ST 520
Statistical Principles of Clinical Trials

Lecture Notes

Anastasios (Butch) Tsiatis

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 later in the course we will we will give a brief introduction to statistical issues in Epidemiology.

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 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 1.1: *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.

There has been a great deal of research on Phase I clinical trials especially with regards to dose escalation with the goal of giving as high an efficacious dose as possible without undue toxicity; the so-called maximum tolerated dose (MTD). Designs for such phase I trials focus on using as few patients as possible while still being able to estimate the MTD with some degree of accuracy. This is necessary as we take treatments from Phase I to the later phases. In this course we will

focus primarily on Phase II and Phase III clinical trials with the most emphasis on Phase III clinical trials.

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 $\mathcal{C}(k)$.

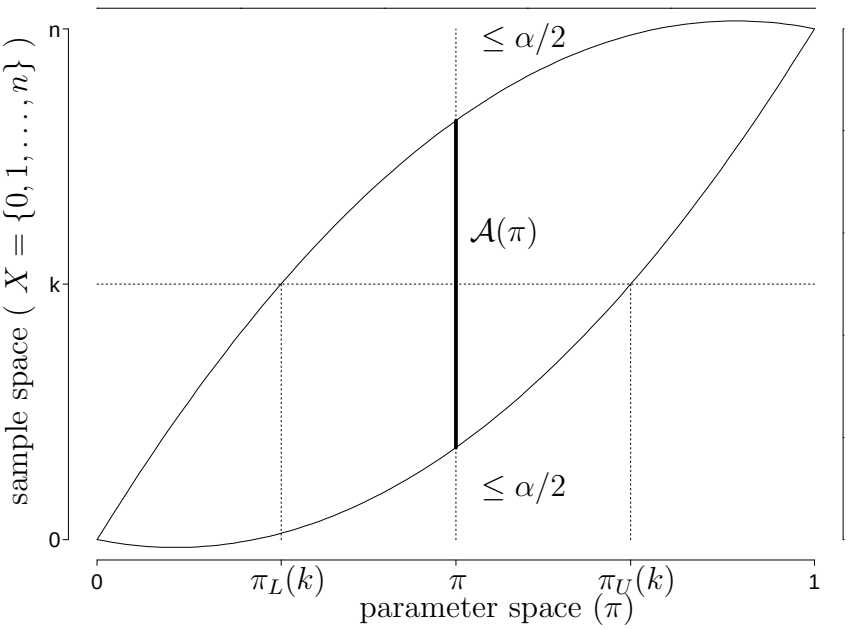
This definition is equivalent to defining an acceptance region (of the sample space) for each value π, denoted as $\mathcal{A}(\pi)$, that has probability greater than equal to $1-\alpha$, i.e.

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

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

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

and

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

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 $\pm 15\%$ 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.

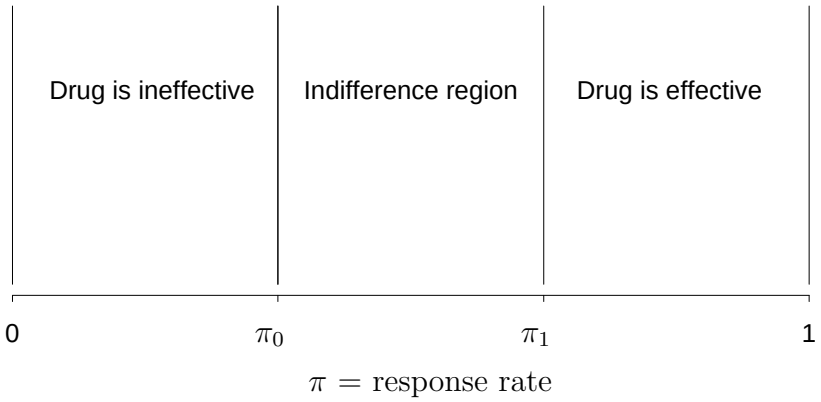
2.2.3 Simon's Two-Stage Design

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

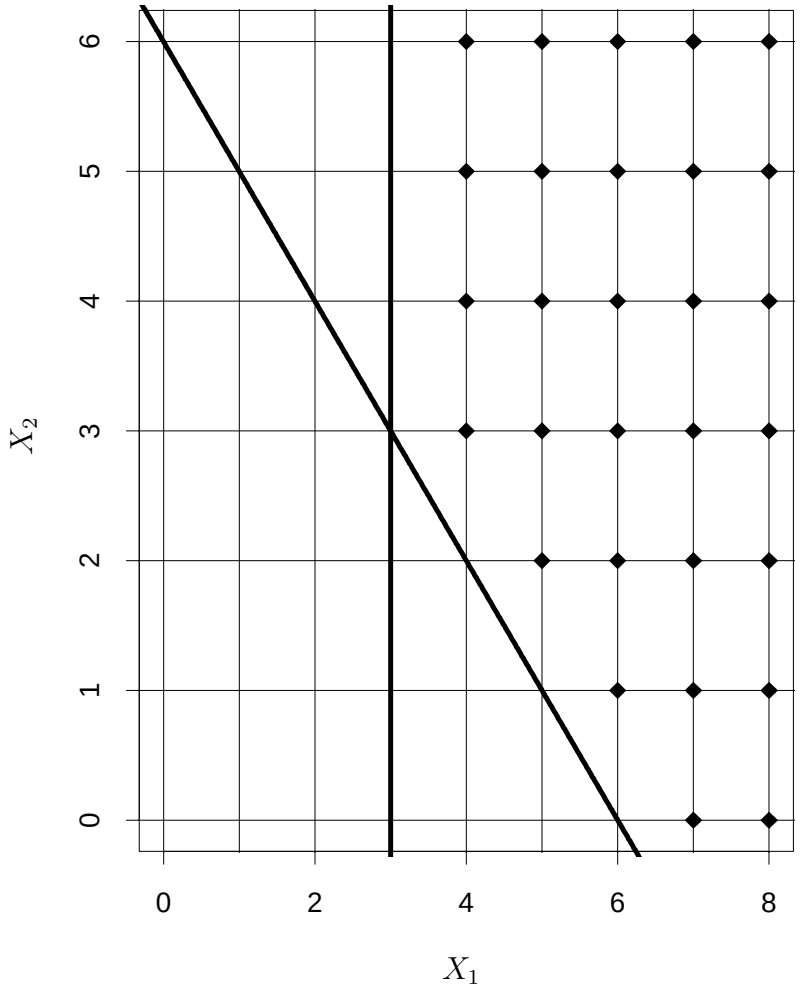
$$P(X_1 = m_1, X_2 = m_2|\pi) = P(X_1 = m_1|\pi) \times P(X_2 = m_2|\pi)$$

$$= \left\{ \binom{n_1}{m_1} \pi^{m_1}(1-\pi)^{n_1-m_1} \right\} \left\{ \binom{n_2}{m_2} \pi^{m_2}(1-\pi)^{n_2-m_2} \right\}.$$

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	$\leq r_1/n_1$	$\leq r/n$	EN(p_0)	PET(p_0)	$\leq r_1/n_1$	$\leq r/n$	EN(p_0)	PET(p_0)
0.05	0.20	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.