such a difference if it can be achieved with a new therapy using a logrank test at the .05 (two-sided) level of significance. How should we design a clinical trial comparing the standard treatment to a promising new treatment?

If the survival distributions for both treatments were exponentially distributed, then the clinically important hazard ratio would be

$$\frac{\lambda_1}{\lambda_0} = \frac{m_0}{m_1} = \frac{6}{9} = 2/3.$$

Hence $\gamma_A = \log(2/3) = -.4055$. Using the formula given by equation (9.4) we determine that we need to carry out a study that will ultimately require a total of 256 deaths.

Since, for this example, patients do not survive for very long, one strategy is to enter some number of patients greater than 256 and continue the study until 256 deaths occur. For example, we may enter, say 350 patients, whom we randomize with equal probability to the two treatments and then analyze the data after 256 deaths.

Note The 350 patients was chosen arbitrarily. We now give more careful consideration to some of the design aspects of a clinical trial with a survival endpoint.

Design Specifications

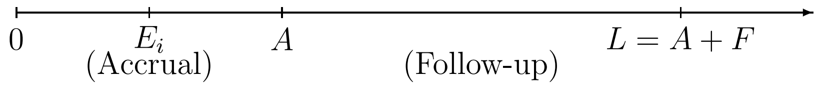
In most clinical trials, arbitrarily picking a number of patients and waiting for the requisite number of events to occur will not be adequate for the proper planning of the trial. More often we need to consider the following in the proper design of a clinical trial with a time to event endpoint.

- number of patients
- accrual period
- follow-up time

It was shown by Schoenfeld that to obtain reasonable approximations for the power, we need the expected number of events (deaths), computed under the alternative, to equal the number given by equation (9.3). Specifically, the expected number of deaths need to be computed separately for each treatment arm, under the assumption that the alternative hypothesis were true, and the sum from both treatments should equal (9.3) to achieve the desired power.

In order to compute the expected number of deaths for each treatment during the design stage, we must consider the following. Also, see figure below.

Figure 9.6: *Illustration of accrual and follow-up*



Remark: We will assume that any censoring that may occur is due to end of study censoring resulting from staggered entry. We will sometimes refer to this as "administrative censoring". If, in addition, we wanted to consider other forms of censoring, such as lost to follow-up due to withdrawal or censoring due to death from other causes, then the formulas given below would have to be modified.

We define the following notation:

- A denotes the accrual period; that is, the calendar period of time that patients are entering the study
- F denotes the calendar period of time after accrual has ended before the final analysis is conducted
- $L = A + F$ denotes the total calendar time of the study from the time the study opens until the final analysis

- Denote the accrual rate at calendar time u by $a(u)$; more precisely as

$$a(u) = \lim_{h \to 0} \left\{ \frac{\text{Expected no. of patients entering between } [u, u+h)}{h} \right\}.$$

The total expected number of patients in the clinical trial is given by

$$\int_0^A a(u) du.$$

If we have a constant accrual rate (this is the most common assumption made during the design stage), then $a(u) = a$ and the expected number of patients is aA.

In a randomized study where patients are assigned with equal probability to each of two treatments the accrual rate to each treatment would be $\frac{a(u)}{2}$. If we denote the distribution function of the survival time for each treatment (0 or 1) by

$$F_j(u) = P(T \leq u | A = j) = 1 - S_j(u), j = 0, 1,$$

then the expected number of observed deaths from treatment j is

$$d_j = \int_0^A \frac{a(u)}{2} F_j(L-u) du, j = 0, 1. \tag{9.5}$$

In words, we expect $\frac{a(u)}{2} du$ patients to enter between times $[u, u+du)$, of which the proportion $F_j(L-u)$ are expected to die by the end of the study (at time L). This number summed (integrated) over u, for values of u during the accrual period $[0, A]$, will yield the expected number of deaths on treatment $j = 0, 1$. To get the desired power we want

$$d_1 + d_0 = \frac{4(\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta)^2}{\gamma_A^2},$$

Note: The number of failures can be affected by

- the accrual rate
- accrual period (sample size)
- follow-up period
- the failure (hazard) rate (survival distribution)

Some, or all, of these factors can be controlled by the investigator and have to be considered during the design stage.

Example:

Assume the accrual rate is constant a patients per year and we randomize equally to two treatments so that the accrual rate is $a/2$ patients per year on each treatment. Also assume that the treatment specific survival distributions are exponentially distributed with hazards $\lambda_j, j = 0, 1$. Then

$$\begin{aligned} d_j &= \int_0^A \frac{a}{2}\left[1 - \exp\{-\lambda_j(L-u)\}\right] du \\ &= \frac{a}{2}\left[A - \frac{\exp(-\lambda_j L)}{\lambda_j}\{\exp(\lambda_j A) - 1\}\right]. \end{aligned}$$

During the design stage, we expect 100 patients per year to be recruited into the study. The median survival for treatment 0 is expected to be about 4 years and assumed to follow an exponential distribution ($\lambda_0 = .173$). We want the power to be 90% if the new treatment (treatment 1) can increase the median survival to 6 years ($\lambda_1 = .116$) using a logrank test at the .05 (two-sided) level of significance.

For this problem the clinically important hazard ratio is 2/3 corresponding to $\gamma_A = \log(2/3)$. Using equation (9.3), the total number of deaths necessary is 256. Hence, we need

$$d_1(A, L) + d_0(A, L) = 256.$$

Note: We use the notation $d_1(A, L)$ to emphasize the fact that the number of deaths will depend on our choice of accrual period A and length of study L.

According to the formulas we derived, we need A and L to satisfy the equation

$$\begin{gathered} 50\left[A - \frac{\exp(-.116L)}{.116}\{\exp(.116A) - 1\}\right] \\ +50\left[A - \frac{\exp(-.173L)}{.173}\{\exp(.173A) - 1\}\right] \\ = 256. \end{gathered}$$

There are many combinations of A and L that would satisfy the equation above. Clearly, the minimum length of the study would occur if we continued accruing patients continuously until we obtained the desired number of events. This corresponds to the case where $A = L$. When $A = L$ there is only one unknown above and solving the equation above yields $A = L = 7$ years. Such a design would require a total of 700 patients.

Note: This equation does not have a closed form solution and the solution must be obtained numerically by iterative techniques.

In contrast, if we accrued for five years, $A = 5$, or a total of 500 patients, then solving for L yields $L = 7.65$.

Clearly, to obtain 256 deaths, we need to accrue at least 256 patients. Therefore, A must be greater than 2.56 years. However, choosing L that small in order to accrue as few patients as possible would result in a trial that would take an exceedingly long time to complete.

A good strategy is to experiment with a variety of combinations of A and L that can be used to present to the clinical investigators and then the choice which best suits the needs of the study can be made.

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 Tests

We now consider the case where we randomize to $K > 2$ treatments and are interested in testing the null hypothesis that the survival distributions are the same for all treatments versus the alternative that there is some difference. With right censoring, the data from such a clinical trial can be represented as realizations of the iid triplets $(U_i, \Delta_i, A_i), i = 1, \ldots, n$, where for the i-th individual, $U_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, and $A_i = (1, \ldots, K)$ denotes the treatment group that the individual was randomized to. Letting $S_j(t) = P(T \geq t | A = j), j = 1, \ldots, K$ denote the treatment-specific survival distributions, the null hypothesis of no treatment difference is

$$H_0 : S_1(t) = \ldots = S_K(t),\ t \geq 0,$$

or equivalently

$$H_0 : \lambda_1(t) = \ldots = \lambda_K(t),\ t \geq 0,$$

where $\lambda_j(t), j = 1, \ldots, K$ denote the treatment-specific hazard rates.

The test of the null hypothesis for this K-sample problem will be a direct generalization of the logrank test used for the two-sample test. Using the same notation as in the previous section where $n_j(u)$ denotes the number of individuals at risk at time u in treatment group j and $d_j(u)$ the number of observed deaths in treatment group j at time u, we view the data as a series of $2 \times K$ tables at the distinct failure times u where $d(u) \geq 1$. For example,

	Treatments				
	1	2	...	K	Total
No. of deaths	$d_1(u)$	$d_2(u)$	...	$d_K(u)$	$d(u)$
No. alive	$n_1(u) - d_1(u)$	$n_2(u) - d_2(u)$	...	$n_K(u) - d_K(u)$	$n(u) - d(u)$
No. at risk	$n_1(u)$	$n_2(u)$	...	$n_K(u)$	$n(u)$

Generalization of the two-sample test

At each time u where $d(u) \geq 1$, we now consider a vector of observed minus expected number of deaths under the null hypothesis for each treatment group j. Namely,

$$\begin{pmatrix} d_1(u) - \frac{n_1(u)}{n(u)} d(u) \\ \cdot \\ \cdot \\ \cdot \\ d_K(u) - \frac{n_K(u)}{n(u)} d(u) \end{pmatrix}^{K \times 1}$$

Note: The sum of the elements in this vector is equal to zero. Because of this, such vectors can only vary in $(K-1)$-dimensional space. Consequently, it suffices to use only $K - 1$ of these vectors for further consideration in constructing test statistics. It doesn't matter which $K-1$ of these we use, thus, for convention, we will take from now on $j = 1, \ldots, K-1$.

Analogous to the two-sample problem, if we condition on the marginal counts of these $2 \times K$ contingency tables, then, under the null hypothesis, the distribution of the counts $\{d_1(u), \ldots, d_K(u)\}$ is a multivariate version of the hypergeometric distribution. Specifically, this implies that, conditional on the marginal counts, the conditional expectation of $d_j(u)$ is

$$E_C\{d_j(u)\} = \frac{n_j(u)}{n(u)} d(u), j = 1, \ldots, K.$$

Note: We use the notation $E_C(\cdot)$ to denote conditional expectation.

Also the conditional variance-covariance matrix of the $d_j(u)$'s is given by

$$var_C\{d_j(u)\} = \frac{d(u)\{n(u) - d(u)\} n_j(u)\{n(u) - n_j(u)\}}{n^2(u)\{n(u)-1\}}, j = 1, \ldots, K,$$

and for $j \neq j'$

$$cov_C\{d_j(u), d_{j'}(u)\} = -\left[\frac{d(u)\{n(u) - d(u)\} n_j(u) n_{j'}(u)}{n^2(u)\{n(u)-1\}}\right].$$

Again, analogous to the construction of the logrank test in two-samples, consider the $K-1$ dimensional vector $\mathcal{T}_n$, made up by the sum of the observed minus expected counts for treatments $j = 1, \ldots, K-1$, summed over distinct death times u such that $d(u) \geq 1$. That is

$$\mathcal{T}_n = \begin{pmatrix} \sum_{\text{death times } u} \left\{d_1(u) - \frac{n_1(u)}{n(u)} d(u)\right\} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{\text{death times } u} \left\{d_{K-1}(u) - \frac{n_{K-1}(u)}{n(u)} d(u)\right\} \end{pmatrix}^{(K-1)\times 1}$$

The corresponding $(K-1) \times (K-1)$ covariance matrix of the vector $\mathcal{T}_n$ is given by $\mathcal{V}_n^{(K-1)\times(K-1)}$, where for $j = 1, \ldots, K-1$ the diagonal terms are given by

$$\mathcal{V}_{njj} = \sum_{\text{death times } u} \left[\frac{d(u)\{n(u) - d(u)\} n_j(u)\{n(u) - n_j(u)\}}{n^2(u)\{n(u)-1\}}\right],$$

and for $j \neq j'$, the off-diagonal terms are

$$\mathcal{V}_{njj'} = -\sum_{\text{death times } u} \left[\frac{d(u)\{n(u) - d(u)\} n_j(u) n_{j'}(u)}{n^2(u)\{n(u)-1\}}\right].$$

The test statistic used to test the null hypothesis is given by the quadratic form

$$T_n = \mathcal{T}_n^T [\mathcal{V}_n]^{-1} \mathcal{T}_n,$$

where we use the notation $(\cdot)^T$ to denote the transpose of a vector. Under H_0, the statistic T_n is distributed asymptotically as a central chi-square distribution with $K-1$ degrees of freedom. This test statistic is called the K-sample logrank test statistic. If the null hypothesis is true, we would expect the elements in the vector $\mathcal{T}_n$ to be near zero. Hence the quadratic form, T_n, should also be near zero. If however, there were treatment differences, then we would expect some of the elements in the vector $\mathcal{T}_n$ to deviate from zero and thus expect the quadratic form, T_n, to be larger. Thus, we will reject the null hypothesis if T_n is sufficiently large. For a level α test, we would reject H_0 when

$$T_n = \mathcal{T}_n^T [\mathcal{V}_n]^{-1} \mathcal{T}_n \geq \chi^2_{\alpha, K-1}.$$

Proc lifetest in SAS implements this test and we will illustrate shortly using the data from CALGB 8541. But first let us end this section with a short discussion on how sample size calculations can be implemented during the design stage when one is comparing the survival distribution of more than two treatments with the logrank test.

9.7 Sample-size considerations for the K-sample logrank test

The computations of the sample size necessary to attain the desired power to detect clinically important treatment differences using the logrank test generalize from the two-sample problem to the K-sample problem in a manner similar to that considered for the comparison of proportions or means discussed in Chapter 7.

Specifically, if we randomize with equal probability to K treatments, then the total number of deaths necessary for the K-sample logrank test at the α level of significance to have power at least $1-\beta$ to detect a hazard ratio (assumed constant over time) between any two treatments greater than equal to $\exp(\gamma_A)$ is given by

$$d = \frac{2K\phi^2(\alpha, \beta, K-1)}{\gamma_A^2},$$

where

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

is the value of the non-centrality parameter necessary so that a non-central chi-square distributed random variable with $K-1$ degrees of freedom and non-centrality parameter $\phi^2(\alpha, \beta, K-1)$ will exceed the value $\chi^2_{\alpha;K-1}$ with probability $(1-\beta)$. Tables of $\phi^2(\alpha, \beta, K-1)$ for $\alpha = .05$ were provided in chapter 7.

For example, if we take $K = 3$, then in order to ensure that we have at least 90% power to detect a hazard ratio between any two treatments that may exceed 1.5, using a logrank test at the .05 level of significance, we would need the total number of deaths to exceed

$$d = \frac{2 \times 3 \times 12.654}{\{\log(1.5)\}^2} = 462.$$

We can contrast this to a two-sample comparison which needs 256 events. As in the two-sample problem, the computations during the design stage will involve the best guesses for the accrual rate, accrual period, follow-up period, and underlying treatment-specific survival distributions which can be translated to the desired number of failures. Thus we can experiment with different values of

- accrual rate $a(u)$
- underlying treatment-specific failure time distributions $F_j(t) = P(T \leq t|A = j) = 1 - S_j(t), j = 1, \ldots, K$ under the alternative hypothesis of interest (we may take these at the least favorable configuration)
- the accrual period A
- the length of study L

so that

$$\sum_{j=1}^{K} d_j\{a(\cdot), F_j(\cdot), A, L\} = \frac{2K\phi^2(\alpha, \beta, K-1)}{\gamma_A^2},$$

where $d_j\{a(\cdot), F_j(\cdot), A, L\}$ denotes the expected number of deaths in treatment group j as a function of $a(\cdot)$, $F_j(\cdot)$, A, L, computed using equation (9.5).

CALGB 8541 Example

We now return to the data from CALGB 8541 which compared three treatments in a randomized study of node positive stage II breast cancer patients. The three treatments were

- 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. As well as testing for overall differences in the three treatments, we shall also look at the three pairwise comparisons.

`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

                                                 Pr >
          Test      Chi-Square      DF    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

Up to now we have considered the design and analysis of clinical trials assuming the data would be analyzed at only one point in time; i.e. the final analysis. However, due to ethical as well as practical considerations, the data are monitored periodically during the course of the study and for a variety of reasons may be stopped early. In this section we will discuss some of the statistical issues in the design and analysis of clinical trials which allow the possibility of early stopping. These methods fall under the general title of **group-sequential methods**.

Some reasons a clinical trial may be stopped early include

- Serious toxicity or adverse events
- Established benefit
- No trend of interest
- Design or logistical difficulties too serious to fix

Since there is lot invested (scientifically, emotionally, financially, etc.) in a clinical trial by the investigators who designed or are conducting the trial, they may not be the best suited for deciding whether the clinical trial should be stopped. It has become common practice for most large scale phase III clinical trials to be monitored by an independent data monitoring committee; often referred to as a Data Safety Monitoring Board (DSMB). It is the responsibility of this board to monitor the data from a study periodically (usually two to three times a year) and make recommendations on whether the study should be modified or stopped. The primary responsibility of this board is to ensure the safety and well being of the patients that have enrolled into the trial.

The DSMB generally has members who represent the following disciplines:

- Clinical
- Laboratory
- Epidemiology
- Biostatistics
- Data Management
- Ethics

The members of the DSMB should have no conflict of interest with the study or studies they are monitoring; e.g. no financial holdings in the company that is developing the treatments by member or family. All the discussions of the DSMB are confidential. The charge of the DSMB includes:

- 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

Later in the study, as the data mature, the analyses focus on treatment comparisons. One of the important issues in deciding whether a study should be stopped early is whether a treatment difference during an interim analysis is sufficiently large or small to warrant early termination. **Group-sequential** methods are rules for stopping a study early based on treatment differences that are observed during interim analyses. The term group-sequential refers to the fact that the data are monitored sequentially at a finite number of times (calendar) where a group of new data are collected between the interim monitoring times. Depending on the type of study, the new data may come from new patients entering the study or additional information from patients already in the study or a combination of both. In this chapter we will study statistical issues in the design and analysis of such group-sequential methods. We will take a general approach to this problem that can be applied to many different clinical trials. This approach is referred to as **Information-based design and monitoring of clinical trials**.

The typical scenario where these methods can be applied is as follows:

- A study in which data are collected over calendar time, either data from new patients entering the study or new data collected on patients already in the study
- Where the interest is in using these data to answer a research question. Often, this is posed as a decision problem using a hypothesis testing framework. For example, testing whether a new treatment is better than a standard treatment or not.
- The investigators or "the monitoring board" monitor the data periodically and conduct interim analyses to assess whether there is sufficient "strong evidence" in support of the research hypothesis to warrant early termination of the study

- At each monitoring time, a test statistic is computed and compared to a stopping boundary. The stopping boundary is a prespecified value computed at each time an interim analysis may be conducted which, if exceeded by the test statistic, can be used as sufficient evidence to stop the study. Generally, a test statistic is computed so that its distribution is well approximated by a normal distribution. (This has certainly been the case for all the statistics considered in the course)
- 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 problem where

- t-tests are used to compare the mean of continuous random variables between treatments
- proportions test 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 tests using maximum likelihood estimators

10.2 Information based design and monitoring

The underlying structure that is assumed here is that the data are generated from a probability model with population parameters Δ, θ, where Δ denotes the parameter of primary interest, in our case, this will generally be treatment difference, and θ denote the nuisance parameters. We will focus primarily on two-sided tests where we are testing the null hypothesis

$$H_0 : \Delta = 0$$

versus the alternative

$$H_A : \Delta \neq 0,$$

however, the methods are general enough to also consider one-sided tests where we test

$$H_0 : \Delta \leq 0$$

versus

$$H_A : \Delta > 0.$$

Remark This is the same framework that has been used throughout the course.

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 Δ 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.

Example 1. (Dichotomous response)

Let π_1, π_0 denote the population response rates for treatments 1 and 0 (say new treatment and control) respectively. Let the treatment difference be given by

$$\Delta = \pi_1 - \pi_0$$

The test of the null hypothesis 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 $j = 0, 1$.

Example 2. (Time to event)

Suppose we assume a proportional hazards model. Letting A denote treatment indicator, we consider the model

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(-\Delta),$$

and we want to test the null hypothesis of no treatment difference

$$H_0 : \Delta = 0$$

versus the two-sided alternative

$$H_A : \Delta \neq 0,$$

or the one-sided test that treatment 1 does not improve survival

$$H_0 : \Delta \leq 0$$

versus the alternative that it does improve survival

$$H_A : \Delta > 0.$$

Using all the survival data up to time t (some failures and some censored observations), 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 Δ that was derived by D. R. Cox and $se\{\hat{\Delta}(t)\}$ is the corresponding standard error. 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 $T(t)$.

Example 3. (Parametric models)

Any parametric model where we assume the underlying density of the data is given by $p(z; \Delta, \theta)$, and use for $\hat{\Delta}(t)$ the maximum likelihood estimator for Δ and for $se\{\hat{\Delta}(t)\}$ compute the estimated standard error using the square-root of the inverse of the observed information matrix, with the data up to time t.

In most important applications the test statistic has the property that the distribution when $\Delta = \Delta^*$ follows a normal distribution, namely,

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}} \overset{\Delta=\Delta^*}{\sim} N(\Delta^* I^{1/2}(t, \Delta^*), 1),$$

where $I(t, \Delta^*)$ denotes the statistical information at time t. Statistical information refers to Fisher information, but for those not familiar with these ideas, for all practical purposes, we can equate (at least approximately) information with the standard error of the estimator; namely,

$$I(t, \Delta^*) \approx \{se(\hat{\Delta}(t)\}^{-2}.$$

Under the null hypothesis $\Delta = 0$, the distribution follows a standard normal, that is

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

and is used a basis for a group-sequential test. For instance, if we considered a two-sided test, we would reject the null hypothesis whenever

$$|T(t)| \geq b(t),$$

where $b(t)$ is some critical value or what we call a **boundary** value. If we only conducted one analysis and wanted to construct a test at the α level of significance, then we would choose $b(t) = \mathcal{Z}_{\alpha/2}$. Since under the null hypothesis the distribution of $T(t)$ is $N(0,1)$ then

$$P_{H_0}\{|T(t)| \geq \mathcal{Z}_{\alpha/2}\} = \alpha.$$

If, however, the data were monitored at K different times, say, $t_1, \ldots, t_K$, then we would want to have the opportunity to reject the null hypothesis if the test statistic, computed at any of these times, was sufficiently large. That is, we would want to reject H_0 at the first time $t_j, j = 1, \ldots, K$ such that

$$|T(t_j)| \geq b(t_j),$$

for some properly chosen set of boundary values $b(t_1), \ldots, b(t_K)$.

Note: In terms of probabilistic notation, using this strategy, rejecting the null hypothesis corresponds to the event

$$\bigcup_{j=1}^{K} \{|T(t_j)| \geq b(t_j)\}.$$

Similarly, accepting the null hypothesis corresponds to the event

$$\bigcap_{j=1}^{K} \{|T(t_j)| < b(t_j)\}.$$

The crucial issue is how large should the treatment differences be during the interim analyses before we reject H_0; that is, how do we choose the boundary values $b(t_1), \ldots, b(t_K)$? Moreover, what are the consequences of such a strategy of sequential testing on the level and power of the test and how does this affect sample size calculations?

10.3 Type I error

We start by first considering the effect that group-sequential testing has on type I error; i.e. on the level of the test. Some thought must be given on the choice of boundary values. For example, since we have constructed the test statistic $T(t_j)$ to have approximately a standard normal distribution, if the null hypothesis were true, at each time t_j, if we naively reject H_0 at the first monitoring time that the absolute value of the test statistic exceeds 1.96 (nominal p-value of .05), then the type I error will be inflated due to multiple comparisons. That is,

$$\text{type I error} = P_{H_0}[\bigcup_{j=1}^{K}\{|T(t_j)| \geq 1.96\}] > .05,$$

if $K \geq 2$.

This is illustrated in the following table:

Table 10.1: Effect of multiple looks on type I error

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
∞	1.000

The last entry in this table was described by J. Cornfield as

"Sampling to a foregone conclusion".

Our first objective is to derive group-sequential tests (i.e. choice of boundaries $b(t_1), \dots, b(t_K)$) that have the desired prespecified type I error of α.

Level α test

We want the probability of rejecting H_0, when H_0 is true, to be equal to α, say .05. The strategy for rejecting or accepting H_0 is as follows:

- Stop and reject H_0 at the first interim analysis if

$$\{|T(t_1)| \geq b(t_1)\}$$

- or stop and reject H_0 at the second interim analysis if

$$\{|T(t_1)| < b(t_1), |T(t_2)| \geq b(t_2)\}$$

- or …
- or stop and reject at the final analysis if

$$\{|T(t_1)| < b(t_1), \dots, |T(t_{K-1})| < b(t_{K-1}), |T(t_K)| \geq b(t_K)\}$$

- otherwise, accept H_0 if

$$\{|T(t_1)| < b(t_1), \dots, |T(t_K)| < b(t_K)\}.$$

This representation partitions the sample space into mutually exclusive rejection regions and an acceptance region. In order that our testing procedure have level α, the boundary values $b(t_1), \dots, b(t_K)$ must satisfy

$$P_{\Delta=0}\{|T(t_1)| < b(t_1), \dots, |T(t_K)| < b(t_K)\} = 1 - \alpha. \tag{10.1}$$

Remark:

By construction, the test statistic $T(t_j)$ will be approximately distributed as a standard normal, if the null hypothesis is true, at each time point t_j. However, to ensure that the group-sequential test have level α, the equality given by (10.1) must be satisfied. The probability on the left hand side of equation (10.1) involves the joint distribution of the multivariate statistic $(T(t_1), \ldots, T(t_K))$. Therefore, we would need to know the joint distribution of this sequentially computed test statistic at times $t_1, \ldots, t_K$, under the null hypothesis, in order to compute the necessary probabilities to ensure a level α test. Similarly, we would need to know the joint distribution of this multivariate statistic, under the alternative hypothesis to compute the power of such a group-sequential test.

The major result which allows the use of a general methodology for monitoring clinical trials is that

"Any efficient based test or estimator for Δ, properly normalized, when computed sequentially over time, has, asymptotically, a normal independent increments process whose distribution depends only on the parameter Δ and the statistical information."

Scharfstein, Tsiatis and Robins (1997). JASA. 1342-1350.

As we mentioned earlier the test statistics are constructed so that when $\Delta = \Delta^*$

$$T(t) = \frac{\hat{\Delta}(t)}{se\{\hat{\Delta}(t)\}} \sim N(\Delta^* I^{1/2}(t, \Delta^*), 1),$$

where we can approximate statistical information $I(t, \Delta^*)$ by $[se\{\hat{\Delta}(t)\}]^{-2}$. If we normalize by multiplying the test statistic by the square-root of the information; i.e.

$$W(t) = I^{1/2}(t, \Delta^*) T(t),$$

then this normalized statistic, computed sequentially over time, will have the normal independent increments structure alluded to earlier. Specifically, if we compute the statistic at times $t_1 < t_2 < \ldots < t_K$, then the joint distribution of the multivariate vector $\{W(t_1), \ldots, W(t_K)\}$ is asymptotically normal with mean vector $\{\Delta^* I(t_1, \Delta^*), \ldots, \Delta^* I(t_K, \Delta^*)\}$ and covariance matrix where

$$var\{W(t_j)\} = I(t_j, \Delta^*), \; j = 1, \ldots, K$$

and

$$cov[W(t_j), \{W(t_\ell) - W(t_j)\}] = 0, \; j < \ell, j, \ell = 1, \ldots, K.$$

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

$$\begin{aligned}
W(t_1) &= W(t_1)\\
W(t_2) &= W(t_1)+\{W(t_2)-W(t_1)\}\\
&\cdot\\
&\cdot\\
&\cdot\\
W(t_j) &= W(t_1)+\{W(t_2)-W(t_1)\}+\ldots+\{W(t_j)-W(t_{j-1})\}
\end{aligned}$$

has the same distribution as a partial sum of independent normal random variables

This structure implies that the covariance matrix of $\{W(t_1),\ldots,W(t_K)\}$ is given by

$$var\{W(t_j)\}=I(t_j,\Delta^*),\ j=1,\ldots,K$$

and for $j<\ell$

$$\begin{aligned}
& cov\{W(t_j),W(t_\ell)\}\\
={}& cov[W(t_j),\{W(t_\ell)-W(t_j)\}+W(t_j)]\\
={}& cov[W(t_j),\{W(t_\ell)-W(t_j)\}]+cov\{W(t_j),W(t_j)\}\\
={}& 0+var\{W(t_j)\}\\
={}& I(t_j,\Delta^*).
\end{aligned}$$

Since the test statistic

$$T(t_j)=I^{-1/2}(t_j,\Delta^*)W(t_j),\ j=1,\ldots,K$$

this implies that the joint distribution of $\{T(t_1),\ldots,T(t_K)\}$ is also multivariate normal where the mean

$$E\{T(t_j)\}=\Delta^* I^{1/2}(t_j,\Delta^*),\ j=1,\ldots,K \tag{10.2}$$

and the covariance matrix is such that

$$var\{T(t_j)\}=1,\ j=1,\ldots,K \tag{10.3}$$

and for $j < \ell$, the covariances are

$$\begin{aligned} cov\{T(t_j), T(t_\ell)\} &= cov\{I^{-1/2}(t_j, \Delta^*)W(t_j), I^{-1/2}(t_\ell, \Delta^*)W(t_\ell)\} \\ &= I^{-1/2}(t_j, \Delta^*)I^{-1/2}(t_\ell, \Delta^*)cov\{W(t_j), W(t_\ell)\} \\ &= I^{-1/2}(t_j, \Delta^*)I^{-1/2}(t_\ell, \Delta^*)I(t_j, \Delta^*) \\ &= \frac{I^{1/2}(t_j, \Delta^*)}{I^{1/2}(t_\ell, \Delta^*)} = \sqrt{\frac{I(t_j, \Delta^*)}{I(t_\ell, \Delta^*)}}. \end{aligned} \tag{10.4}$$

In words, the covariance of $T(t_j)$ and $T(t_\ell)$ is the square-root of the relative information at times t_j and t_ℓ. Hence, under the null hypothesis $\Delta = 0$, the sequentially computed test statistic $\{T(t_1), \ldots, T(t_K)\}$ is multivariate normal with mean vector zero and covariance matrix (in this case the same as the correlation matrix, since the variances are all equal to one)

$$V_T = \left[\sqrt{\frac{I(t_j, 0)}{I(t_\ell, 0)}}\right], \; j \le \ell. \tag{10.5}$$

The importance of this result is that the joint distribution of the sequentially computed test statistic, under the null hypothesis, is completely determined by the relative proportion of information at each of the monitoring times $t_1, \ldots, t_K$. This then allows us to evaluate probabilities such as those in equation (10.1) that are necessary to find appropriate boundary values $b(t_1), \ldots, b(t_K)$ that achieve the desired type I error of α.

10.3.1 Equal increments of information

Let us consider the important special case where the test statistic is computed after equal increments of information; that is

$$I(t_1, \cdot) = I, \; I(T_2, \cdot) = 2I, \; \ldots, \; I(t_K, \cdot) = KI.$$

Remark: For problems where the response of interest is instantaneous, whether this response be discrete or continuous, the information is proportional to the number of individuals under study. In such cases, calculating the test statistic after equal increments of information is equivalent to calculating the statistic after equal number of patients have entered the study. So, for instance, if we planned to accrue a total of 100 patients with the goal of comparing the response rate between two treatments, we may wish to monitor five times after equal increments of information; i.e after every 20 patients enter the study.

In contrast, if we were comparing the survival distributions with possibly right censored data, then it turns out that information is directly proportional to the number of deaths. Thus, for such a study, monitoring after equal increments of information would correspond to conducting interim analyses after equal number of observed deaths.

In any case, monitoring a study K times after equal increments of information imposes a very specific distributional structure, under the null hypothesis, for the sequentially computed test statistic that can be exploited in constructing group-sequential tests. Because $I(t_j, 0) = jI,\ j = 1, \ldots, K$, this means that the joint distribution of the sequentially computed test statistic $\{T(t_1), \ldots, T(t_K)\}$ is a multivariate normal with mean vector equal to zero and by (10.5) with a covariance matrix equal to

$$V_T = \left[\sqrt{\frac{I(t_j, 0)}{I(t_\ell, 0)}} = \sqrt{\frac{j}{\ell}} \right],\ j \leq \ell. \tag{10.6}$$

This means that under the null hypothesis the joint distribution of the sequentially computed test statistic computed after equal increments of information is completely determined once we know the total number K of interim analyses that are intended. Now, we are in a position to compute probabilities such as

$$P_{\Delta=0}\{|T(t_1)| < b_1, \ldots, |T(t_K)| < b_K\}$$

in order to determine boundary values $b_1, \ldots, b_K$ where the probability above equals $1 - \alpha$ as would be necessary for a level-α group-sequential test.

Remark: The computations necessary to compute such integrals of multivariate normals with the covariance structure (10.6) can be done quickly using using recursive numerical integration that was first described by Armitage, McPherson and Rowe (1969). This method takes advantage of the fact that the joint distribution is that of a standardized partial sum of independent normal random variables. This integration allows us to search for different combinations of $b_1, \ldots, b_K$ which satisfy

$$P_{\Delta=0}\{|T(t_1)| < b_1, \ldots, |T(t_K)| < b_K\} = 1 - \alpha.$$

There are infinite combinations of such boundary values that lead to level-α tests; thus, we need to assess the statistical consequences of these different combinations to aid us in making choices in which to use.

10.4 Choice of boundaries

Let us consider the flexible class of boundaries proposed by Wang and Tsiatis (1987) *Biometrics.* For the time being we will restrict attention to group-sequential tests computed after equal increments of information and later discuss how this can be generalized. The boundaries by Wang and Tsiatis were characterized by a power function which we will denote by Φ. Specifically, we will consider boundaries where

$$b_j = (\text{constant}) \times j^{(\Phi - .5)}, .$$

Different values of Φ will characterize different shapes of boundaries over time. We will also refer to Φ as the shape parameter.

For any value Φ, we can numerically derive the the constant necessary to obtain a level-α test. Namely, we can solve for the value c such that

$$P_{\Delta=0}\{\bigcap_{j=1}^{K} |T(t_j)| < cj^{(\Phi - .5)}\} = 1 - \alpha.$$

Recall. Under the null hypothesis, the joint distribution of $\{T(t_1), \ldots, T(t_K)\}$ is completely known if the times $t_1, \ldots, t_K$ are chosen at equal increments of information. The above integral is computed for different c until we solve the above equation. That a solution exists follows from the monotone relationship of the above probability as a function of c. The resulting solution will be denoted by $c(\alpha, K, \Phi)$. Some of these are given in the following table.

Table 10.2: *Group-sequential boundaries for two-sided tests for selected values of α, K, Φ*

	K			
Φ	2	3	4	5
		$\alpha = .05$		
0.0	2.7967	3.4712	4.0486	4.5618
0.1	2.6316	3.1444	3.5693	3.9374
0.2	2.4879	2.8639	3.1647	3.4175
0.3	2.3653	2.6300	2.8312	2.9945
0.4	2.2636	2.4400	2.5652	2.6628
0.5	2.1784	2.2896	2.3616	2.4135
		$\alpha = .01$		
0.0	3.6494	4.4957	5.2189	5.8672
0.1	3.4149	4.0506	4.5771	5.0308
0.2	3.2071	3.6633	4.0276	4.3372
0.3	3.0296	3.3355	3.5706	3.7634
0.4	2.8848	3.0718	3.2071	3.3137
0.5	2.7728	2.8738	2.9395	2.9869

Examples: Two boundaries that have been discussed extensively in the literature and have been used in many instances are special cases of the class of boundaries considered above. These are when $\Phi = .5$ and $\Phi = 0$. The first boundary when $\Phi = .5$ is the Pocock boundary; Pocock (1977) *Biometrika* and the other when $\Phi = 0$ is the O'Brien-Fleming boundary; O'Brien and Fleming (1979) *Biometrics*.

10.4.1 Pocock boundaries

The group-sequential test using the Pocock boundaries rejects the null hypothesis at the first interim analysis time $t_j, j = 1, \ldots, K$ (remember equal increments of information) that

$$|T(t_j)| \geq c(\alpha, K, 0.5),\ j = 1, \ldots, K.$$

That is, the null hypothesis is rejected at the first time that the standardized test statistic using all the accumulated data exceeds some constant.

For example, if we take $K = 5$ and $\alpha = .05$, then according to Table 10.2 $c(.05, 5, 0.5) = 2.41$. Therefore, the .05 level test which will be computed a maximum of 5 times after equal increments of information will reject the null hypothesis at the first time that the standardized test statistic exceeds 2.41; that is, we reject the null hypothesis at the first time $t_j,\ j = 1, \ldots, 5$ when $|T(t_j)| \geq 2.41$. This is also equivalent to rejecting the null hypothesis at the first time $t_j,\ j = 1, \ldots, 5$ that the nominal p-value is less than .0158.

10.4.2 O'Brien-Fleming boundaries

The O'Brien-Fleming boundaries have a shape parameter $\Phi = 0$. A group-sequential test using the O'Brien-Fleming boundaries will reject the null hypothesis at the first time $t_j,\ j = 1, \ldots, K$ when

$$|T(t_j)| \geq c(\alpha, K, 0.0)/\sqrt{j},\ j = 1, \ldots, K.$$

For example, if we again choose $K = 5$ and $\alpha = .05$, then according to Table 10.2 $c(.05, 5, 0.0) = 4.56$. Therefore, using the O'Brien-Fleming boundaries we would reject at the first time $t_j,\ j = 1, \ldots, 5$ when

$$|T(t_j)| \geq 4.56/\sqrt{j},\ j = 1, \ldots, 5.$$

Therefore, the five boundary values in this case would be $b_1 = 4.56$, $b_2 = 3.22$, $b_3 = 2.63$, $b_4 = 2.28$, and $b_5 = 2.04$

The nominal p-values for the O'Brien-Fleming boundaries are contrasted to those from the Pocock boundaries, for $K = 5$ and $\alpha = .05$ in the following table.

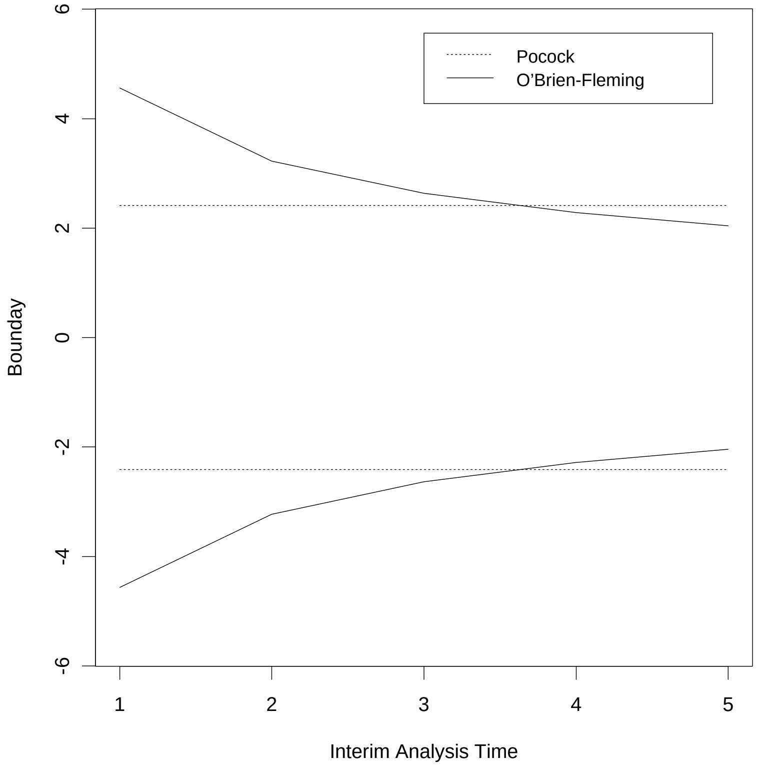
Table 10.3: *Nominal p-values for $K = 5$ and $\alpha = .05$*

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

The shape of these boundaries are also contrasted in the following figure.

Figure 10.1: *Pocock and O'Brien-Fleming Boundaries*



10.5 Power and sample size in terms of information

We have discussed the construction of group-sequential tests that have a pre-specified level of significance α. We also need to consider the effect that group-sequential tests have on power and its implications on sample size. To set the stage, we first review how power and sample size are determined with a single analysis using information based criteria.

As shown earlier, the distribution of the test statistic computed at a specific time t; namely $T(t)$, under the null hypothesis, is

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

and for a clinically important alternative, say $\Delta = \Delta_A$ is

$$T(t) \overset{\Delta=\Delta_A}{\sim} N(\Delta_A I^{1/2}(t,\Delta_A),1),$$

where $I(t,\Delta_A)$ denotes statistical information which can be approximated by $[se\{\hat{\Delta}(t)\}]^{-2}$, and $\Delta_A I^{1/2}(t,\Delta_A)$ is the noncentrality parameter. In order that a two-sided level-α test have power $1-\beta$ to detect the clinically important alternative Δ_A, we need the noncentrality parameter

$$\Delta_A I^{1/2}(t,\Delta_A) = \mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta},$$

or

$$I(t,\Delta_A) = \left\{\frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A}\right\}^2. \tag{10.7}$$

From this relationship we see that the power of the test is directly dependent on **statistical information**. Since information is approximated by $[se\{\hat{\Delta}(t)\}]^{-2}$, this means that the study should collect enough data to ensure that

$$[se\{\hat{\Delta}(t)\}]^{-2} = \left\{\frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A}\right\}^2.$$

Therefore one strategy that would guarantee the desired power to detect a clinically important difference is to monitor the standard error of the estimated difference through time t as data were being collected and to conduct the one and only final analysis at time t^F where

$$[se\{\hat{\Delta}(t^F)\}]^{-2} = \left\{\frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A}\right\}^2$$

using the test which rejects the null hypothesis when

$$|T(t^F)| \geq \mathcal{Z}_{\alpha/2}.$$

Remark Notice that we didn't have to specify any of the nuisance parameters with this information-based approach. The accuracy of this method to achieve power depends on how good the approximation of the distribution of the test statistic is to a normal distribution and how good the approximation of $[se\{\hat{\Delta}(t^F)\}]^{-2}$ is to the Fisher information. Some preliminary numerical simulations have shown that this information-based approach works very well if the sample sizes are sufficiently large as would be expected in phase III clinical trials.

In actuality, we cannot launch into a study and tell the investigators to keep collecting data until the standard error of the estimated treatment difference is sufficiently small (information large) without giving them some idea how many resources they need (i.e. sample size, length of study, etc.). Generally, during the design stage, we posit some guesses of the nuisance parameters and then use these guesses to come up with some initial design.

For example, if we were comparing the response rate between two treatments, say treatment 1 and treatment 0, and were interested in the treatment difference $\pi_1 - \pi_0$, where π_j denotes the population response probability for treatment $j = 0, 1$, then, at time t, we would estimate the treatment difference using $\hat{\Delta}(t) = p_1(t) - p_0(t)$, where $p_j(t)$ denotes the sample proportion that respond to treatment j among the individuals assigned to treatment j by time t for $j = 0, 1$. The standard error of $\hat{\Delta}(t) = p_1(t) - p_0(t)$ is given by

$$se\{\hat{\Delta}(t)\} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1(t)} + \frac{\pi_0(1-\pi_0)}{n_0(t)}}.$$

Therefore, to obtain the desired power of $1 - \beta$ to detect the alternative where the population response rates were π_1 and π_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: The sample size formula given above is predicated on the use of the test statistic

$$T(t) = \frac{p_1(t) - p_0(t)}{\sqrt{\frac{p_1(t)\{1-p_1(t)\}}{n_1(t)} + \frac{p_0(t)\{1-p_0(t)\}}{n_0(t)}}}$$

to test for treatment difference in the response rates. Strictly speaking, this is not the same as the proportions test

$$T(t) = \frac{p_1(t) - p_0(t)}{\sqrt{\bar{p}(t)\{1-\bar{p}(t)\}\left\{\frac{1}{n_1(t)} + \frac{1}{n_0(t)}\right\}}},$$

although the difference between the two tests is inconsequential with equal randomization and large samples. What we discussed above is essentially the approach taken for sample size calculations used in Chapter 6 of the notes. The important point here is that power is driven by the amount of statistical information we have regarding the parameter of interest from the available data. The more data the more information we have. To achieve power $1-\beta$ to detect the clinically important difference Δ_A using a two-sided test at the α level of significance means that we need to have collected enough data so that the statistical information equals

$$\left\{\frac{\mathcal{Z}_{\alpha/2}+\mathcal{Z}_{\beta}}{\Delta_A}\right\}^2.$$

Let us examine how issues of power and information relate to group-sequential tests. If we are planning to conduct K interim analyses after equal increments of information, then the power of the group-sequential test to detect the alternative $\Delta=\Delta_A$ is given by

$$1-P_{\Delta=\Delta_A}\{|T(t_1)|<b_1,\ldots,|T(t_K)|<b_K\}.$$

In order to compute probabilities of events such as that above we need to know the joint distribution of the vector $\{T(t_1),\ldots,T(t_K)\}$ under the alternative hypothesis $\Delta=\Delta_A$.

It will be useful to consider the maximum information at the final analysis which we will denote as MI. A K-look group-sequential test with equal increments of information and with maximum information MI would have interim analyses conducted at times t_j where $j\times MI/K$ information has occurred; that is,

$$I(t_j,\Delta_A)=j\times MI/K,\ j=1,\ldots,K. \tag{10.8}$$

Using the results (10.2)-(10.4) and (10.8) we see that the joint distribution of $\{T(t_1),\ldots,T(t_K)\}$, under the alternative hypothesis $\Delta=\Delta_A$, is a multivariate normal with mean vector

$$\Delta_A\sqrt{\frac{j\times MI}{K}},\ j=1,\ldots,K$$

and covariance matrix V_T given by (10.6). If we define

$$\delta=\Delta_A\sqrt{MI},$$

then the mean vector is equal to

$$(\delta\sqrt{\frac{1}{K}},\delta\sqrt{\frac{2}{K}},\ldots,\delta\sqrt{\frac{K-1}{K}},\delta). \tag{10.9}$$

A group-sequential level-α test from the Wang-Tsiatis family rejects the null hypothesis at the first time $t_j, j = 1, \ldots, K$ where

$$|T(t_j)| \geq c(\alpha, K, \Phi) j^{(\Phi - .5)}.$$

For the alternative $H_A : \Delta = \Delta_A$ and maximum information MI, the power of this test is

$$1 - P_\delta[\bigcap_{j=1}^{K} \{|T(t_j)| < c(\alpha, K, \Phi) j^{(\Phi - .5)}\}],$$

where $\delta = \Delta_A \sqrt{MI}$, and $\{T(t_1), \ldots, T(t_K)\}$ is multivariate normal with mean vector (10.9) and covariance matrix V_T given by (10.6). For fixed values of α, K, and Φ, the power is an increasing function of δ which can be computed numerically using recursive integration. Consequently, we can solve for the value δ that gives power $1 - \beta$ above. We denote this solution by $\delta(\alpha, K, \Phi, \beta)$.

Remark: The value δ plays a role similar to that of a noncentrality parameter.

Since $\delta = \Delta_A \sqrt{MI}$, this implies that a group-sequential level-α test with shape parameter Φ, computed at equal increments of information up to a maximum of K times needs the maximum information to equal

$$\Delta_A \sqrt{MI} = \delta(\alpha, K, \Phi, \beta)$$

or

$$MI = \left\{ \frac{\delta(\alpha, K, \Phi, \beta)}{\Delta_A} \right\}^2$$

to have power $1 - \beta$ to detect the clinically important alternative $\Delta = \Delta_A$.

10.5.1 Inflation Factor

A useful way of thinking about the maximum information that is necessary to achieve prespecified power with a group-sequential test is to relate this to the information necessary to achieve prespecified power with a fixed sample design. In formula (10.7), we argued that the information necessary to detect the alternative $\Delta = \Delta_A$ with power $1 - \beta$ using a fixed sample test at level α is

$$I^{FS} = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta}{\Delta_A} \right\}^2.$$

In contrast, the maximum information necessary at the same level and power to detect the same alternative using a K-look group-sequential test with shape parameter Φ is

$$MI = \left\{ \frac{\delta(\alpha, K, \Phi, \beta)}{\Delta_A} \right\}^2 .$$

Therefore

$$\begin{aligned} MI &= \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta}{\Delta_A} \right\}^2 \left\{ \frac{\delta(\alpha, K, \Phi, \beta)}{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta} \right\}^2 \\ &= I^{FS} \times IF(\alpha, K, \Phi, \beta), \end{aligned}$$

where

$$IF(\alpha, K, \Phi, \beta) = \left\{ \frac{\delta(\alpha, K, \Phi, \beta)}{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta} \right\}^2$$

is the inflation factor, or the relative increase in information necessary for a group-sequential test to have the same power as a fixed sample test.

Note: The inflation factor does not depend on the endpoint of interest or the magnitude of the treatment difference that is considered clinically important. It only depends on the level (α), power ($1 - \beta$), and the group-sequential design (K, Φ). The inflation factors has been tabulated for some of the group-sequential tests which are given in the following table.

Table 10.4: *Inflation factors as a function of K, α, β and the type of boundary*

		α=0.05			α=0.01		
		Power=1-β			Power=1-β		
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

This result is convenient for designing studies which use group-sequential stopping rules as it can build on techniques for sample size computations used traditionally for fixed sample tests. For example, if we determined that we needed to recruit 500 patients into a study to obtain some prespecified power to detect a clinically important treatment difference using a traditional fixed sample design, where information, say, is proportional to sample size, then in order that we have the same power to detect the same treatment difference with a group-sequential test we would need to recruit a maximum number of $500 \times IF$ patients, where IF denotes the corresponding inflation factor for that group-sequential design. Of course, interim analyses would be conducted after every $\frac{500 \times IF}{K}$ patients had complete response data, a maximum of K times, with the possibility that the trial could be stopped early if any of the interim test statistics exceeded the corresponding boundary. Let us illustrate with a specific example.

Example with dichotomous endpoint

Let π_1 and π_0 denote the population response rates for treatments 1 and 0 respectively. Denote the treatment difference by $\Delta = \pi_1 - \pi_0$ and consider testing the null hypothesis $H_0 : \Delta = 0$ versus the two-sided alternative $H_A : \Delta \neq 0$. We decide to use a 4-look O'Brien-Fleming boundary; i.e. $K = 4$ and $\Phi = 0$, at the .05 level of significance ($\alpha = .05$). Using Table 10.2, we derive the boundaries which correspond to rejecting H_0 whenever

$$|T(t_j)| \geq 4.049/\sqrt{j},\ j = 1, \ldots, 4.$$

The boundaries are given by

Table 10.5: *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

In designing the trial, the investigators tell us that they expect the response rate on the control treatment (treatment 0) to be about .30 and want to have at least 90% power to detect a significant difference if the new treatment increases the response by .15 (i.e. from .30 to .45) using a two-sided test at the .05 level of significance. They plan to conduct a two arm randomized study with equal allocation and will test the null hypothesis using the standard proportions test.

The traditional fixed sample size calculations using the methods of chapter six, specifically formula (6.4), results in the desired fixed sample size of

$$n^{FS} = \left\{ \frac{1.96 + 1.28\sqrt{\frac{.3\times.7+.45\times.55}{2\times.375\times.625}}}{.15} \right\}^2 \times 4 \times .375 \times .625 = 434,$$

or 217 patients per treatment arm.

Using the inflation factor from Table 10.4 for the 4-look O'Brien-Fleming boundaries at the .05 level of significance and 90% power i.e. 1.02, we compute the maximum sample size of 434 × 1.02=444, or 222 per treatment arm. To implement this design, we would monitor the data after every 222/4 ≈ 56 individuals per treatment arm had complete data regarding their response for a maximum of four times. At each of the four interim analyses we would compute the test statistic, i.e. the proportions test

$$T(t_j) = \frac{p_1(t_j) - p_0(t_j)}{\sqrt{\bar{p}(t_j)\{1-\bar{p}(t_j)\}\left\{\frac{1}{n_1(t_j)} + \frac{1}{n_0(t_j)}\right\}}},$$

using all the data accumulated up to the j-th interim analysis. If at any of the four interim analyses the test statistic exceeded the corresponding boundary given in Table 10.5 or, equivalently, if the two-sided p-value was less than the corresponding nominal p-value in Table 10.5, then we would reject H_0. If we failed to reject at all four analyses we would then accept H_0.

10.5.2 Information based monitoring

In the above example it was assumed that the response rate on the control treatment arm was .30. This was necessary for deriving sample sizes. It may be, in actuality, that the true response rate for the control treatment is something different, but even so, if the new treatment can increase the probability of response by .15 over the control treatment we may be interested in detecting such a difference with 90% power. We've argued that power is directly related to information. For a fixed sample design, the information necessary to detect a difference $\Delta = .15$ between the response probabilities of two treatments with power $1-\beta$ using a two-sided test at the α level of significance is

$$\left\{\frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta}{\Delta_A}\right\}^2.$$

For our example, this equals

$$\left\{\frac{1.96 + 1.28}{.15}\right\}^2 = 466.6.$$

For a 4-look O-F design this information must be inflated by 1.02 leading to $MI = 466.6 \times 1.02 = 475.9$. With equal increments, this means that an analysis should be conducted at times t_j when the information equals $\frac{j \times 475.9}{4} = 119 \times j,\ j = 1, \ldots, 4$. Since information is approximated by $[se\{\hat{\Delta}(t)\}]^{-2}$, and in this example (comparing two proportions) is equal to

$$\left[\frac{p_1(t)\{1-p_1(t)\}}{n_1(t)} + \frac{p_0(t)\{1-p_0(t)\}}{n_0(t)}\right]^{-1},$$

we could monitor the estimated standard deviation of the treatment difference estimator and conduct the four interim analyses whenever

$$[se\{\hat{\Delta}(t)\}]^{-2} = 119 \times j,\ j = 1, \ldots, 4,$$

i.e. at times t_j such that

$$\left[\frac{p_1(t_j)\{1-p_1(t_j)\}}{n_1(t_j)}+\frac{p_0(t_j)\{1-p_0(t_j)\}}{n_0(t_j)}\right]^{-1}=119\times j,\ j=1,\ldots,4.$$

At each of the four analysis times we would compute the test statistic $T(t_j)$ and reject H_0 at the first time that the boundary given in Table 10.5 was exceeded.

This information-based procedure would yield a test that has the correct level of significance ($\alpha = .05$) and would have the desired power ($1-\beta = .90$) to detect a treatment difference of .15 in the response rates regardless of the underlying true control treatment response rate π_0. In contrast, if we conducted the analysis after every 112 patients (56 per treatment arm), as suggested by our preliminary sample size calculations, then the significance level would be correct under H_0 but the desired power would be achieved only if our initial guess (i.e. $\pi_0 = .30$) were correct. Otherwise, we would over power or under power the study depending on the true value of π_0 which, of course, is unknown to us.

10.5.3 Average information

We still haven't concluded which of the proposed boundaries (Pocock, O-F, or other shape parameter Φ) should be used. If we examine the inflation factors in Table 10.4 we notice that K-look group-sequential tests that use the Pocock boundaries require greater maximum information that do K-look group-sequential tests using the O-F boundaries at the same level of significance and power; but at the same time we realize that Pocock tests have a better chance of stopping early than O-F tests because of the shape of the boundary. How do we assess the trade-offs?

One way is to compare the average information necessary to stop the trial between the different group-sequential tests with the same level and power. A good group-sequential design is one which has a small average information.

Remark: Depending on the endpoint of interest this may translate to smaller average sample size or smaller average number of events, for example.

How do we compute average information?

We have already discussed that the maximum information MI is obtained by computing the information necessary to achieve a certain level of significance and power for a fixed sample design and multiplying by an inflation factor. For designs with a maximum of K analyses after equal increments of information, the inflation factor is a function of α (the significance level), β (the type II error or one minus power), K, and Φ (the shape parameter of the boundary). We denote this inflation factor by $IF(\alpha, K, \Phi, \beta)$.

Let V denote the number of interim analyses conducted before a study is stopped. V is a discrete integer-valued random variable that can take on values from $1, \ldots, K$. Specifically, for a K-look group-sequential test with boundaries $b_1, \ldots, b_K$, the event $V = j$ (i.e. stopping after the j-th interim analysis) corresponds to

$$(V = j) = \{|T(t_1)| < b_1, \ldots, |T(t_{j-1})| < b_{j-1}, |T(t_j)| \geq b_j\},\ j = 1, \ldots, K.$$

The expected number of interim analyses for such a group-sequential test, assuming $\Delta = \Delta^*$ is given by

$$E_{\Delta^*}(V) = \sum_{j=1}^{K} j \times P_{\Delta^*}(V = j).$$

Since each interim analysis is conducted after increments MI/K of information, this implies that the average information before a study is stopped is given by

$$AI(\Delta^*) = \frac{MI}{K} E_{\Delta^*}(V).$$

Since $MI = I^{FS} \times IF(\alpha, K, \Phi, \beta)$, then

$$AI(\alpha, K, \Phi, \beta, \Delta^*) = I^{FS} \left[\left\{ \frac{IF(\alpha, K, \Phi, \beta)}{K} \right\} E_{\Delta^*}(V) \right].$$

Note: We use the notation $AI(\alpha, K, \Phi, \beta, \Delta^*)$ to emphasize the fact that the average information depends on the level, power, maximum number of analyses, boundary shape, and alternative of interest. For the most part we will consider the average information at the null hypothesis $\Delta^* = 0$ and the clinically important alternative $\Delta^* = \Delta_A$. However, other values of the parameter may also be considered.

Using recursive numerical integration, the $E_{\Delta^*}(V)$ can be computed for different sequential designs at the null hypothesis, at the clinically important alternative Δ_A, as well as other values for the treatment difference. For instance, if we take $K = 5$, $\alpha = .05$, power equal to 90%, then under $H_A : \Delta = \Delta_A$, the expected number of interim analyses for a Pocock design is equal to $E_{\Delta_A}(V) = 2.83$. Consequently, the average information necessary to stop a trial, if the alternative H_A were true would be

$$I^{FS}\left[\left\{\frac{IF(.05,5,.5,.10)}{5}\right\}\times 2.83\right]$$
$$= I^{FS}\left\{\frac{1.21}{5}\times 2.83\right\}$$
$$= I^{FS}\times .68.$$

Therefore, on average, we would reject the null hypothesis using 68% of the information necessary for a fixed-sample design with the same level (.05) and power (.90) as the 5-look Pocock design, if indeed, the clinically important alternative hypothesis were true. This is why sequential designs are sometimes preferred over fixed-sample designs.

Remark: If the null hypothesis were true, then it is unlikely ($< .05$) that the study would be stopped early with the sequential designs we have been discussing. Consequently, the average information necessary to stop a study early if the null hypothesis were true would be close to the maximum information (i.e. for the 5-look Pocock design discussed above we would need almost 21% more information than the corresponding fixed-sample design).

In contrast, if we use the 5-look O-F design with $\alpha = .05$ and power of 90%, then the expected number of interim analyses equals $E_{\Delta_A}(V) = 3.65$ under the alternative hypothesis H_A. Thus, the average information is

$$I^{FS}\left[\left\{\frac{IF(.05,5,0.0,.10)}{5}\right\}\times 3.65\right]$$
$$= I^{FS}\left\{\frac{1.03}{5}\times 3.65\right\}$$
$$= I^{FS}\times .75.$$

Summarizing these results: For tests at the .05 level of significance and 90% power, we have

	Maximum	Average
Designs	information	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}

Recall:

$$I^{FS} = \left\{ \frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_\beta}{\Delta_A} \right\}^2 .$$

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 Φ which minimizes the average information under the clinically important alternative Δ_A for different values of α, K, and power $(1-\beta)$. Some these optimal Φ are provided in the paper by Wang and Tsiatis (1987) *Biometrics*. For example, when $K = 5$, $\alpha = .05$ and power of 90% the optimal shape parameter $\Phi = .45$, which is very close to the Pocock boundary.

- Keep in mind, however, that the designs with better stopping properties under the alternative need greater maximum information which, in turn, implies greater information will be needed 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.

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.

10.5.4 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 Φ. We've already argued the pros and cons of conservative boundaries such as O-F versus more aggressive boundaries such as Pocock. As mentioned previously, for a variety of statistical and non-statistical reasons, conservative boundaries have been preferred in practice. In terms of the number of looks K, it turns out that the properties of a group-sequential test is for the most part insensitive to the number of looks after a certain point. We illustrate this point using the following table which looks at the maximum information and the average information under the alternative for the O'Brien-Fleming boundaries for different values of K.

Table 10.6: *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	$I^{FS} \times 1.03$	$I^{FS} \times .75$

We note from Table 10.6 that there is little change in the early stopping properties of the group-sequential test once K exceeds 4. Therefore, the choice of K should be chosen based on logistical and practical issues rather than statistical principles (as long as K exceeds some lower threshold; i.e. 3 or 4). For example, the choice might be determined by how many times one can feasibly get a data monitoring committee to meet.

2. Compute the information necessary for a fixed sample design and translate this into a physical design of resource use. You will need to posit some initial guesses for the values of the nuisance parameters as well as defining the clinically important difference that you want to detect with specified power using a test at some specified level of significance in order to derive sample sizes or other design characteristics. This is the usual "sample size considerations" that were discussed throughout the course.

3. The fixed sample information must be inflated by the appropriate inflation factor $IF(\alpha, K, \Phi, \beta)$ to obtain the maximum information

$$MI = I^{FS} \times IF(\alpha, K, \Phi, \beta).$$

Again, this maximum information must be translated into a feasible resource design using initial guesses about the nuisance parameters. For example, if we are comparing the response rates of a dichotomous outcome between two treatments, we generally posit the response rate for the control group and we use this to determine the required sample sizes as was illustrated in the example of section 10.5.1.

Analysis

4. After deriving the maximum information (most often translated into a maximum sample size based on initial guesses), the actual analyses will be conducted a maximum of K times after equal increments of MI/K information.

Note: Although information can be approximated by $[se\{\hat{\Delta}(t)\}]^{-2}$, in practice, this is not generally how the analysis times are determined; but rather, the maximum sample size (determined based on best initial guesses) is divided by K and analyses are conducted after equal increments of sample size. Keep in mind, that this usual strategy may be under or over powered if the initial guesses are incorrect.

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 analyses are conducted after equal increments of information. So, for instance, if we have a problem where information is proportional to sample size, then as long as the analyses are conducted after equal increments of sample size we are guaranteed to have the correct type I error. Therefore, when we compute sample sizes based on initial guesses for the nuisance parameters and monitor after equal increments of this sample size, the corresponding test has the correct level of significance under the null hypothesis.

However, in order that this test have the correct power to detect the clinically important difference Δ_A, it must be computed after equal increments of statistical information MI/K where

$$MI = \left\{\frac{\mathcal{Z}_{\alpha/2} + \mathcal{Z}_{\beta}}{\Delta_A}\right\}^2 IF(\alpha, K, \Phi, \beta).$$

If the initial guesses were correct, then the statistical information obtained from the sample sizes (derived under these guesses) corresponds to that necessary to achieve the correct power. If, however, the guesses were incorrect, then the resulting test may be under or over powered depending on whether there is less or more statistical information associated with the given sample size.

Although an information-based monitoring strategy, such as that outlined in section 10.5.2, is not always practical, I believe that information (i.e. $[se\{\hat{\Delta}(t)\}]^{-2}$) should also be monitored as the study progresses and if this deviates substantially from that desired, then the study team should be made aware of this fact so that possible changes in design might be considered. The earlier in the study that problems are discovered, the easier they are to fix.