Introduction to Clinical trials

Instructor:
Constantin Yiannoutsos, Ph.D.
Professor of Biostatistics
Indiana University
School of R.M. Fairbanks School of Public Health, Department of Biostatistics

Clinical trials in context

Clinical trials in context

The clinical trial apparatus is diffused variably dependent on the context of the application. The specific context of clinical trials discussed in this lecture are:

- Drug trials
- Development of devices
- Complementary and alternative medicine (CAM)
- Prevention trials
- Surgery and skill-dependent therapies

Characteristics with implications for clinical trials

	Context				
Characteristic	Drugs	Devices	Prevention	CAM	Surgery
Regulation	Strong	Moderate	Strong	Minimal	None
Bias control	Easy	Difficult	Easy	Easy	Difficult
Uniformity	High	High	High	Low	Low
Effect size	Small	Large	Moderate	Small	Large
Incremental improvement	Common	Common	Minimal	None	Common
Short-term risk/benefit	Favorable	Varied	Favorable	Unknown	Varied
Long-term risk/benefit	Varied	Varied	Favorable	Unknown	Favorable
Tradition	Strong	Weak	Strong	None	Varied

Drug trials

Clinical trial methodology has been particularly fit for drug development. The reasons are as follows:

- **Regulation**
  Due to well-known disasters (e.g., thalidomide) regulation has been explicit albeit non-uniform regarding drug trials.
- **Tradition**
  There is strong tradition for experimental evaluation of drugs
- **Confounding**
  Drug trials are just as susceptible to confounding and selection bias as other contexts

Drug trials (continued)

- **Incremental improvement**
  This is the rule rather than the exception with drugs (e.g., better absorption, lower toxicity, etc.) and must undergo rigorous testing
- **Placebos**
  Drugs are uniquely suited to the use of placebos
- **Economics**
  Several economic reasons based on the structure of health care in the US are supportive to rigorous testing of drugs.
- **Psychology**
  For patients, knowing that a drug has been rigorously tested improves its psychological acceptance

Devices

Borrowing the informal definition by Piantadosi (Piantadosi, 2005), any object used in or on the body for which a health claim is made can be considered a "medical device". Development of medical devices is subtantively different from drugs as follows:

- **Regulation**
  Regulation is substantially uneven for marketing of devices. FDA approval has been required as of 1976 and requires the satisfaction of two standards:
  - Substantial equivalence (501(k)) to a device approved prior to 1976 (and some after 1976 as well)
  - Premarket approval application (PMA)

  As a result, only a minority of devices have supporting clinical data.

Devices (continued)

- **Tradition**
  Device development should not be different from drugs, but there is no strong tradition for rigorous testing of devices and device functionality is often confused for efficacy.
- **Placebos**
  Device assessment is not amenable to the use of placebos, with some notable exceptions (deep brain stimulation in Parkinson's disease)

Prevention trials

Prevention trials are divided into three categories

- **Primary prevention**
  These are prevention approaches for healthy subjects
- **Secondary**
  Prevention of subjects at high risk for the disease
- **Tertiary**
  These are strategies to prevent recurrences in patients that already have the disease

Methodological challenges in prevention trials

The main methodological challenges in prevention trials is

- Long-term adherence and delayed treatment effect distort the treatment effect
- Perception of risk and benefit is different depending on whether prevention is primary or secondary/tertiary prevention (i.e., whether prevention is applied to healthy versus sick individuals). A classic example is the study is tamoxifen for the prevention of breast cancer recurrence. Another is the lessening support for vaccines among the US public.

CAM

It is difficult to define alternative medicine. A coarse definition given by Piantadosi (Piantadosi, 2005) is "a treatment whose mechanism of action is poorly defined or incompatible with established biology". This is a limited and perhaps short-sighted definition that does not apply to all situations.

The problem with evaluating CAM interventions is that many CAM treatments are not amenable to traditional (see measurable) scientific methods. As a result, clinical trials have not been frequently used to assess CAM approaches.

CAM

- **Regulation**
  There is no governmental regulation requiring CAM approaches to be safe and effective like there are for drugs and some devices
- **Tradition**
  There is no strong tradition among CAM practitioners for experimental trials
- **Incremental improvement**
  There is no formalized development of CAM treatments as there is for drugs, although CAM methods are amenable to testing and placebo controls.

CAM (continued)

- **Economics**
  Economics such as small overhead for development of these therapies, disincentive for voluntary testing as well as the lack of patents are against large clinical trials in evaluation of CAM approaches.
- **Psychology**
  Patients are usually very accepting of CAM approaches

Surgery and skill-based therapies

Surgical methods have not been applied to rigorous testing because practitioners see their methods as intuitively justified and have favorable risk-benefit ratio especially when applied to the right patients. Some of the contextual similarities and differences between surgical and therapeutic interventions are as follows:

- **Regulation**
  There is no governmental regulation requiring surgical approaches to be safe and effective, because they are not marketed directly to the public. A notable exception is eye corrective (laser) surgery that is marketed no differently than drugs.

Surgery and skill-based techniques (continued)

- **Tradition**
  There is a tradition of reverence for experience and not for rigorous testing (that involves questioning standard approaches) in surgery, which goes against the broad appeal of rigorous testing in surgical interventions.
- **Incremental improvement**
  Incremental improvement is the norm for surgical and skill-based techniques.

Surgery and skill-based techniques (continued)

- **Confounding**
  Surgical methods are susceptible to observer bias and patient selection bias. These are grouped together as follows:
  - Prognosis of patients (weaker patients may not survive surgery artificially distorting the efficacy of the technique
  - Surgical techniques are strongly dependent on the skill, experience and supportive care quality of the practitioner
  - Efficacy of the procedure (early promising results may establish the technique and make testing virtually impossible)

Surgery and skill-based techniques (continued)

- **Placebos**
  Placebos are generally seen as unethical when applied to surgery from the perspective of risk-benefit for the patient. Some notable exceptions have been observed.
- **Economics**
  Economics lead in the early adoption of promising surgical techniques. However, this also has the effect that the individual uncertainty that needs to be in place to allow practitioners to participate in clinical trials is not there and thus, these techniques cannot be rigorously tested once adopted broadly.
- **Psychology**
  Patients are usually very accepting of surgical procedures because they hold the promise of quick alleviation of the problem.

Statistical perspectives

Clinical trials as hybrid of clinical and statistical reasoning

It is very important for all those involved in clinical trials to understand the statistical perspective involved in their conduct. Statistics are used as

- A descriptive tool
- An analytical science
- An aid in making decisions

There are different "schools of thought" in statistics and following the methodology of one or the other may lead to different (but not necessarily less valid) interpretations of the same data.

Models and parameters

Essentially, all models are wrong, but some are useful – George Box

A model describes the relationships between two observable quantities. An example is the following simple (linear) model:

$$y = \overbrace{a + bx}^{\text{deterministic part}} + \overbrace{\epsilon}^{\text{random part}}$$

Data yield information about model parameters (a and b above), which, in turn, provide insight into nature. Whether one views a and b as constants or random variables has important consequences in inference.

Schools of thought

There are two schools of thought or philosophies.

- The “frequentist” or “classical” approach
- The Bayesian school

The frequentist school

The Frequentist or Classical School of thought is the predominant approach to statistical thought.

For example, the probability of an unbiased coin coming up "Heads" is 50% because this is the frequency of this phenomenon if an unbiased coin is tossed infinite times.

One problem of the frequentist ideology is that it assumes infinite repetitions of an experiment (or the current experiment) under identical conditions (in a sort of infinite imaginary universes identical to ours). For example, a 30% probability that it will rain tomorrow is understood as the fact that, under identical conditions (in identical but imaginary universes like our own), rain will come 30% of the time.

The Bayesian school of thought

The Bayesian approach provides a sound mathematical framework to update *a priori* (prior) beliefs (e.g. the probability that the drug will be effective) by introduction of the data, into an *a posteriori* (posterior) probability (e.g., what is the probability of efficacy after completion of the clinical study).

Problems and controversies

There are two problems that have limited the appeal of Bayesian techniques:

- **The controversy over prior beliefs**
  Prior belief can weigh substantially on the outcome so that the posterior probabilities based on the same data can be much different or even contradictory given disparate prior beliefs. While this is not an inconsistency inherent in the method, reaching different conclusions from the same data has been problematic to say the least.
- **Computational issues**
  Bayesian techniques are not always amenable to analytical (i.e., exact) calculations and thus have been inaccessible to statisticians until fairly recently.

Binomial analysis example

We provide the following example in order to outline the three methods and focus on their areas of agreement and, more importantly, disagreement or deviation from each other.

Assume we have 20 independent Bernoulli observations (i.e., like coin tosses) with probability of success θ and failure $1 - \theta$. We will try to make inferences about the probability of success θ. All three schools of thought agree that θ, the parameter of interest, follows a Binomial distribution with probability density function

$$P(X = r) = \left(\begin{array}{c} n \\ r \end{array} \right) \theta^r (1 - \theta)^{n-r}$$

The Binomial likelihood

If we obtain $n = 20$ samples and observe $k = 8$ successes we may ask: What is the chance that, under this model, we would obtain 8 successes? This is called the "likelihood" of the data and, for the binomial probability model, it is given by the relationship

$$L(\theta) = \left(\begin{array}{c} 20 \\ 8 \end{array} \right) \theta^8 (1-\theta)^{20-8}$$

The notation $L(\theta)$ suggests that the likelihood is a function of θ (the data $\mathbf{x} = (n, k) = (20, 8)$ are considered fixed at their observed values).

It is important to note that the likelihood has all the information that we have about θ.

Maximum likelihood

The likelihood function $L(\theta)$ is maximized for a particular value $\hat{\theta}$. This value of theta, which is called *maximum likelihood estimate* of θ, is considered as the most consistent value with the data.

The value of θ maximizing the likelihood or, more routinely, the log-likelihood, can be found using straightforward algebraic methods.

Maximum likelihood in the Binomial example

The log-likelihood in the binomial example is proportional to (excluding the log of the factorials)

$$\log L(\theta) \propto r \log \theta + (n - r) \log(1 - \theta)$$

For $n = 20$ and $k = 8$ is maximized for $\hat{\theta} = \frac{8}{20} = 0.4$ this is:

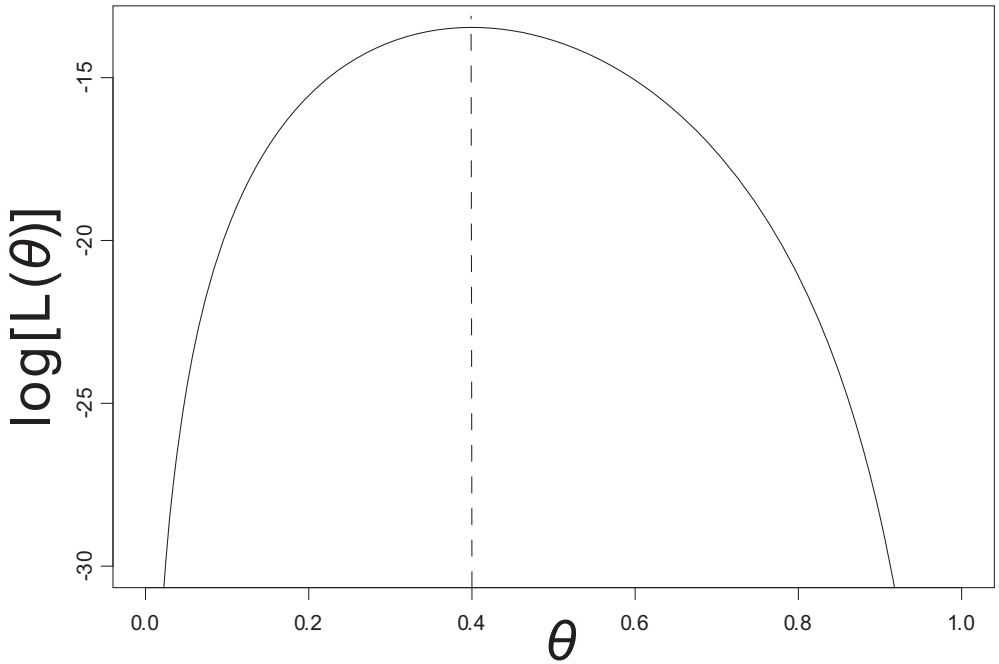

Figure 1: Log likelihood function in the binomial experiment)

Frequentist approach

Frequentist thought understands θ as having a constant value in nature. Variability is totally associated with the sampling procedure. For example, if the hypothesized value of $\theta = 0.5$ the sampling distribution of $\hat{\theta}$ is as follows:

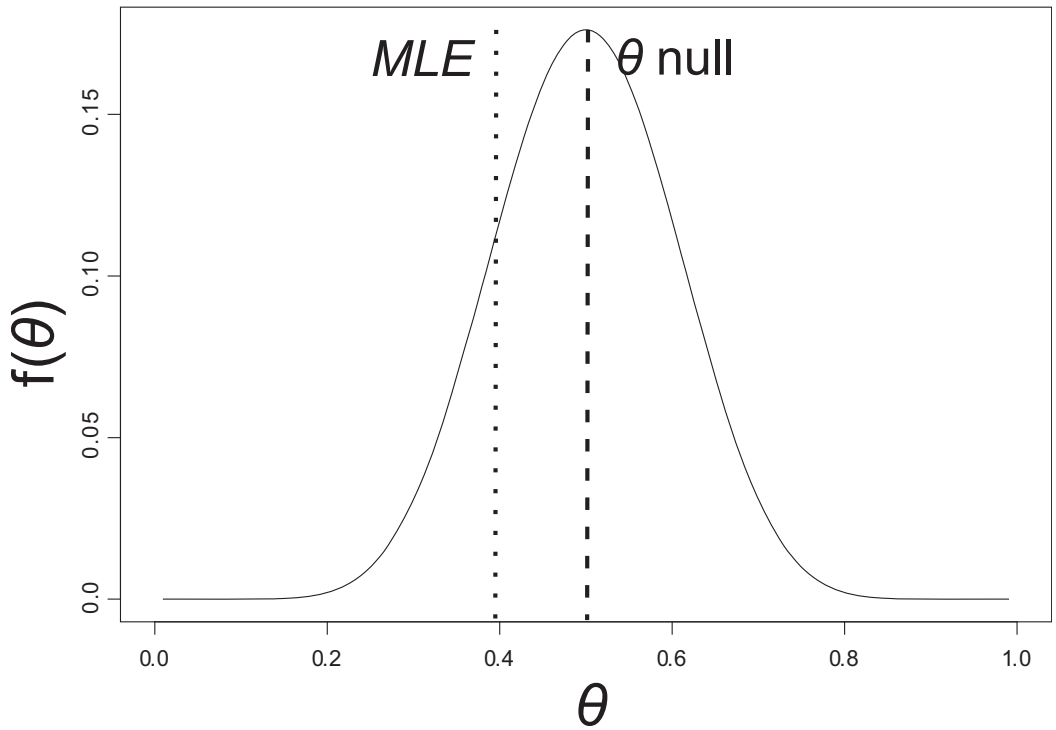

Figure 2: Sampling distribution of $\hat{\theta}$ under the null hypothesis $\theta = 0.5$

Frequentist inference

Hypothesis testing

Frequentist inference occurs when probability statements are made about the plausibility that the data arose from a model with a specific hypothetical value for θ.

A "hypothesis test" assesses how plausible it is that the observed data arose from a model with a specific value for θ. In the above example, this would be based on the quantity

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} = \frac{0.4 - 0.5}{0.5/\sqrt{20}} = -0.894$$

The two-sided area corresponding to $Z = -0.894$ (i.e., the area under the standard normal curve below -0.894 and above 0.894) is 0.57, which implies that the data could have arisen from a $\theta = 0.5$.

Understanding p values

P-values are the probability that the observed value or a more extreme value would be possibly observed even if the null hypothesis were correct.

This means that the p-value is the probability of a *false positive* event. In other words, the p-value is the probability that we reject the null hypothesis when it is true!

This is the reason that, if the p value is less than the alpha level, we reject the null hypothesis.

Note that p values *rank* the strength of the evidence provided by a test. If we have two p values $p_1 < p_2$, then the experiment that resulted in p_1 provided stronger evidence than the one resulting in p_2. We will see the implications of this again later in this course.

Frequentist inference

Confidence intervals

Another concept is the "confidence interval". A 95% confidence interval contains the true value for θ 95% of the time.

The point of reference of the confidence interval is $\hat{\theta} = 0.4$. A 95% confidence interval in the example above would have as a lower bound the value below which lies at most 2.5% of the binomial probability and as an upper value the one above which lies at most 2.5% of the binomial probability (with $n = 20$ and $r = 8$).

A 95% confidence interval in the example is given by the expression

$$\hat{\theta} \pm z_{97.5\%} \sqrt{\hat{\theta}(1-\hat{\theta})/n}$$

leading to the confidence interval $(0.185, 0.615)$, which contains the hypothesized value of $\theta = 0.5$.

Problems with the frequentist procedure

The problem with frequentist procedures is that they literally assume that the current (and thus unique) experiment is replicated under identical conditions infinite times.

To report the results of the inference to a non-statistical audience we deliberately misstate this and report that the unknown parameter θ will be inside the confidence interval 95% of the time.

The Bayesian approach

Where the Frequentist school of thought considers the observed data as one realization of infinite replicates of the experiment and the parameter θ fixed, the Bayesian approach considers the observed data as unique (and thus fixed) and the parameters θ as variable and assigns probabilities to them.

The Bayes theorem

The basis of the Bayesian school of thought is the, so called, Bayes theorem. This theorem is a mathematical way to invert conditional probabilities as follows:

If B and its complement B^c are disjoint and mutually exclusive events and A is any other event, then

$$Pr\{B|A\} = \frac{Pr\{A|B\}Pr\{B\}}{Pr\{A|B\}Pr\{B\} + Pr\{A|B^c\}Pr\{B^c\}}$$

This theorem allows one to "update" the probability associated with an event B by incorporating evidence subsumed in event A (e.g., beliefs about the efficacy of a drug or device before – $Pr\{B\}$ – and after – $Pr\{B|A\}$ – a clinical trial – $Pr\{A|B\}$).

Summarizing prior evidence

The problem with Bayesian methods is the attempt to summarize all information available *prior* to conducting the study. This is done in the form of a prior distribution.

Translating expert clinical opinion into a prior distribution ("eliciting a prior") is a difficult process in itself. In addition, different practitioners will have different levels of enthusiasm about the likely efficacy of a therapy or intervention. This in turn will translate into different prior distributions which may result in different conclusions.

This is against the idea that scientific inference is an objective process that should reach the same conclusions under the same data.

Eliciting different priors

Different types of prior distributions can be used depending on the strength of the prior evidence or within sensitivity analyses later on.

- Reference priors

  These represent minimal amount of information

- Clinical priors

  These represent the beliefs of particularly well informed experts in the field

- Skeptical or enthusiastic priors

  The former consider large treatment effects unlikely, while the latter urge continuation of the trial even when the results do not support efficacy of the treatment

Priors in the Binomial case study

Suppose that, prior to the experiment, the most likely value for θ is 0.6 and that this value is not strongly favored over values such as 0.5 or 0.7 but is strongly favored over lower values such as 0.4.

This evidence could be summarized by considering a prior distribution of the form

$$f(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

where $B(a,b)$ is the Beta function and a and b are shape and location constants.

Beta priors in the Binomial case study

For example, consider three beta distribution priors with a and b respectively as A: $a = 13$, $b = 9$, B: $a = 4$, $b = 3$ and C: $a = 1$, $b = 1$. In particular, prior C gives equal chance to all values of θ from zero to one. This is called a *uninformative* prior. These are given in the following figure:

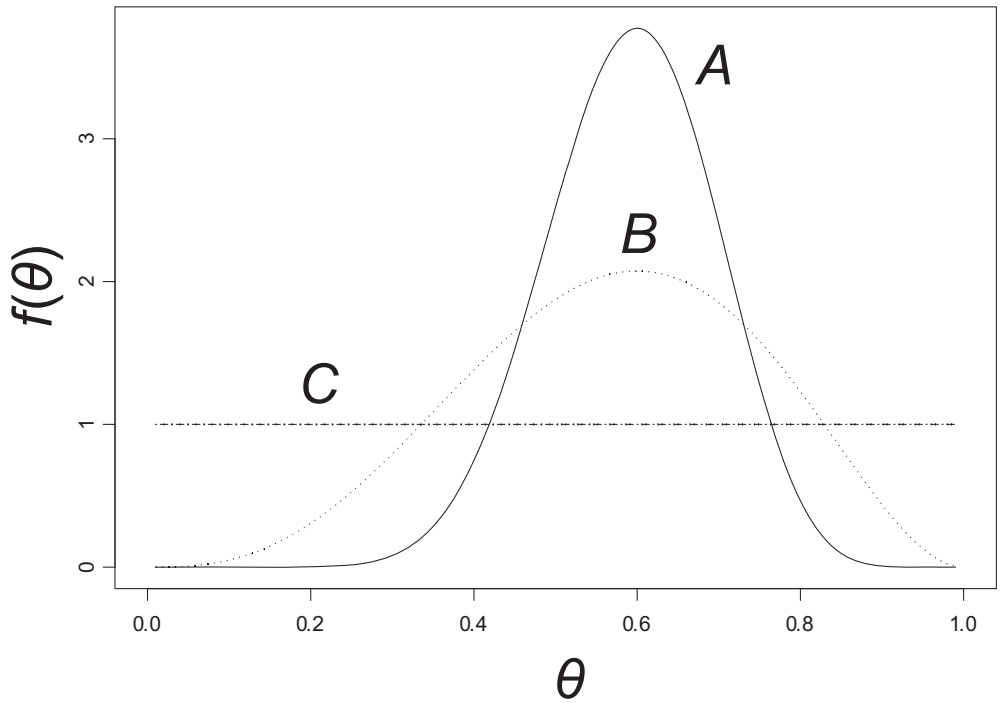

Figure 3: Beta priors in the binomial experiment

Updating prior beliefs

The posterior distribution

The next step is to combine the prior distribution $f(\theta)$ above with the binomial likelihood $L(\theta)$ (or more consistently to the Bayesian notation $L(\mathbf{x}|\theta)$). From Bayes' theorem, the posterior distribution is

$$g(\theta|\mathbf{x}) = \frac{L(\mathbf{x}|\theta)f(\theta)}{\int_{-\infty}^{+\infty} L(\mathbf{x}|\theta)f(\theta)f(\theta)d\theta}$$

Calculating $g(\theta|\mathbf{x})$ for arbitrary combinations of priors and likelihoods can be computationally demanding, a fact that has negatively impacted the appeal of Bayesian techniques prior to the advent of sufficient computing power and recent theoretical developments.

Posterior distribution in the Binomial experiment

Using the Beta distribution in the Binomial example makes calculations tractable because (see Piantadosi, 2005)

$$g(\theta|\mathbf{x}) = \frac{\theta^{r+a-1}(1-\theta)^{n-r+b-1}}{B(r+a, n-r+b)}$$

which is itself a Beta distribution. Choosing the beta distribution as the prior (what is called a *compatible* prior to the binomial) simplified the calculations.

Note the effect of the prior on the posterior distribution. The prior is like having observed a more successes and b more failures. The larger these numbers are the stronger prior beliefs are (and the more difficult it is to reverse them!).

Posterior distributions in the Binomial case study

The posterior distributions resulting from the above priors are given in the following figure:

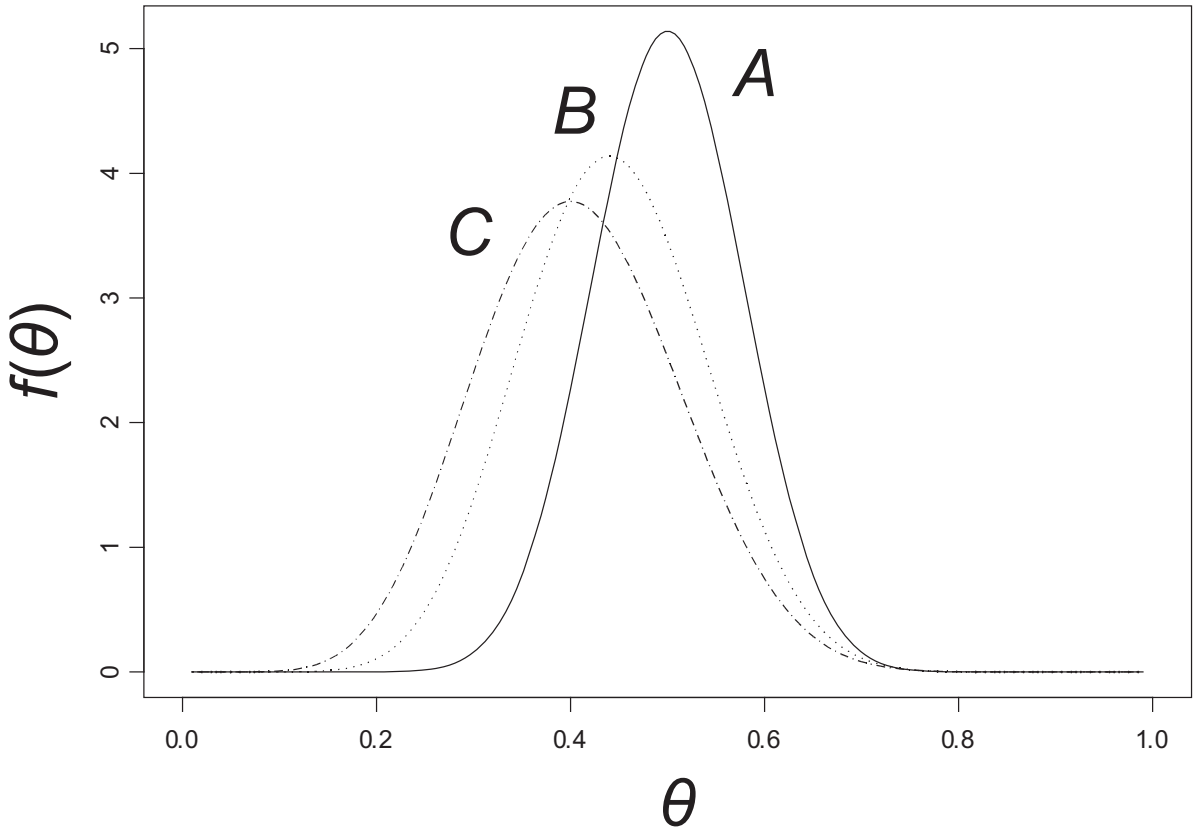

Figure 4: Beta posteriors in the binomial experiment)

Comments

It is clear from the Figure that the posterior distribution A has not moved a great deal towards the value $\hat{\theta} = 0.4$.

By contrast, B has moved somewhere between 0.4 and 0.6.

The noninformative prior produced a distribution most closely related to the frequentist approach (centered around $\hat{\theta} = 0.4$).

Bayesian credible intervals

The Bayesian version of the confidence interval in the frequentist approach is the *credible interval*. We show below some credible intervals based on posterior distribution C above for plausible values of θ.

Coverage	Bounds			
%	Symmetric		Equal tailed	
50%	0.337	0.479	0.338	0.480
90%	0.245	0.583	0.240	0.578
95%	0.218	0.616	0.211	0.608
99%	0.171	0.677	0.155	0.663

Note the difference between Bayesian credible intervals and frequentist confidence intervals. The former, as they do not assume a symmetric distribution (as the latter most often do), so that we can choose intervals which have the same tails, but this does not mean that their bounds will be equidistant from the point estimate.

Homo Bayesianis

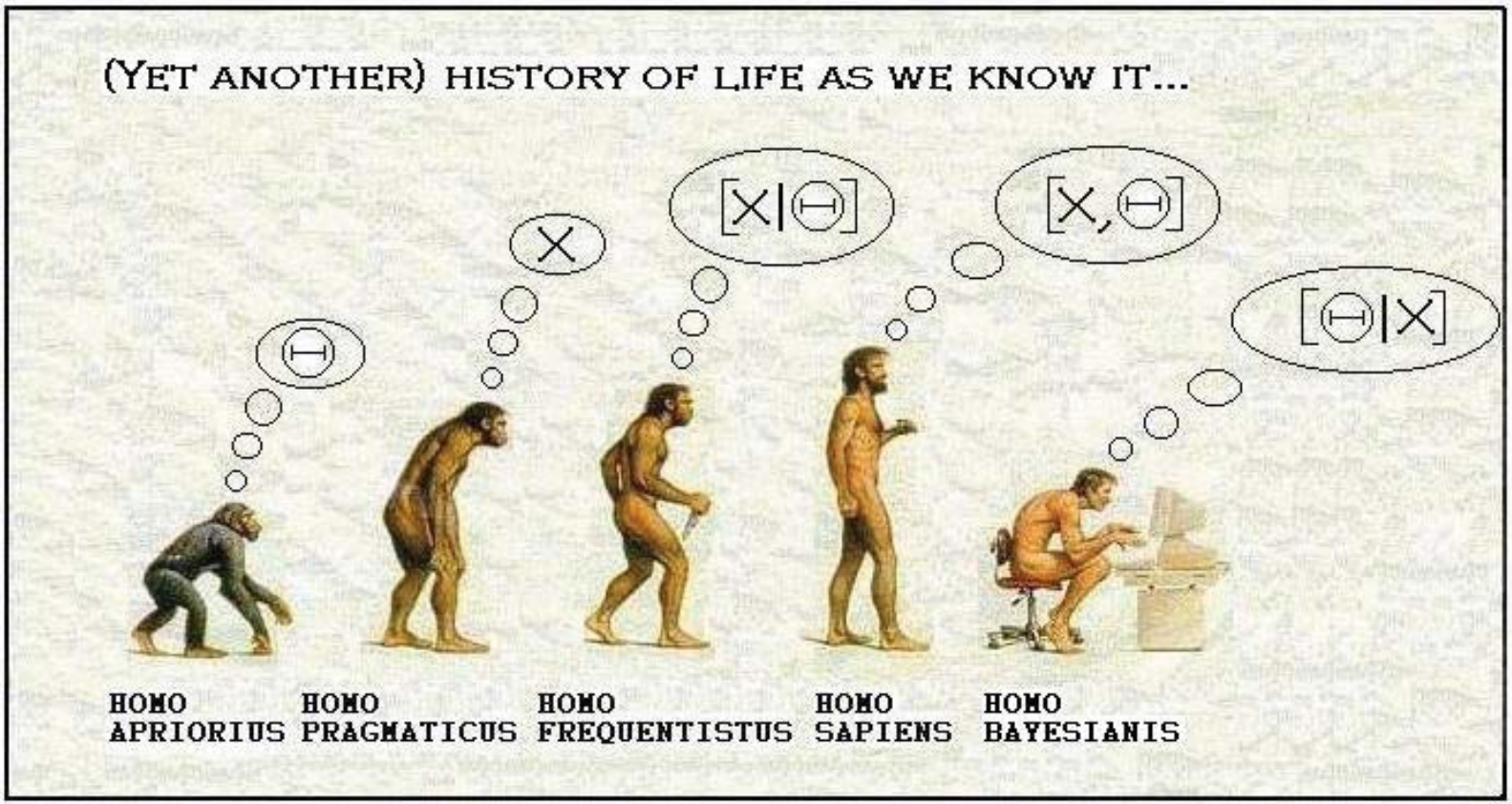

Figure 5: An alternative world view

Clinical trials as experimental designs

Classical principles of experimental design

The three classical principles of experimental design are applicable to clinical trials as well:

- *Local control*
  This is to reduce variability by blocking or stratification by site in multi-center trials
- *Replication*
  Replication of experimental units is used to estimate variability. As obvious as it seems by today's standards, it has been slowly accepted by physicians.
- *Randomization*
  Randomization is a universally accepted method to reduce bias in clinical trials

Clinical trials versus other experimental studies

Clinical trials have unique characteristics compared to other experimental studies. Some of these are:

- The clinical investigator cannot control as much variability as one could in the laboratory
- Experiments on non-human subjects require few constraints
- The clinical investigator cannot usually test all his subjects simultaneously, which leads to long follow-up duration and introduces other sources of variability into the study
- Therapeutic trials make it impossible to test multiple treatments on the same subject due to combined toxicity considerations
- The use of uncertainty in clinical trials is a more fundamental concept than in non-human experiments

Proper design is critical for correct inference

Skilful analysis cannot overcome inappropriate design (we will discuss the need for statistical input up front in the design of the trial later). Proper design:

- Allows investigators to satisfy ethical constraints
- Permits efficient use of resources
- Isolates the treatment effect from confounding effects
- Controls variability and reduces selection and observation bias
- Simplifies and validates the analysis and increases the external validity of the trial

Trial terminology
Drug development

Therapeutic drug development trials are classified into four *phases*.

- *Phase I*
  These are dose-finding studies
- *Phase II*
  These look for evidence of activity, efficacy and safety at a fixed dose (determined by Phase I studies)
- *Phase III*
  In these trials, new treatments, with evidence of safety and efficacy, are compared with alternatives, no therapy or placebo.
- *Phase IV*
  Phase IV studies occur after regulatory approval to look for uncommon side effects

Trial terminology
Cancer prevention

In the context of cancer prevention, the Phase system is different than therapeutic trials

- *Phase II*
  This phase is sub-divided into
  - *Phase IIa* trials, that are small scale feasibility studies using intermediate endpoints (e.g., precursor lesions or biomarkers)
  - *Phase IIb* trials are randomized comparative studies using intermediate endpoints

Trial terminology
Cancer prevention (continued)

- *Phase III*
  These prevention studies employ comparative designs like Phase IIb trials and use definitive endpoints (e.g., cancer incidence)
- *Phase IV*
  These are defined population studies
- *Phase V*
  Phase V are demonstration and implementation trials

Trial terminology
Descriptive terminology

Differences in terminology (e.g., cancer therapeutics versus prevention) highlight the need for a more general, descriptive terminology. One such system is as follows (Piantadosi, 2005 pp. 133):

Developmental	Terminology	
Stage	Old	Descriptive
Early	None	Translational
↓	Phase I	Treatment mechanism, TM
	↓	Dose-finding, DF
		Dose ranging
Middle	Phase II	Safety and activity, SA
↓	Phase IIa	↓
Late	Phase IIb	Comparative, CTE
↓	Phase III	↓
	Phase IV	Expanded safety, ES
	Large simple	Large scale

Classical components of design present in clinical trials

Three classical components of experimental design are applicable in clinical trials. These are *treatment design*, *error control design* and *sampling design*.

Trial Type	Treatment Deisgn	Error Design	Sampling Design
Translational	Simple	Simple	Simple
Dose-finding	Complex	Simple	Simple
SA	Simple	Simple	Moderate
CTE	Moderate	Complex	Moderate
Factorial	Very complex	Complex	Complex
Large scale	MOderate	Moderate	Simple
Crossover	Complex	Moderate	Moderate

Evidence provided by types of medical studies

Byar (1978) provided the following list regarding the types of clinical studies and the evidence they provide:

Study type	Evidence obtained
Case report	Demonstration that some event is possible
Case series	Demonstration of possibly related clinical events
Database analysis	Treatment not determined by experimental design but by physician or patient preference. Data are unlikely to have been collected to assess efficacy.
Observational study	Investigators exploit "natural" exposure or treatment selection and choose comparison group by design.
Controlled clinical trial	Treatment assigned by design. Endpoints actively ascertained and analyses are planned in advance.
Replicated clinical trials	Independent verification of efficacy estimates.

Developmental trial designs – Early development
Translational and mechanistic studies

Translational trials refer to the communication of ideas and treatments from the laboratory to the clinical field.

Mechanistic trials have as their goal to understand the drug mechanism. In addition, early studies are concerned with the relationship between dose and safety (dose-finding trials), using pharmacokinetic and dose-response models.

Developmental trial designs – Middle develolpment

SA trials of feasibility and treatment effect estimation

Middle development studies address the issues of "tolerability" (feasibility, safety and activity) to generate risk-benefit statements. These involve typically 25-30 patients. The classic design features of SA trials are as follows:

- Focused eligibility
- Fixed dose or treatment algorithm
- Single cohort compared to external reference standard
- Use of surrogate clinical outcomes to shorten trial time
- Modest number of patients
- Explicit decision parameters about continued development

Developmental trial designs
Skipping the SA step

Sometimes the middle development step may be skipped. However this has considerable risks and a number of points should be considered before abandoning the middle-development step:

- The sponsor must be willing to accept the financial risk
- The risk of a long comparative study with null results must be acceptable
- There should be enough information regarding safety and the likelihood of efficacy
- Compelling biological rationale
- Calendar time and the potential for a lost opportunity can make this choice compelling
- The chance of unforeseen events must be low
- The cost of middle development must high relative to the information gained

Developmental trial designs – Late development
Comparative studies

Comparative studies employ a concurrent comparison group (internal control) and are designed to provide precise estimates of treatment difference. CTE trials correspond to the Phase III part of the original classification.

CTE trials may be carried out in a very large scale (e.g., cancer trials that are designed to assess a very small treatment effect).

CTE trials employ fixed sample size, staged or fully sequential designs or crossover designs. They are most frequently performed by multiple institutions simultaneously.

Developmental trial designs

ES (Phase IV) studies of uncommon treatment effects

Even after regulatory approval there is still time to learn about uncommon side effects, drug interactions and unusual complications. Phase IV studies target unusual side effects but may not precisely determine the exact number of persons receiving treatment.

Infrequently, these studies result in the removal of an approved drug from the market. There are 10 such cases since 1974 where approved drugs were removed due to safety concerns.

Documents necessary to run a clinical trial

There are a number of documents that are produced in order to direct the conduct of a clinical trial. These are:

- *The protocol*
  This is the main document that describes both the conceptual and logistical aspects of the trial
- *The concept sheet*
  Before putting together a protocol, a summary statement can be created that can prepare the investigators for the structure and content of the final protocol document and communicate basic scientific aspects of the trial
- *Investigator's brochure*
  This describes the investigational products or device and contains documentation of ethical review, case report or data forms and other regulatory documents

Components of the protocol

1. Title page
2. Contents/index (optional)
3. Synopsis (optional)
4. Schema (essential for complex SA or CTE protocols)
5. Objectives of the study
6. Introduction and background
7. Study design (essential)
8. Drug information
9. Staging criteria (essential for all studies)
10. Patient eligibility and exclusion criteria (essential for all protocols)

Components of the protocol (continued)

11. Registration, randomization procedures and stratification (essential for all protocols)
12. Treatment program (essential for all protocols)
13. Dose modification/side effects (essential for all protocols)
14. Agent information (essential for all protocols)
15. Treatment evaluation (essential for all protocols)
16. Adverse events and toxicity management (optional as a separate section)
17. Serial measurements/study calendar (essential for SA and CTE studies)
18. Statistical considerations (essential for DF, SA and CTE studies)

Components of the protocol (continued)

19. External collaborations or reviews (essential for all multi-institutional protocols)
20. Data recording, management and monitoring (essential for all protocols)
21. Special instructions
22. Communication and publication data (essential for all studies which involve collaborations, especially data management, outside of the institution, particularly pharma-sponsored studies)
23. Peer review (optional)
24. Patient consent (essential for all protocols)
25. References (essential for all protocols)
26. Data (case report) forms (CRF) (optional)

Components of the protocol (continued)

27. Protocol amendments (essential for all protocols if any amendments exist)
28. Other appendices (optional)
29. Glossary (optional)

Random error and bias

Random error and bias

Error has two components:

- A a random component (random fluctuations beyond the ability of the investigator to explain)
- A systematic component (bias), or differences that are not a product of chance alone

Understanding their difference is the first step in controlling them through experimental designs.

Random error
Hypothesis tests

Hypothesis testing is approach for choosing between two competing hypotheses labeled H_o and H_α by use of a summary statistic T.

Hypothesis testing requires the definition of a *critical region* in advance and then choose H_α or H_o depending on whether T falls within the critical region or not respectively.

Pictorial representation of hypothesis testing

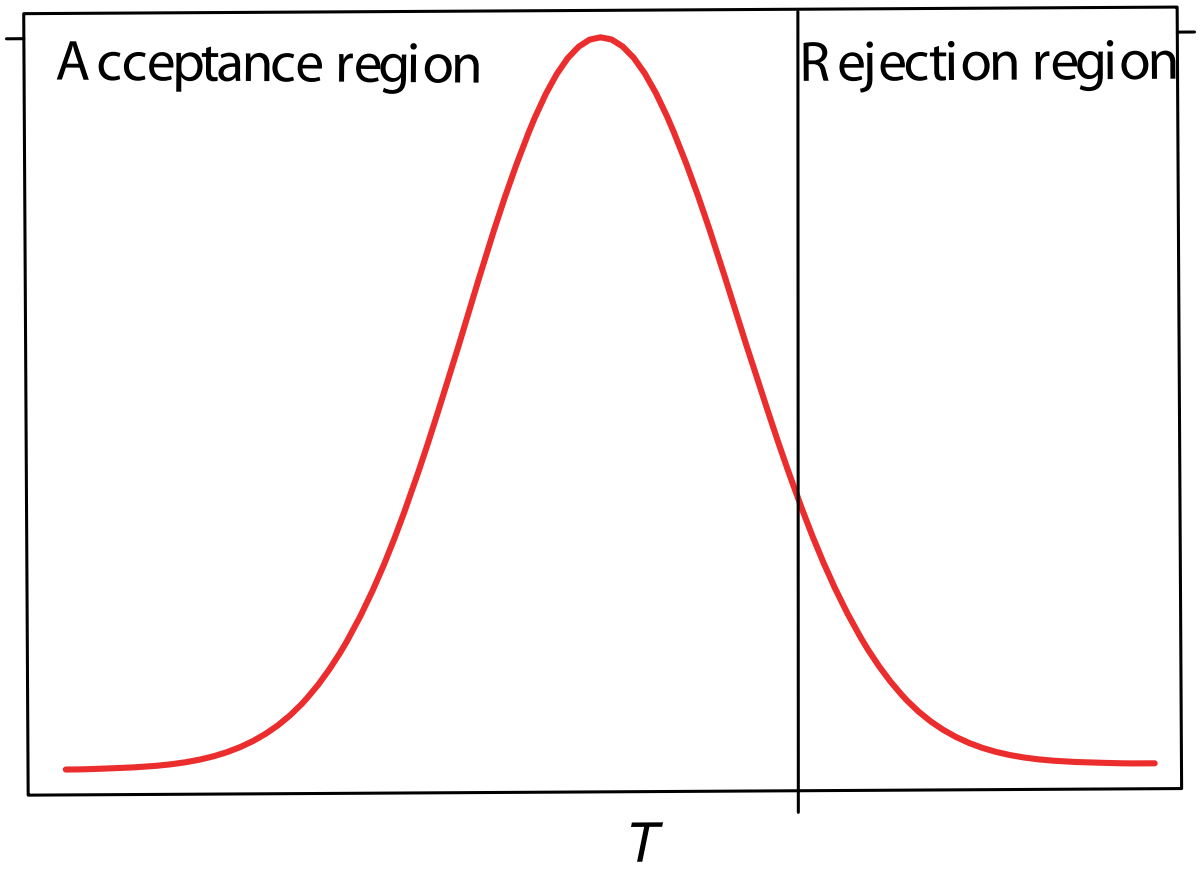

Figure 6: Acceptance and rejection region for a statistic T

Random error
Significance tests

In practice investigators seldom test hypotheses with a hypothesis test described as above.

Instead, a procedure called *significance testing* is used, based on the probability distribution of T if H_o is assumed true. The more extreme T is, relative to this distribution, the less likely H_o is to be true. The significance level is

$$p = Pr\{T^* \geq T | H_o\}$$

where T is the value of the statistic based on observed data.

This test yields a significance level or p value instead of a decision.

Types of error in tests of hypothesis

Hypothesis tests are affected by two types of random error:

Table 1: Random errors from hypothesis tests and other dichotomized inferences

	Truth and consequences	
Result of test	H_o	H_α
Reject H_o	Type I error	No error
Reject H_α	No error	Type II error

Pictorial presentation of error types

Pictorially the four situations shown in the previous table are shown in the following figure:

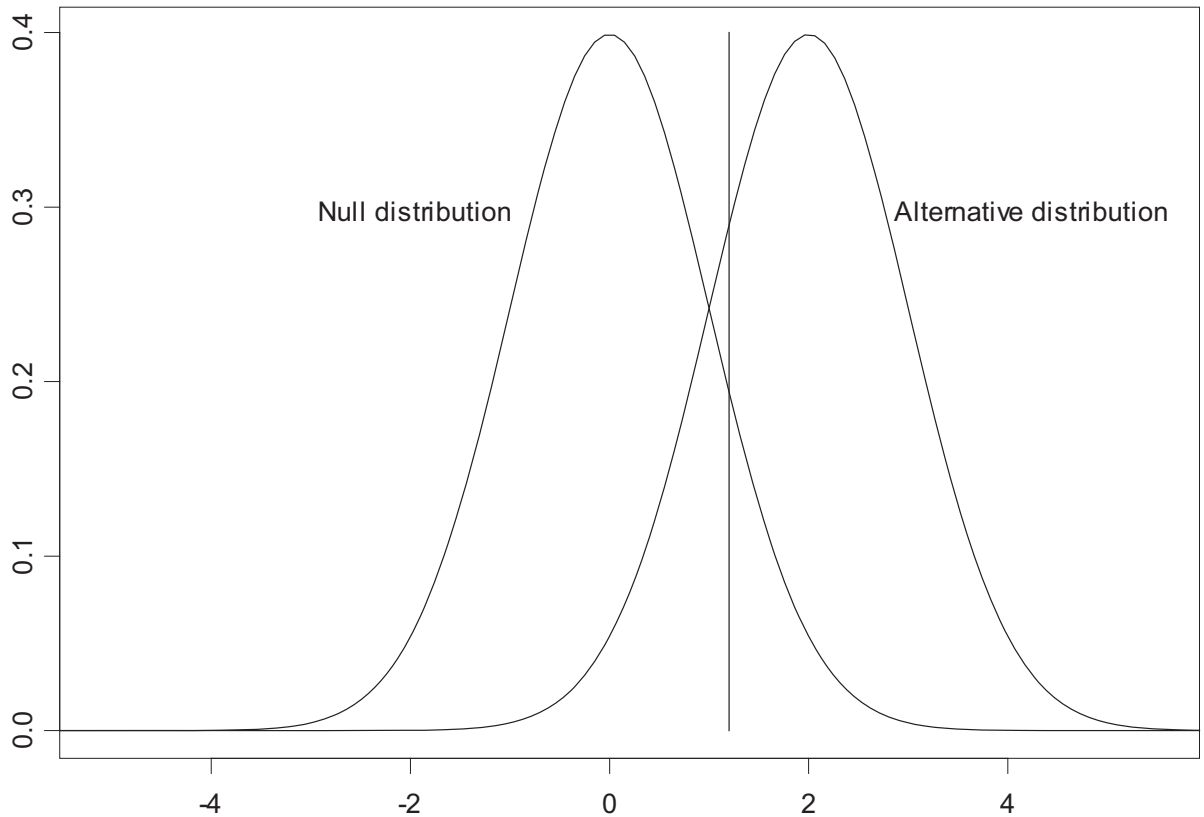

Figure 7: Distribution of T under the null and alternative hypothesis and the resultant Type I and Type II errors

Confidence intervals

Performing testing using point estimates and confidence intervals gives more information than merely using hypothesis tests.

Confidence intervals are centered on teh observed or estimated effect and convey important information regarding the precision of the estimate. We can usually reconstruct hypothesis tests using confidence intervals.

To see how this is possible, consider a confidence interval around the observed quantity Δ. If this confidence interval does not cover the hypothesized (under H_o) value Δ_h, then we would reject this hypothesis. If this is erroneous (i.e., the sample was atypical) corresponds to a Type-I error situation.

Characteristics of Type-II error

While the Type-I error is easier to control as it is based on a single factor (the significance of the test), there are three factors that influence the Type II error:

- The critical value for the rejection of the null hypothesis
- The variability of the estimator under the alternative
- The distance between the centers of the null and the alternative hypothesis

Investigators have control of the former and the second factors (through manipulation of the sample size). However, the magnitude of the alternative hypothesis (distance from the null) is out of investigator control.

Pictorial representation of the test of two means

This is shown pictorially in the following Figure for a t test of two means carried out at the 5% alpha level.

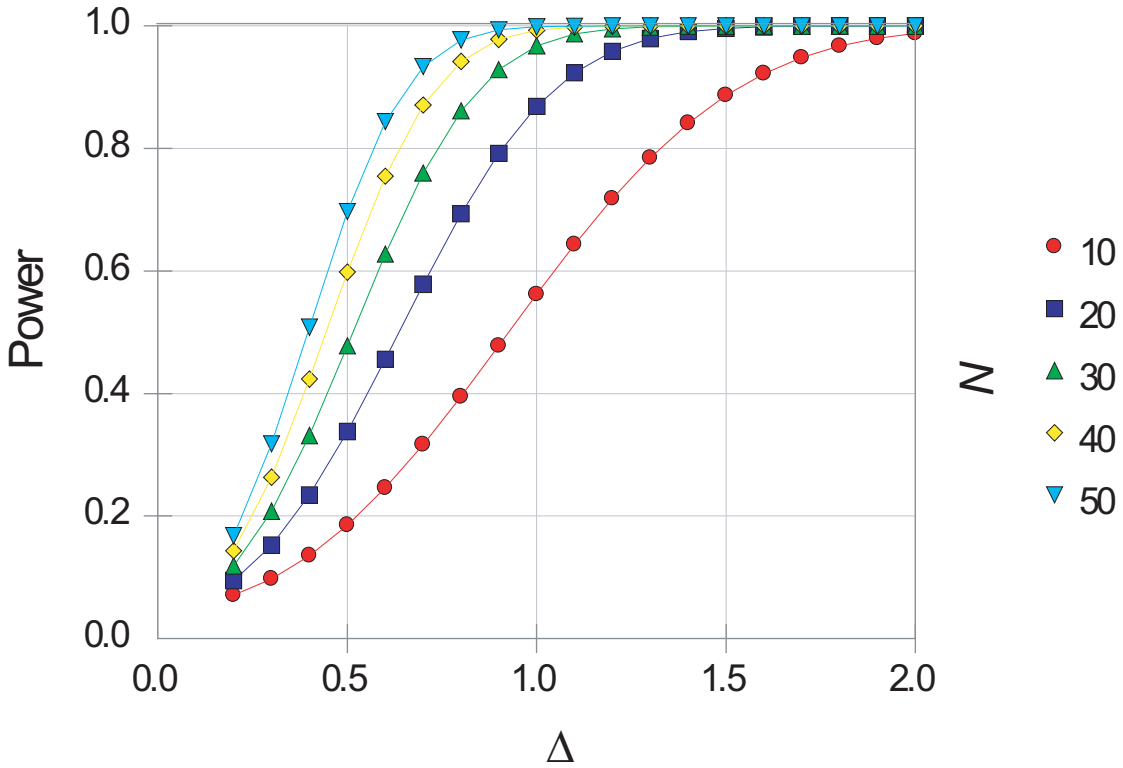

Figure 8: Power curves for progressively larger effect size ($\Delta = \frac{\mu_0 - \mu_\alpha}{\sigma}$)

Post-hoc power calculations

At the conclusion of a trial we know the estimated clinical effect and its variability. In many situations, it has been advocated that a ("post-hoc) power calculation be undertaken, ostensibly to assess the strength of the results. This is a practice that is "fundamentally flawed" (Hoening & Heisey, *The American Statistician*, 2001)

The problem lays in the fact that post-hoc (or "observed power" as the authors put it) is completely determined by the significance level (the p value) and, computing the observed power adds nothing to the interpretation of the results. In other words, non-significant p values are always associated with low observed power values.

A good analogy was made by Lenth (*J Animal Sci*, 2001): "If my car made it to the top of a hill [null hypothesis was rejected], then it is powerful enough to climb that hill; if it didn't make it to the top of that hill [null was not rejected], then it obviously isn't powerful enough."

Clinical bias

When bias is present, its magnitude, compared to random error is important. Bias arises from a number of sources:

Table 2: Sources of bias

Type of bias	Potential consequences
Selection bias	External validity, especially with no internal controls
Procedure selection	Healthier patients may improve results for that treatment
Postentry exclusion	Overoptimistic view of treatment, especially its safety
Selective loss of data	Variety of effects depending on the mechanism of loss
Ascertainment of bias	Treatment effect is brought in line with prior expectation
Uncontrolled covariate	Confounding of treatment effect with covariate effect

Controlling structural bias

The following methods may reduce structural bias in a study:

- *Randomization*
  It is the only method to reduce selection bias and control for unmeasured covariates.
- *Blocking and stratification*
  These methods control for the effects of known covariates by balancing them within treatment groups.
- *Masking*
  Masking (blinding) reduces assessment bias. Single masking is when the patient does not know the treatment, double masking (double blind) is when neither the patient or treating physician know and a triple masking has been proposed to include the DSMB members that assess the progress of the trial (these are usually given hidden treatment codes).

Controlling structural bias (continued)

- *Concurrent controls*
  It eliminates the confounding with calendar time (as would happen if historical controls were used) and facilitates randomization
- *Objective assessments*
  Using objective assessments reduces assessment bias and improves the reproducibility of the study
- *Active follow-up and endpoint ascertainment*
  If trials rely on passive endpoint reporting this may result in bias (e.g., because of lost-to-follow-up issues).
- *No post-hoc exclusions*
  Post-hoc exclusions may increase in selection bias because of unobserved correlation between the selection factor and the outcome of interest. In addition removing subsets from the study undermines the validity of the randomization.

Objectives and endpoints

Objectives versus endpoints

The *objectives* of a trial are the goals of the study both broadly in terms of the development process or specifically in terms of the foci of the study itself.

The *endpoints* or specific objectives of a trial are the measurable outcomes of the study subjects and should not be confused with the general objectives of the trial.

For example, a general objective is to decide whether a treatment is safe and effective. An objective particular to the specific study may be to decide whether the treatment extends survival compared to another treatment. A specific objective (or endpoint) may be to estimate the relative hazard of all-cause mortality in teh two treatment groups based on the survival outcomes of the participating subjects.

Considerations for evaluating and selecting endpoints

"Hard" or objective endpoints should be preferred. Examples include death, relapse or progression, and laboratory measurements. Examples of "soft" outcomes are those that are subjective and do not fulfill the criteria listed below for selection and evaluation of endpoints:

Table 3: Considerations for endpoint evaluation and selection

Characteristic	Meaning
Relevant	Clinically important/useful
Quantifiable	Scored on an appropriate scale
Valid	Measures the intended effect
Objective	Interpreted the same by all observers
Reliable	Same effect yields consistent measurements
Sensitive	Responds to small changes in the effect
Specific	Unaffected by extraneous influences
Precise	Has small variability

Types of data

The following are types of data used in clinical trials

Table 4: Types of data used in clinical trials

ID Number	Age	Sex	Toxicity Grade	Age Rank	Sex Code
1	41	M	3	1	0
2	53	F	2	5	1
3	47	F	4	2	1
4	51	M	1	4	0
5	60	F	1	9	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
Scale:	Ratio	Category	Ordinal	Interval	Nominal

The special case of event data

In event data, the *time* from an starting point (usually enrollment or start of treatment) until the observation of the event of interest (event) or administrative end of the study (censoring) is measured along with a zero/one code corresponding to the former (event=1) or latter (censor=0) situation.

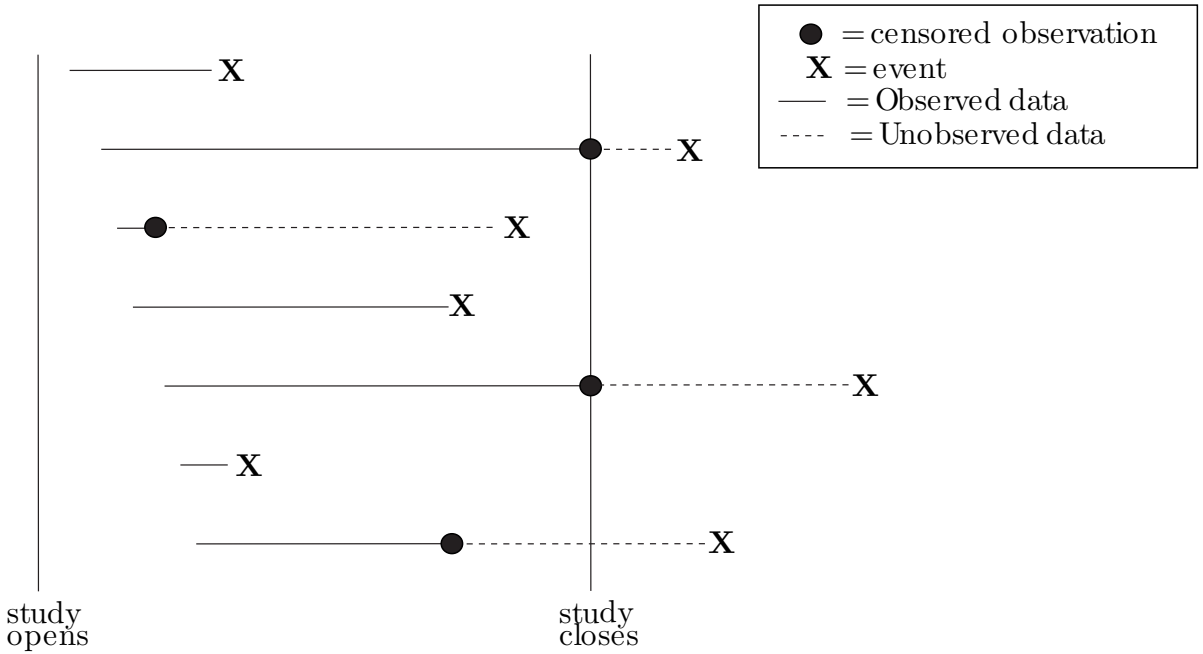

Figure 9: Acceptance and rejection region for a statistic T

Censoring versus lost to follow-up

These two concepts are very easy to confuse but these two concepts are not the same.

Censoring is the inability to observe the event in a fraction of the subjects because the study was completed prior to observing all the events (e.g., not all subjects had died by the end of the study).

Lost to follow-up means that the event cannot be determined even if the study were to be extended or with active follow-up (see third subject from the top in the previous Figure; this subject was lost to follow-up because his event status was not determined despite having disappeared prior to the end of the study).

Data analyses in the presence of LTFU

Data from subjects that were lost to follow-up may be analyzed by censoring them at the time of loss to follow-up.

This would not be appropriate if the loss to follow-up is somehow associated with the outcome (survival), i.e., the observed data are not representative of the not observed data (e.g., because more sicker patients tend to drop out).

Surrogate outcomes

A *surrogate outcome* is one that is measured in place of the biologically definitive outcome. Typically, the definitive outcome measures the clinical benefit, while the surrogate tracks the process and extent of the disease. A rigorous definition of a surrogate outcome was given by Prentice (*Stat Med*, 1989) as

a response variable for which a test of the null hypothesis of no relationshp to the treatment groups under compoarison is also a valid test of the corresponding null hypothesis based on the true endpoint.

We must note that simple correlation between the surrogate and definitive outcome is not a sufficient condition to establish surrogacy. In addition, the surrogate outcome must reflect the effect of treatment on the definitive endpoint.

Types of surrogate outcomes

Consider the following situations shown pictorially on the Figure of the next slide:

1. The relationship of the disease process with the surrogate outcome is completely separate from the definitive outcome. This will not be a surrogate outcome
2. The surrogate outcome lies in the causal pathway of the disease, but there is also an alternative pathway that could invalidate the treatment effect as summarized by the surrogate outcome
3. This is perfect surrogacy as the surrogate outcome summarizes the entire treatment effect on the true outcome.

Pictorial representation of surrogacy

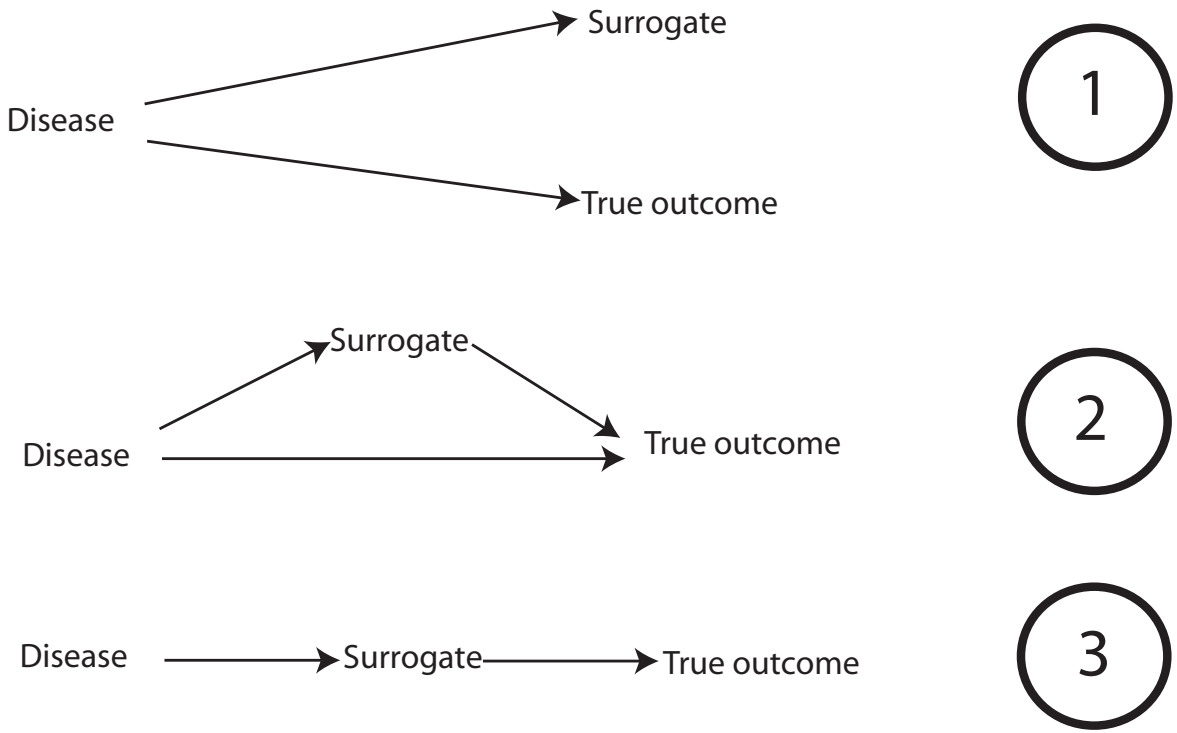

Figure 10: Possible relationships between surrogate, definitive outcomes and treatment effects

Read the article by Choi et al., (*Ann Int Med*, 1993) for a case of an incomplete surrogate marker of HIV and general discussion of surrogate markers.

Surrogate endpoints are disease specific

Surrogate endpoints are disease-specific because they are dependent on the mechanism of action of the treatment under investigation.

	Definitive endpoint	Surrogate endpoint
HIV Infection	AIDS (or death)	Viral load
Cancer	Mortality	Tumor shrinkage
Colon cancer	Disease progression	CEA level
Prostate cancer	Disease progression	PSA level
Cardiovascular disease	hemorrhagic stroke, myocardial infarction	Blood pressure, cholesterol level
Glaucoma	Vision loss	Intraocular pressure