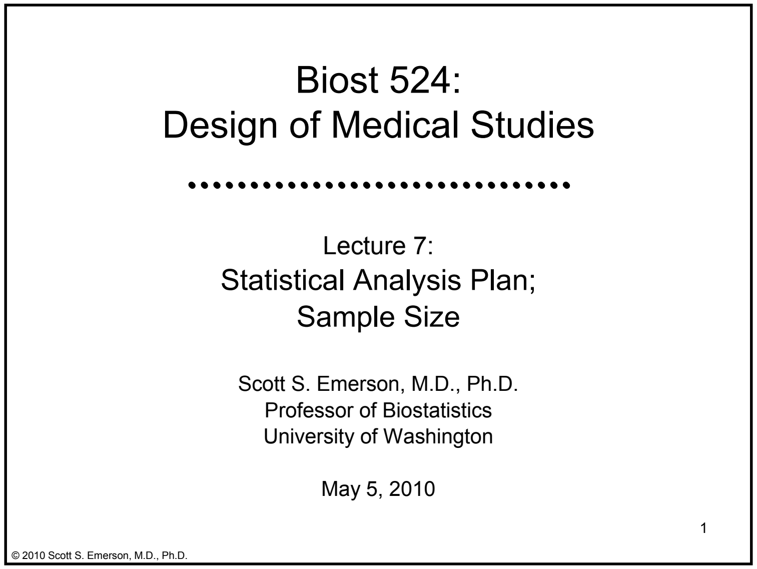
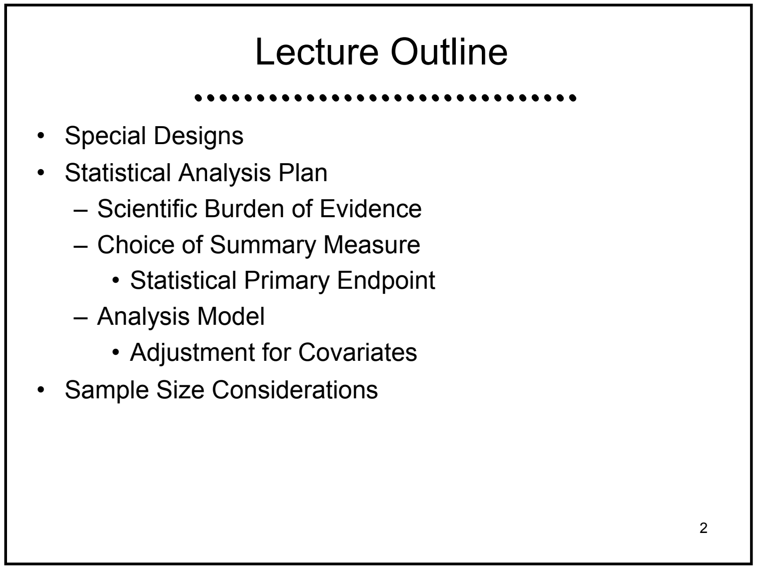


Biost 524: Design of Medical Studies

Lecture 7:
Statistical Analysis Plan;
Sample Size

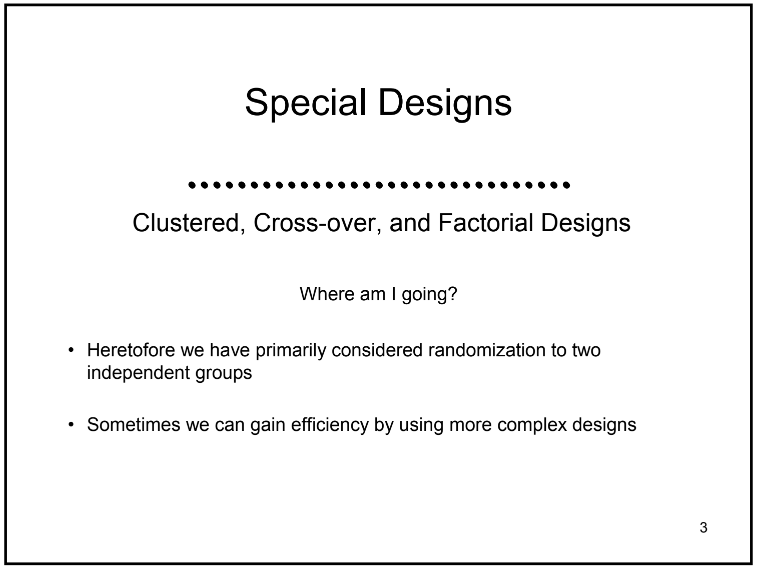
Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

May 5, 2010



Lecture Outline

- Special Designs
- Statistical Analysis Plan
  - – Scientific Burden of Evidence
  - – Choice of Summary Measure
    - Statistical Primary Endpoint
  - – Analysis Model
    - Adjustment for Covariates
- Sample Size Considerations

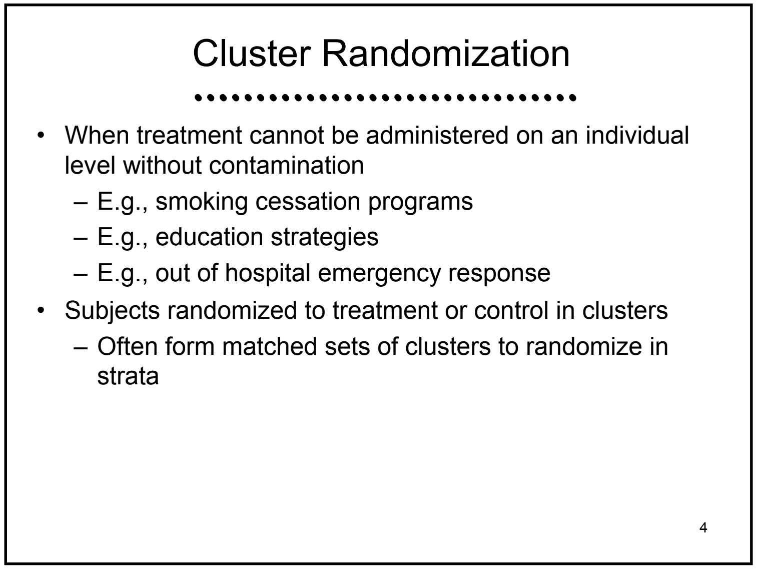


Special Designs

Clustered, Cross-over, and Factorial Designs

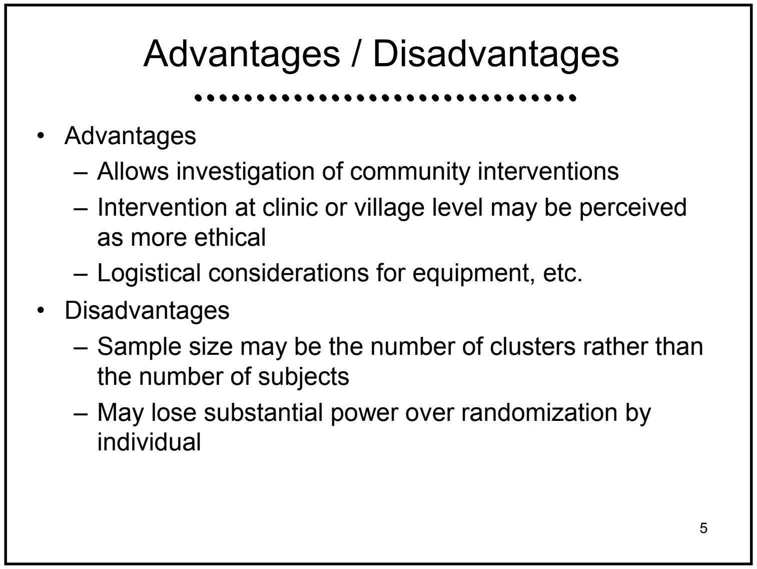
Where am I going?

- Heretofore we have primarily considered randomization to two independent groups
- Sometimes we can gain efficiency by using more complex designs



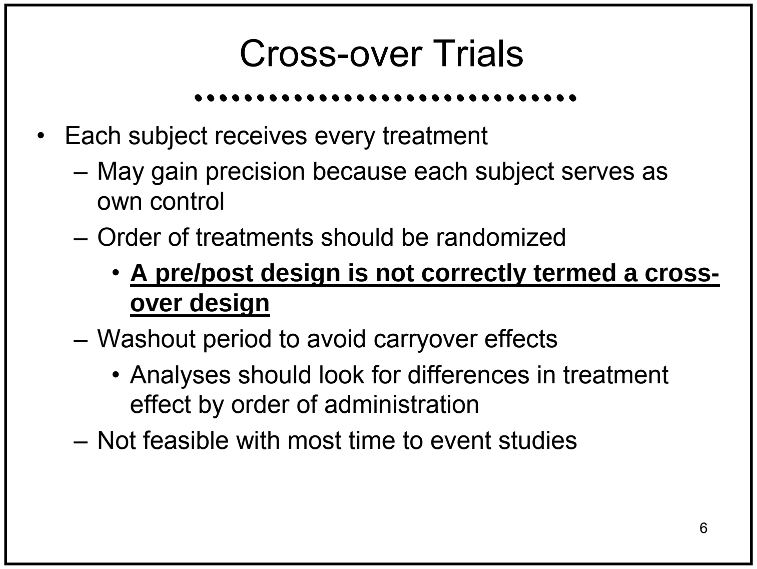
Cluster Randomization

- When treatment cannot be administered on an individual level without contamination
  - – E.g., smoking cessation programs
  - – E.g., education strategies
  - – E.g., out of hospital emergency response
- Subjects randomized to treatment or control in clusters
  - – Often form matched sets of clusters to randomize in strata



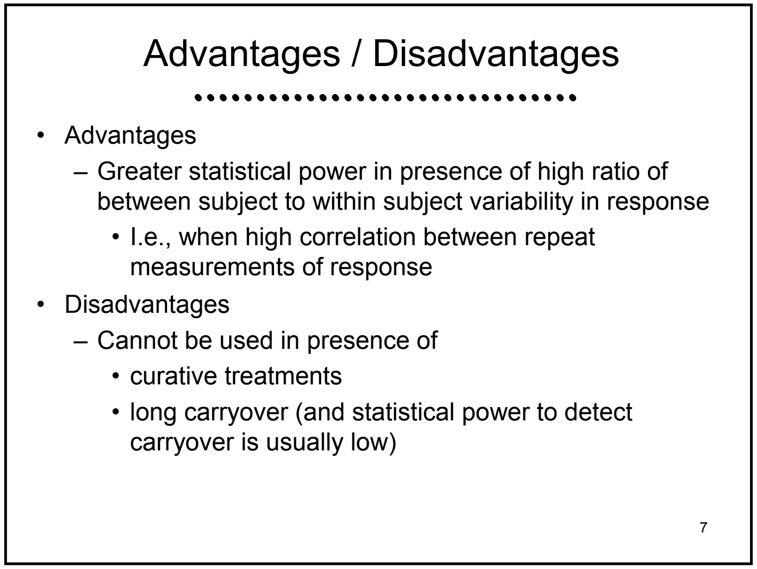
Advantages / Disadvantages

- Advantages
  - Allows investigation of community interventions
  - Intervention at clinic or village level may be perceived as more ethical
  - Logistical considerations for equipment, etc.
- Disadvantages
  - Sample size may be the number of clusters rather than the number of subjects
  - May lose substantial power over randomization by individual



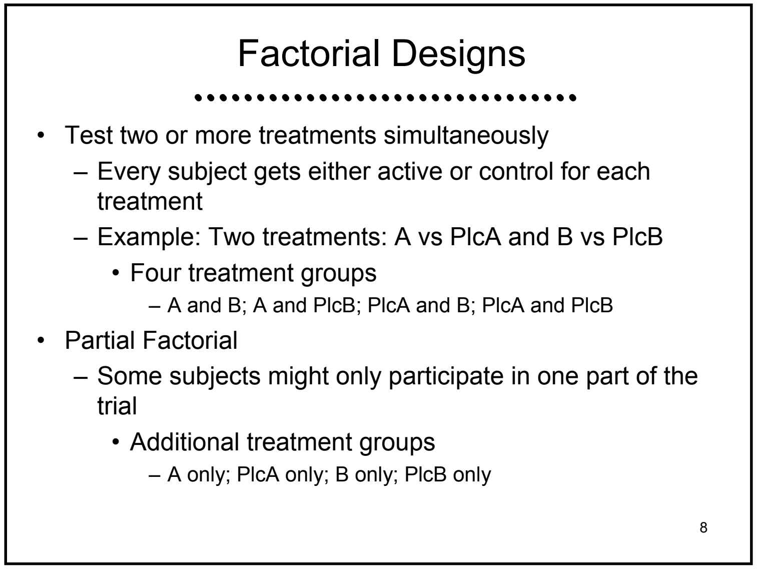
Cross-over Trials

- Each subject receives every treatment
  - May gain precision because each subject serves as own control
  - Order of treatments should be randomized
    - **A pre/post design is not correctly termed a cross-over design**
  - Washout period to avoid carryover effects
    - Analyses should look for differences in treatment effect by order of administration
  - Not feasible with most time to event studies



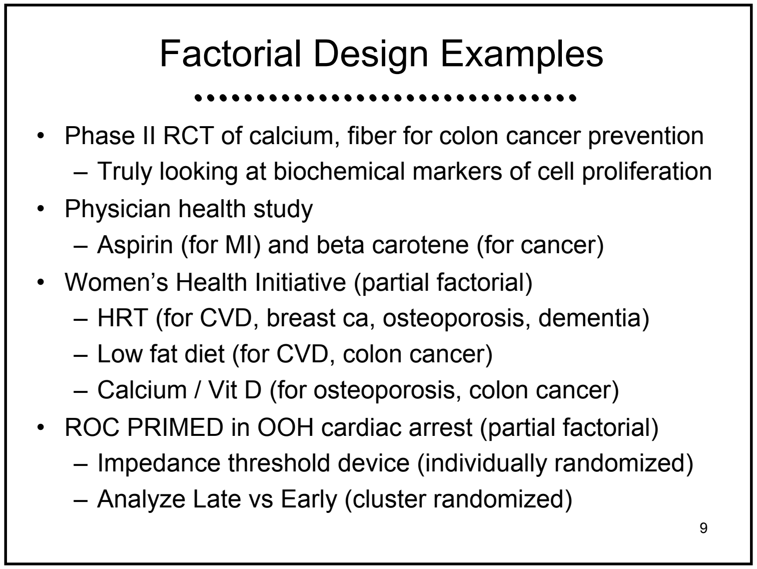
Advantages / Disadvantages

- Advantages
  - Greater statistical power in presence of high ratio of between subject to within subject variability in response
    - I.e., when high correlation between repeat measurements of response
- Disadvantages
  - Cannot be used in presence of
    - curative treatments
    - long carryover (and statistical power to detect carryover is usually low)



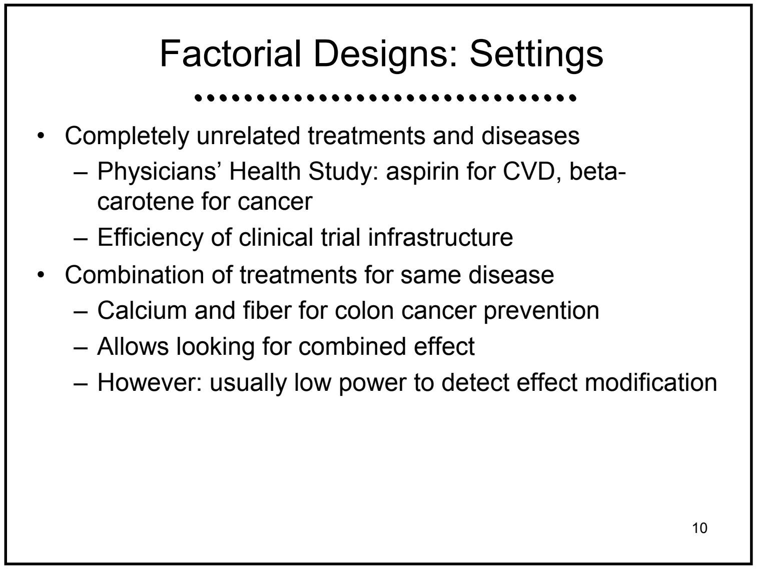
Factorial Designs

- Test two or more treatments simultaneously
  - Every subject gets either active or control for each treatment
  - Example: Two treatments: A vs PlcA and B vs PlcB
    - Four treatment groups
      - A and B; A and PlcB; PlcA and B; PlcA and PlcB
- Partial Factorial
  - Some subjects might only participate in one part of the trial
    - Additional treatment groups
      - A only; PlcA only; B only; PlcB only



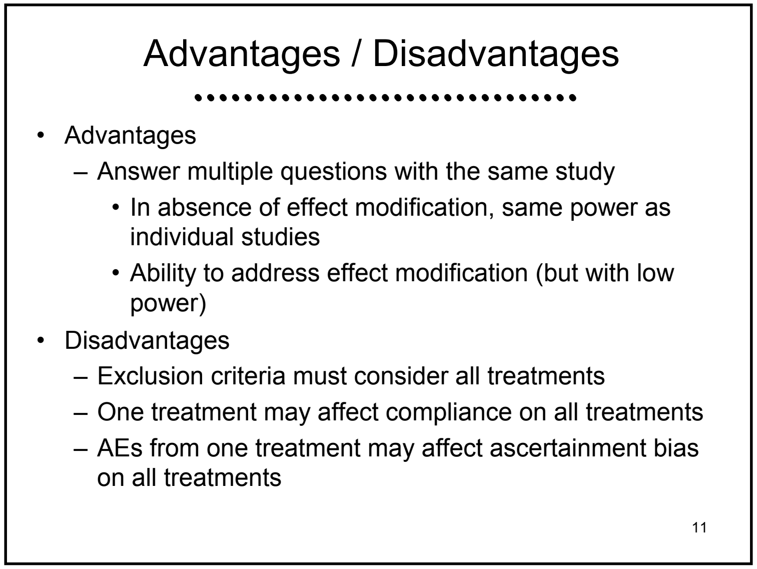
Factorial Design Examples

- Phase II RCT of calcium, fiber for colon cancer prevention
  - – Truly looking at biochemical markers of cell proliferation
- Physician health study
  - – Aspirin (for MI) and beta carotene (for cancer)
- Women's Health Initiative (partial factorial)
  - – HRT (for CVD, breast ca, osteoporosis, dementia)
  - – Low fat diet (for CVD, colon cancer)
  - – Calcium / Vit D (for osteoporosis, colon cancer)
- ROC PRIMED in OOH cardiac arrest (partial factorial)
  - – Impedance threshold device (individually randomized)
  - – Analyze Late vs Early (cluster randomized)



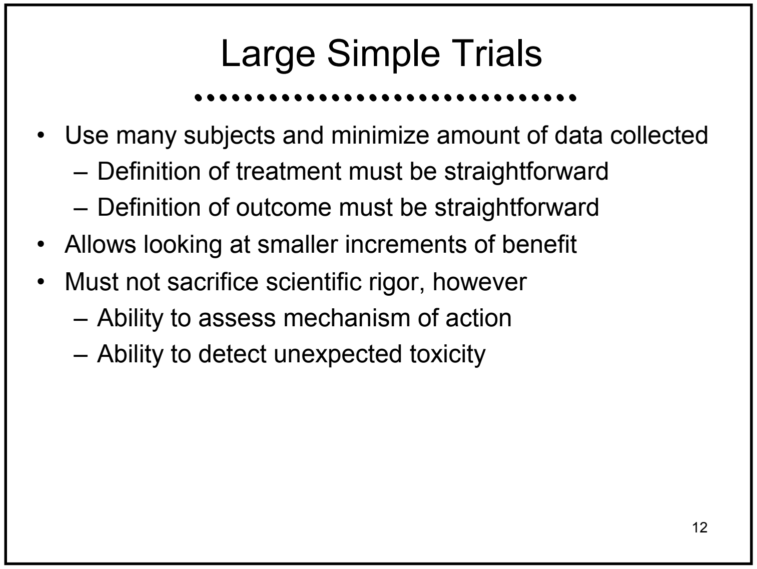
Factorial Designs: Settings

- Completely unrelated treatments and diseases
  - – Physicians' Health Study: aspirin for CVD, beta-carotene for cancer
  - – Efficiency of clinical trial infrastructure
- Combination of treatments for same disease
  - – Calcium and fiber for colon cancer prevention
  - – Allows looking for combined effect
  - – However: usually low power to detect effect modification



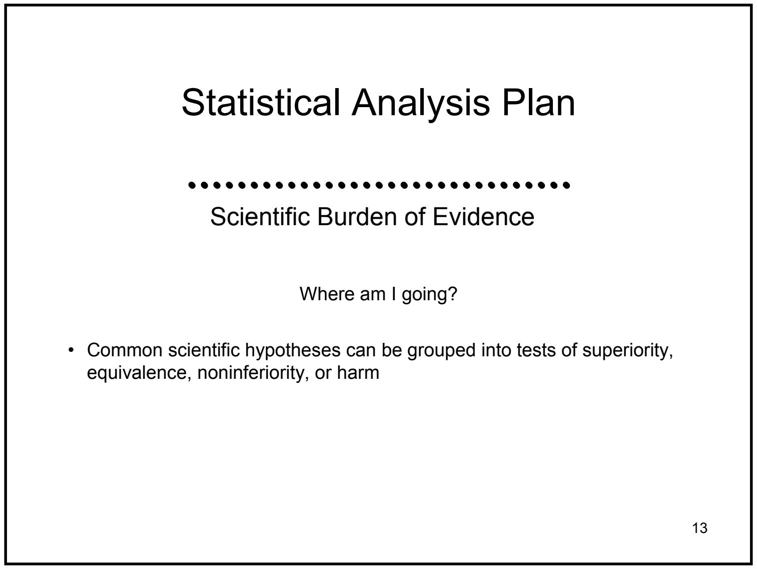
Advantages / Disadvantages

- Advantages
  - – Answer multiple questions with the same study
    - In absence of effect modification, same power as individual studies
    - Ability to address effect modification (but with low power)
- Disadvantages
  - – Exclusion criteria must consider all treatments
  - – One treatment may affect compliance on all treatments
  - – AEs from one treatment may affect ascertainment bias on all treatments



Large Simple Trials

- Use many subjects and minimize amount of data collected
  - – Definition of treatment must be straightforward
  - – Definition of outcome must be straightforward
- Allows looking at smaller increments of benefit
- Must not sacrifice scientific rigor, however
  - – Ability to assess mechanism of action
  - – Ability to detect unexpected toxicity



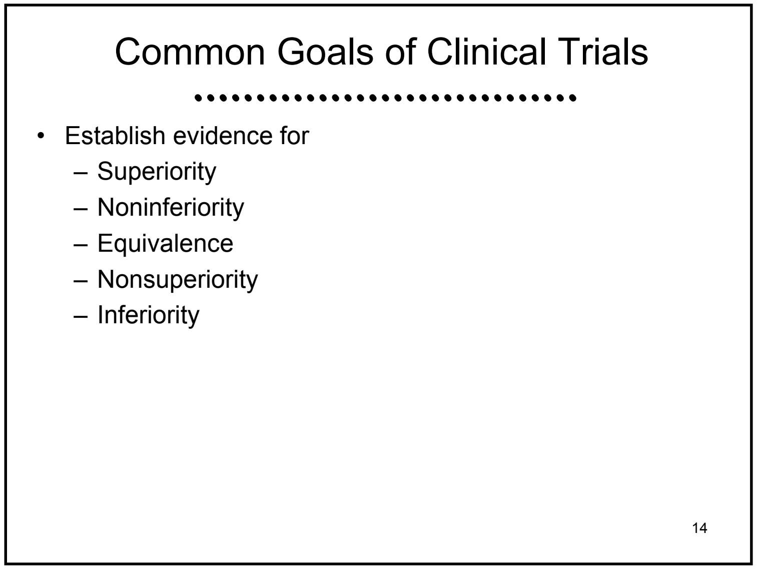
Statistical Analysis Plan

Scientific Burden of Evidence

Where am I going?

- Common scientific hypotheses can be grouped into tests of superiority, equivalence, noninferiority, or harm

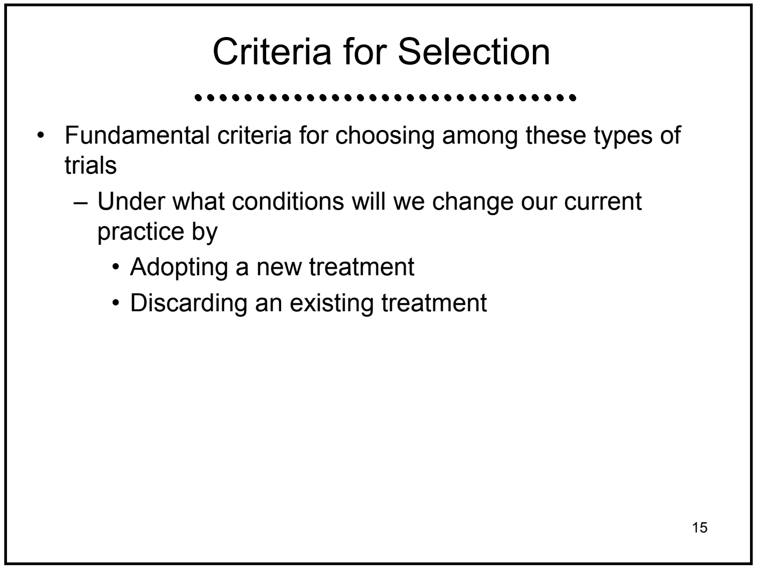
13



Common Goals of Clinical Trials

- Establish evidence for
  - Superiority
  - Noninferiority
  - Equivalence
  - Nonsuperiority
  - Inferiority

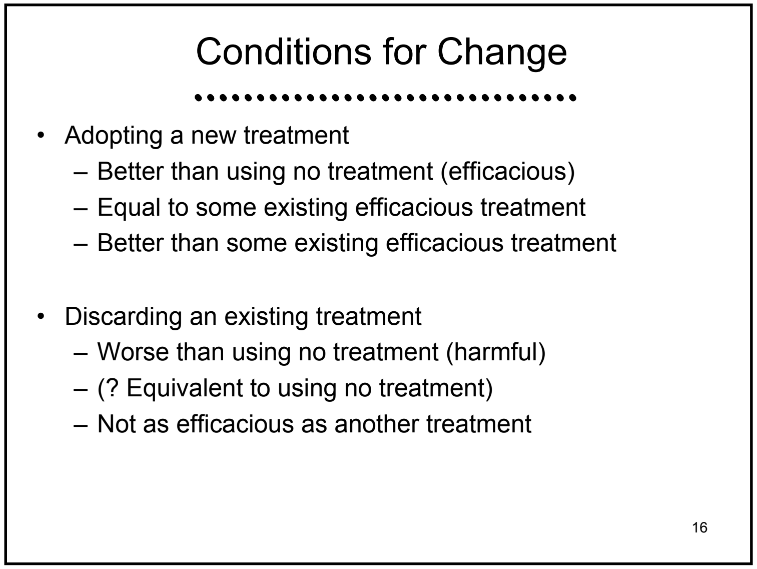
14



Criteria for Selection

- Fundamental criteria for choosing among these types of trials
  - Under what conditions will we change our current practice by
    - Adopting a new treatment
    - Discarding an existing treatment

15

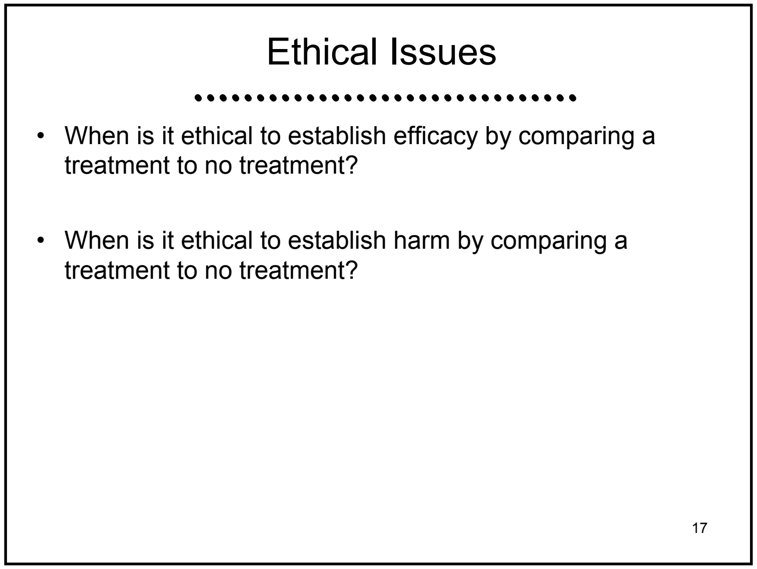


Conditions for Change

- Adopting a new treatment
  - Better than using no treatment (efficacious)
  - Equal to some existing efficacious treatment
  - Better than some existing efficacious treatment

- Discarding an existing treatment
  - Worse than using no treatment (harmful)
  - (? Equivalent to using no treatment)
  - Not as efficacious as another treatment

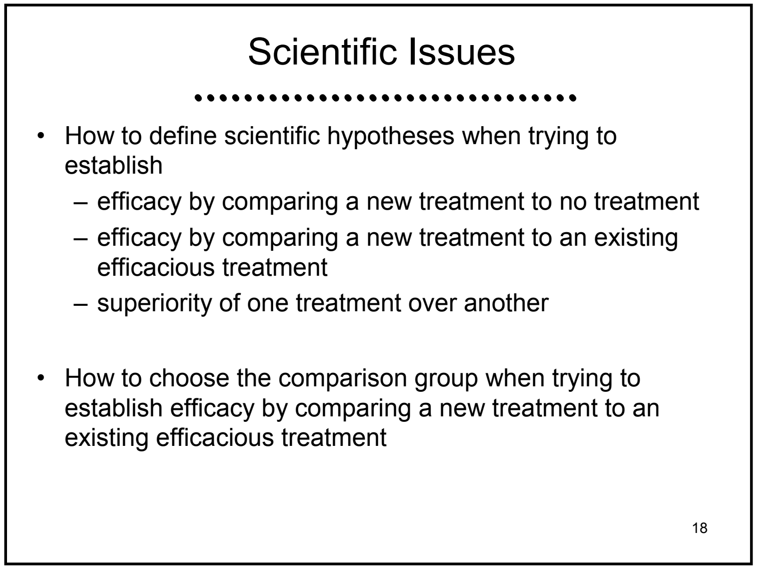
16



Ethical Issues

- When is it ethical to establish efficacy by comparing a treatment to no treatment?

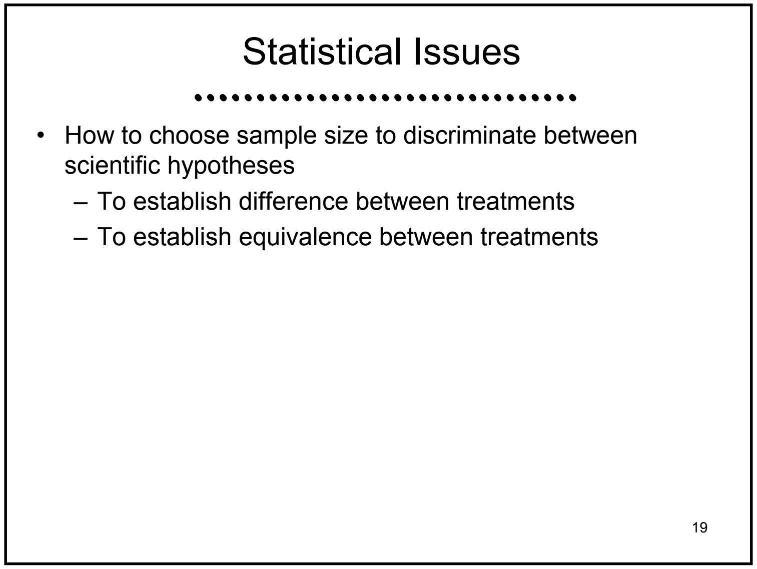
- When is it ethical to establish harm by comparing a treatment to no treatment?



Scientific Issues

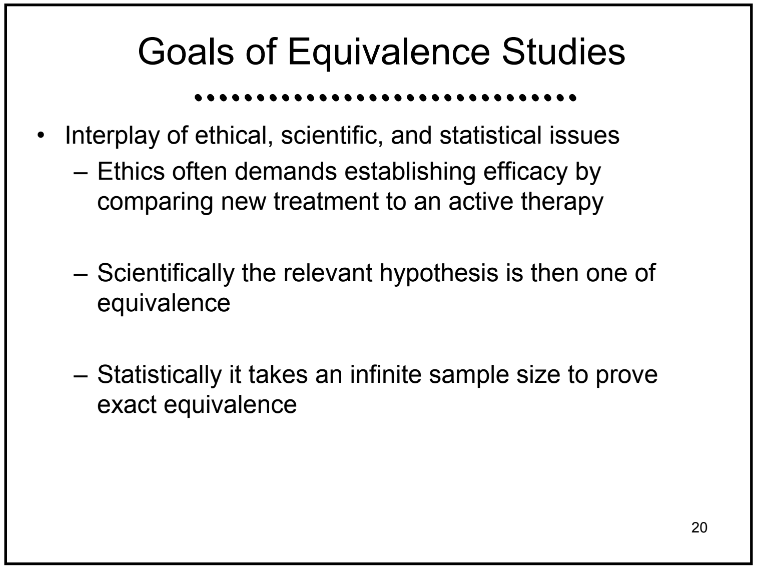
- How to define scientific hypotheses when trying to establish
  - efficacy by comparing a new treatment to no treatment
  - efficacy by comparing a new treatment to an existing efficacious treatment
  - superiority of one treatment over another

- How to choose the comparison group when trying to establish efficacy by comparing a new treatment to an existing efficacious treatment



Statistical Issues

- How to choose sample size to discriminate between scientific hypotheses
  - To establish difference between treatments
  - To establish equivalence between treatments

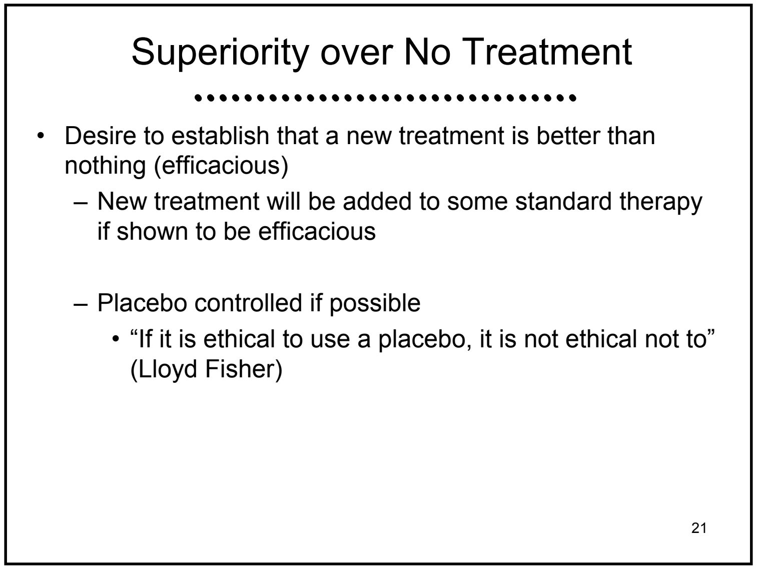


Goals of Equivalence Studies

- Interplay of ethical, scientific, and statistical issues
  - Ethics often demands establishing efficacy by comparing new treatment to an active therapy

  - Scientifically the relevant hypothesis is then one of equivalence

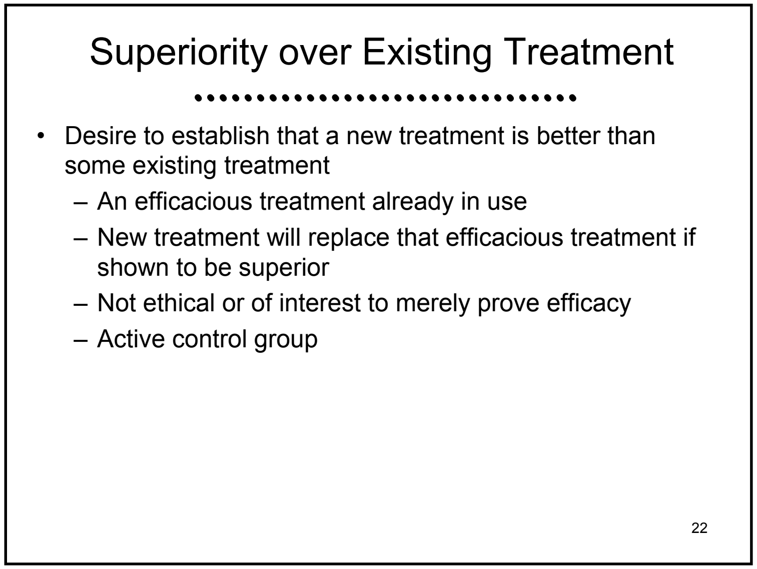
  - Statistically it takes an infinite sample size to prove exact equivalence



Superiority over No Treatment

- Desire to establish that a new treatment is better than nothing (efficacious)
  - New treatment will be added to some standard therapy if shown to be efficacious
  - Placebo controlled if possible
    - “If it is ethical to use a placebo, it is not ethical not to” (Lloyd Fisher)

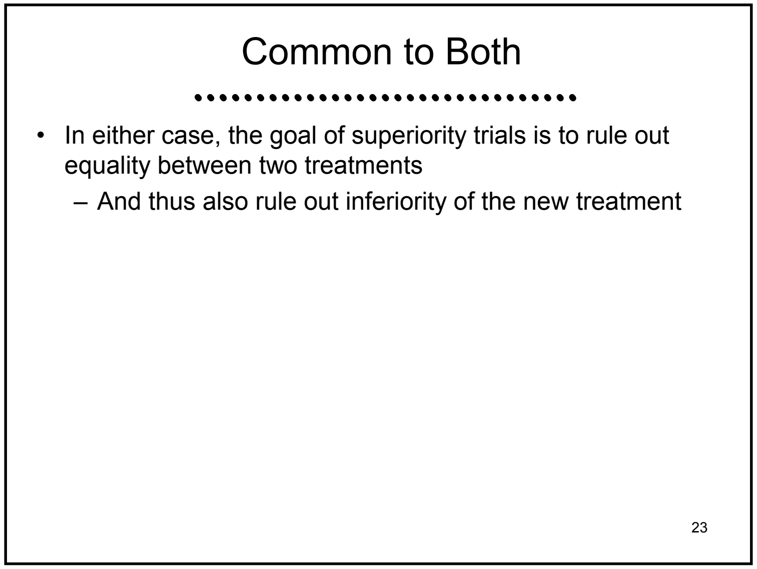
21



Superiority over Existing Treatment

- Desire to establish that a new treatment is better than some existing treatment
  - An efficacious treatment already in use
  - New treatment will replace that efficacious treatment if shown to be superior
  - Not ethical or of interest to merely prove efficacy
  - Active control group

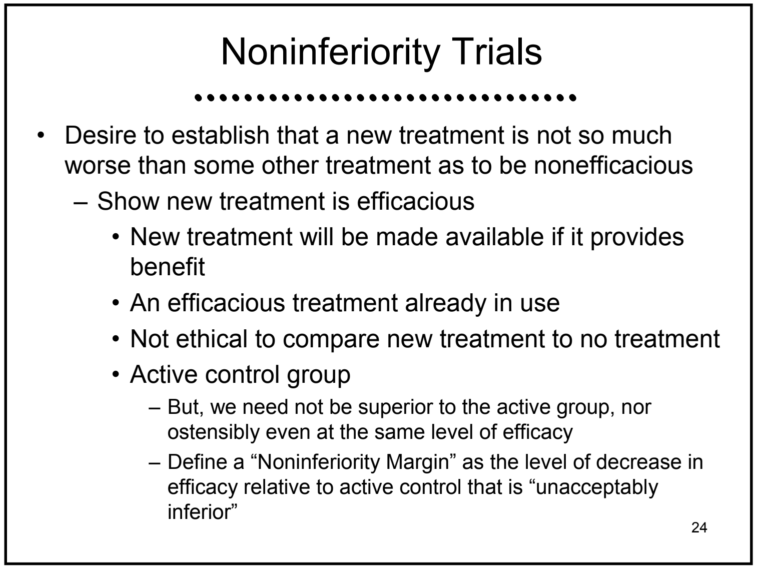
22



Common to Both

- In either case, the goal of superiority trials is to rule out equality between two treatments
  - And thus also rule out inferiority of the new treatment

23



Noninferiority Trials

- Desire to establish that a new treatment is not so much worse than some other treatment as to be nonefficacious
  - Show new treatment is efficacious
    - New treatment will be made available if it provides benefit
    - An efficacious treatment already in use
    - Not ethical to compare new treatment to no treatment
    - Active control group
      - But, we need not be superior to the active group, nor ostensibly even at the same level of efficacy
      - Define a “Noninferiority Margin” as the level of decrease in efficacy relative to active control that is “unacceptably inferior”

24

Use of Noninferiority Trials

- Noninferiority trials of use when
  - – Trying to adopt a new treatment without the expense of proving superiority
    - Often the sponsor actually believes it is superior
  - – Trying to improve secondary endpoints without removing efficacy on primary endpoint
    - E.g., in cancer chemotherapy, adverse events often correlated with efficacy

25

Major Issues with Noninferiority Trials

- Presumption that active control would be efficacious in the current trial
  - – And the need to quantify that level of efficacy
- Establishing the noninferiority margin
  - – How much of a decrease in efficacy is “unacceptably inferior”?
  - – How certain do we have to be that we have not exceeded that limit?

26

Two-sided Equivalence Studies

- Desire to rule out all differences of clinical relevance
  - – Show new treatment is approximately equivalent to existing treatment
    - New treatment will be made available if it provides approximately same level of benefit as existing treatment
    - Goal can be establishing efficacy or just establishing no harm
    - Key is in definition of “approximately equivalent” in a way to rule out the minimal clinically important differences

27

Nonsuperiority Trials

- Desire to establish that a new treatment is not so much better than some other treatment as to be of further interest
  - – Nonsuperiority trials analogous to noninferiority trials
    - Goal to rule out modest levels of benefit
  - – More an issue of futility of continuing to investigate a new treatment

28

Inferiority Trials: Placebo

- Show treatment worse than nothing (harmful)
  - Existing treatment will be discarded if shown to be harmful
  - Usually not ethical with new treatments
  - Proved to be valuable with
    - beta-carotene in cancer prevention
    - hormone replacement therapy in cardiovascular disease

Inferiority Trials: Active Control

- Show one treatment worse than another treatment
  - Inferior treatment will be discarded
  - Usually only ethical with two efficacious treatments
  - Distinction between “treatment” and “control” groups blurred

Unifying Principle

- Type of clinical trial defined by types of hypotheses rejected
  - The hypothesis “accepted” is one of exclusion.
  - Basic statistical principle
    - A distinction must be made between
      - Clinical trial results that are consistent with a hypothesis
      - Clinical trial results that establish a hypothesis
    - We regard a hypothesis as “established” by a clinical trial only if the other hypotheses of interest have been eliminated

Statistical Implications

- Looking Ahead:

- When confidence intervals are used as the criteria for statistical evidence
  - Superiority, noninferiority, equivalence, nonsuperiority, and inferiority trials are distinguished only by
    - defining the hypotheses which you desire to discriminate
    - choosing sample sizes to ensure that confidence intervals will discriminate between those hypotheses

Issues with Active Control Groups

- Choice of hypotheses: “Noninferiority Margin”
  - minimal difference that it is scientifically important to detect
  - maximal difference that it is ethical to allow
- Choice of controls
  - there is an element of historical controls being used

33

Choice of Hypotheses

- The existing treatment has (hopefully) been shown to be efficacious previously
  - Treatment effect was estimated from a sample
  - How should we choose a difference that would still establish that the new treatment is efficacious?
    - Consider the lower bound confidence interval for the treatment effect of the control treatment?
    - Perform an analysis using the estimates and standard errors from the historical studies?
  - Ethical constraints
    - How much of a decrease in efficacy is ethical
    - A treatment that is efficacious may still be proven inferior to another therapy

34

Choice of Patient Eligibility

- In using an existing treatment, we are relying on prior experience to tell us it is efficacious
  - However:
    - Few treatments are truly equally efficacious in all patients
    - Must avoid selecting a subpopulation of patients where the existing treatment is ineffective

35

Choice of Historical Controls

- In using an existing treatment, we are relying on prior experience to tell us it is efficacious
  - However:
    - There may be multiple trials of the active control versus previous standard of care
      - Multiple comparison issues in selection of the trial to use in determining the margin
      - We should expect “regression to the mean”
        - » The prior estimate of treatment effect may be overstated

36

Guiding Principles

- Historically, could active control be relied upon to have worked in current setting
    - Could trial be relied upon to truly declare an inferior treatment inferior?
  - Issues
    - Definitions of disease, outcomes
    - Nonresponders
    - Compliance
    - Ancillary treatments, time trends
    - Multiple comparisons

37

Superiority over Active Control

- When testing for superiority of a new treatment over active control
  - There is an issue that we might not accrue patients for whom the active control is known to work well
    - The “active control” may in fact be harmful in the subset of patients actually accrued
      - All of the toxicity, none of the benefit
  - Nevertheless, if the active control would be the *de facto* standard, superiority of a new treatment is still of population benefit
    - Even if worse than placebo, use of the new treatment is better than the current standard

38

Noninferiority with Active Control

- When testing for noninferiority of a new treatment over active control
  - We must worry that we are allowing adoption of a new treatment that is only proven to be not markedly worse than a possibly ineffective or harmful standard in the accrued patients
  - In choosing the margin, do we formally attenuate the estimated effect to account for this possibility?
  - In testing for noninferiority we must account for variability in prior studies
    - Game theory: We must not allow the sponsor to choose the lowest threshold

39

Statistical Analysis Plan

Choice of Summary Measure

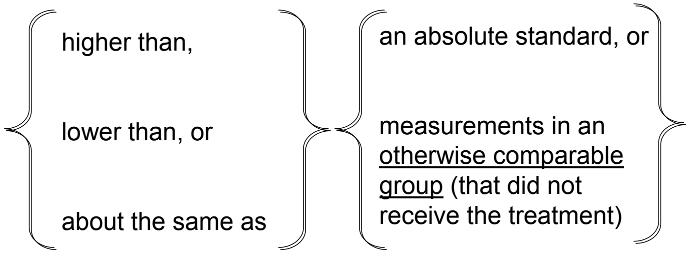
Where am I going?

- We need to refine scientific hypotheses about a clinical endpoint into testable statistical hypotheses about some summary measure of a distribution

40

Second Statistical Refinement

- The group receiving the treatment will tend to have outcome measurements that are



higher than,

lower than, or

about the same as

an absolute standard, or

measurements in an otherwise comparable group (that did not receive the treatment)

41

For Each Outcome Define “Tends To”

- In general, the space of all probability distributions is not totally ordered
  - There are an infinite number of ways we can define a tendency toward a “larger” outcome
  - This can be difficult to decide even when we have data on the entire population
    - Ex: Is the highest paid occupation in the US the one with
      - the higher mean?
      - the higher median?
      - the higher maximum?
      - the higher proportion making $1M per year?

42

Statistical Issues

- Need to choose a primary summary measure or multiple comparison issues result
- Example: Type I error with normal data

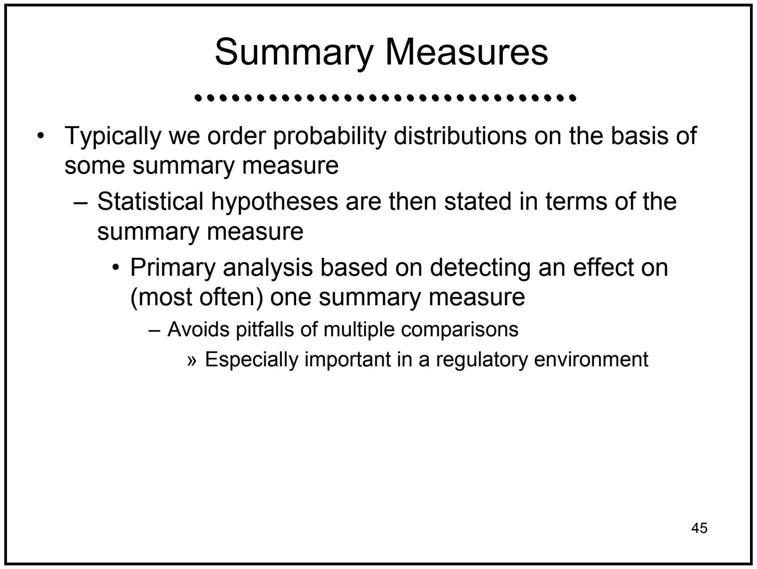
– Any single test:	0.050
– Mean, geometric mean	0.057
– Mean, Wilcoxon	0.061
– Mean, geom mean, Wilcoxon	0.066
– Above plus median	0.085
– Above plus Pr (Y > 1 sd)	0.127
– Above plus Pr (Y > 1.645 sd)	0.169

43

Primary Endpoint: Statistical

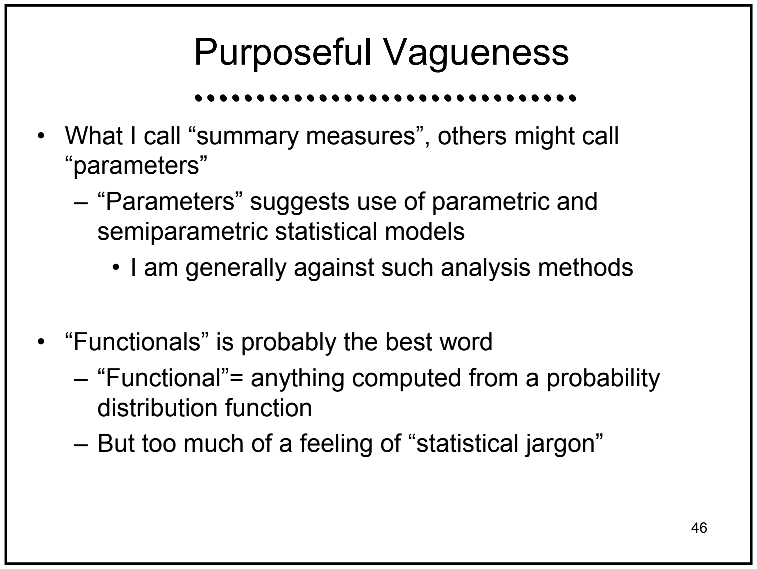
- For a specific clinical endpoint, we still have to summarize its distribution
- Consider (in order or importance)
  - The most relevant summary measure of the distribution of the primary endpoint
    - Based on a loss function?
    - Mean, median, geometric mean, …
  - The summary measurement the treatment is most likely to affect
  - The summary measure that can be that can be assessed most accurately and precisely

44



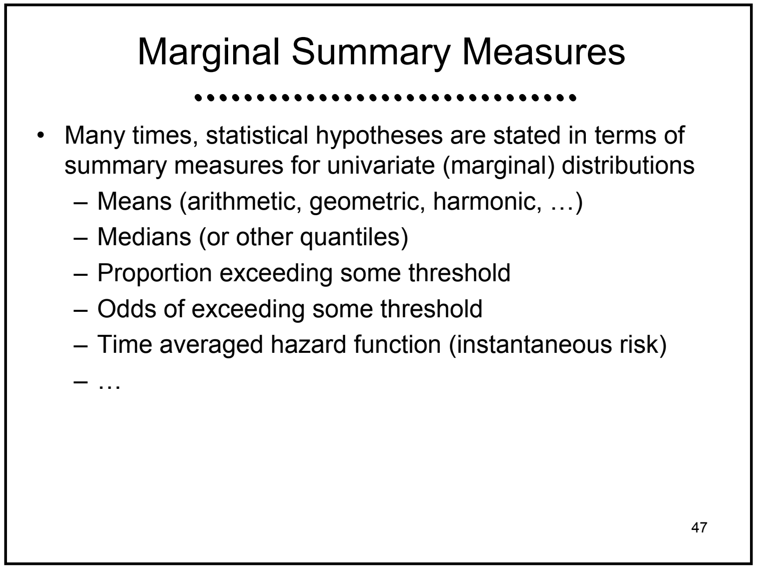
Summary Measures

- Typically we order probability distributions on the basis of some summary measure
  - Statistical hypotheses are then stated in terms of the summary measure
    - Primary analysis based on detecting an effect on (most often) one summary measure
      - Avoids pitfalls of multiple comparisons
        - » Especially important in a regulatory environment



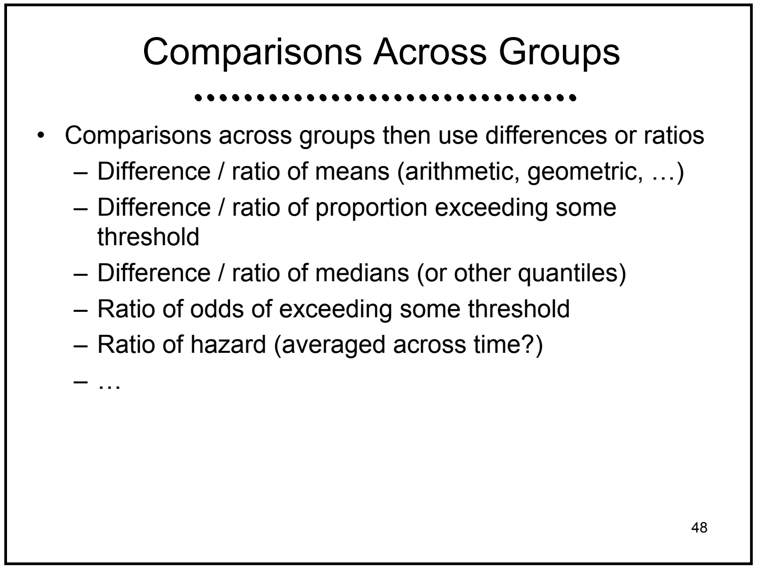
Purposeful Vagueness

- What I call “summary measures”, others might call “parameters”
  - “Parameters” suggests use of parametric and semiparametric statistical models
    - I am generally against such analysis methods
- “Functionals” is probably the best word
  - “Functional”= anything computed from a probability distribution function
  - But too much of a feeling of “statistical jargon”



Marginal Summary Measures

- Many times, statistical hypotheses are stated in terms of summary measures for univariate (marginal) distributions
  - Means (arithmetic, geometric, harmonic, …)
  - Medians (or other quantiles)
  - Proportion exceeding some threshold
  - Odds of exceeding some threshold
  - Time averaged hazard function (instantaneous risk)
  - …



Comparisons Across Groups

- Comparisons across groups then use differences or ratios
  - Difference / ratio of means (arithmetic, geometric, …)
  - Difference / ratio of proportion exceeding some threshold
  - Difference / ratio of medians (or other quantiles)
  - Ratio of odds of exceeding some threshold
  - Ratio of hazard (averaged across time?)
  - …

Joint Summary Measures

- Other times groups are compared using a summary measure for the joint distribution
  - Median difference / ratio of paired observations
  - Probability that a randomly chosen measurement from one population might exceed that from the other
  - …

Transitivity

- Distinction between marginal versus joint summary measures impacts comparisons across studies
  - Comparison across studies important in
    - Phases of drug development
    - Meta-analyses
    - Active controls
  - Most often (always?) transitivity is not guaranteed unless comparisons can be defined using marginal distributions
    - Intransitivity: Pairwise comparisons might suggest
      - A > B, and
      - B > C, but
      - C > A

Can Statistics Help?

- Litmus Test # 2:
  - If the scientific researcher cannot decide on an ordering of probability distributions that would be appropriate when measurements are available on the entire population, there is NO chance that statistics can be of any help.

Can Statisticians Help?

- While I claim that the choice of the definition for “tends to be larger” is primarily a scientific issue, statisticians do usually play an important role
  - Quantifying how different summary measures capture key features of a probability distribution
  - Ensuring that the statistical analysis model truly addresses the scientific goal

Criteria for Summary Measure

- We choose some summary measure of the probability distribution according to the following criteria (in order of importance)
  - Scientifically (clinically) relevant
    - Also reflects current state of knowledge
  - Is likely to vary across levels of the factor of interest
    - Ability to detect variety of changes
  - Statistical precision
    - Only relevant if all other things are equal

Example of Scientific Issues

- E.g., Is the arithmetic mean’s sensitivity to outliers desirable or undesirable?
  - Do we want to detect better infant mortality?
  - Does making one person immortal make up for killing others prematurely?
- E.g., Is the scientific importance of a difference in distribution best measured by the proportion exceeding some threshold?
  - Is an increase in survival time only important if the patient eventually makes it out of intensive care?
- Will we be able to quantify the degree of efficacy?
  - Clinical importance versus statistical significance

Common Practice

- The overwhelming majority of statistical inference is based on means
  - Means of continuous random variables
    - t test, linear regression
  - Proportions (means of binary random variables)
    - chi square test (t test)
  - Rates (means) for count data
    - Poisson analyses

Use of the Mean

- Rationale
  - Scientific relevance
    - Measure of “central tendency” or “location”
    - Related to totals, e.g. total health care costs
  - Plausibility that it would differ across groups
    - Sensitive to many patterns of differences in distributions (especially in tails of distributions)
  - Statistical properties
    - Distributional theory known
    - Optimal (most precise) for many distributions
    - (Ease of interpretation?)

When Not to Use the Mean

- Lack of scientific relevance
  - The mean is not defined for nominal data
  - The mean is sensitive to differences that occur only in the tail of the distribution
    - E.g., increasing the jackpot in Lotto makes one person richer, but most people still lose
  - Small differences may not be of scientific interest
    - Extend life expectancy by 24 hours
    - Decrease average cholesterol in patients with familial hypercholesterolemia by 20 mg/dl

When Not to Use the Mean

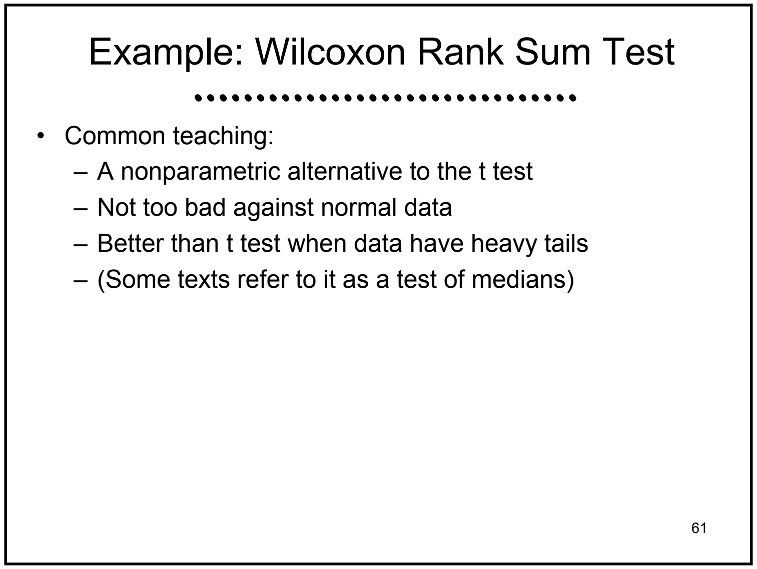
- Intervention unlikely to affect the mean
  - Sometimes we are interested in controlling variability
    - E.g., thermostats are designed to maintain house temperature within a certain range
    - E.g., control of blood glucose in diabetics?
  - (This is not typically a major criterion for avoiding the mean: It is rare that the mean is not affected by an intervention.)

When Not to Use the Mean

- Statistical criteria
  - In the presence of heavy tails (outliers)
    - the mean is not estimated with high precision
    - asymptotic distributional theory may not yet hold
  - When adjusting for covariates, it may be unreasonable to expect the mean to show constant differences across subgroups
    - Especially invoked with binary data
      - (we most often use the odds instead)

Comments on Statistical Criteria

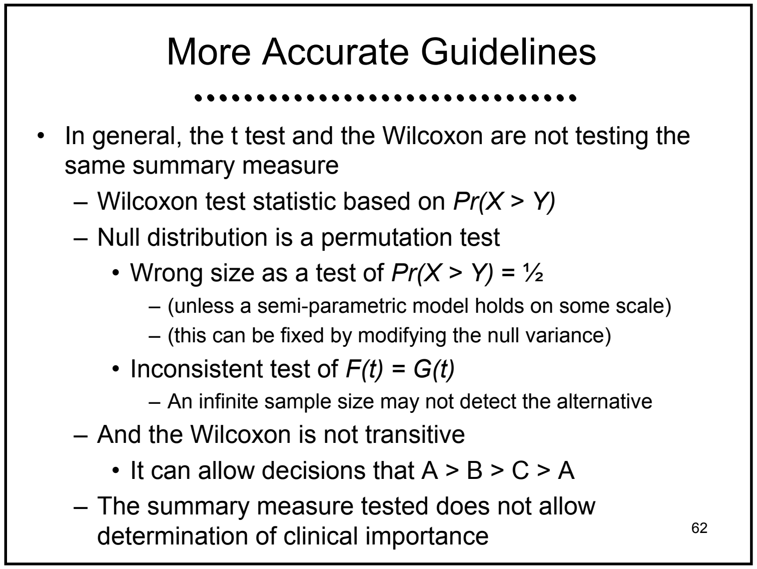
- Many of the reasons used to justify other tests are based on misconceptions
  - The validity of t tests does NOT depend heavily upon normally distributed data
    - We use it all the time with binary data
      - Handling mean-variance relationship is important in small samples
    - Modern computation allows exact small sample inference for means in same manner as used for other tests
  - The statistical theory used to demonstrate inefficiency of the mean is most often based on unreasonable (and sometimes untestable) assumptions



Example: Wilcoxon Rank Sum Test

- Common teaching:
  - A nonparametric alternative to the t test
  - Not too bad against normal data
  - Better than t test when data have heavy tails
  - (Some texts refer to it as a test of medians)

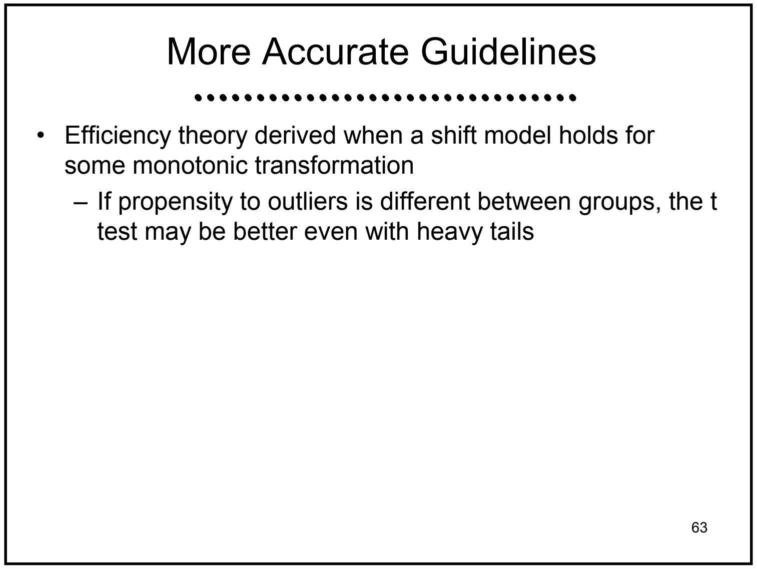
61



More Accurate Guidelines

- In general, the t test and the Wilcoxon are not testing the same summary measure
  - Wilcoxon test statistic based on *Pr(X > Y)*
  - Null distribution is a permutation test
    - Wrong size as a test of *Pr(X > Y)* = ½
      - (unless a semi-parametric model holds on some scale)
      - (this can be fixed by modifying the null variance)
    - Inconsistent test of *F(t) = G(t)*
      - An infinite sample size may not detect the alternative
  - And the Wilcoxon is not transitive
    - It can allow decisions that A > B > C > A
  - The summary measure tested does not allow determination of clinical importance

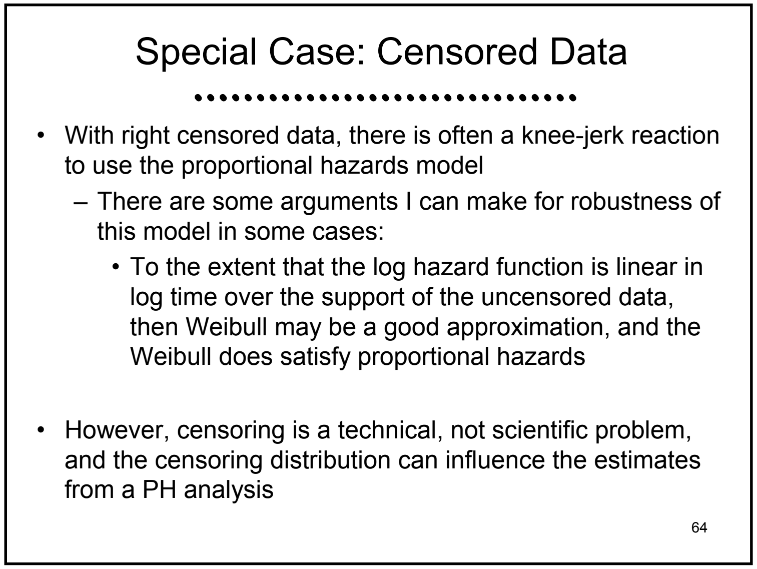
62



More Accurate Guidelines

- Efficiency theory derived when a shift model holds for some monotonic transformation
  - If propensity to outliers is different between groups, the t test may be better even with heavy tails

63



Special Case: Censored Data

- With right censored data, there is often a knee-jerk reaction to use the proportional hazards model
  - There are some arguments I can make for robustness of this model in some cases:
    - To the extent that the log hazard function is linear in log time over the support of the uncensored data, then Weibull may be a good approximation, and the Weibull does satisfy proportional hazards

- However, censoring is a technical, not scientific problem, and the censoring distribution can influence the estimates from a PH analysis

64

Hypothetical Example: Setting

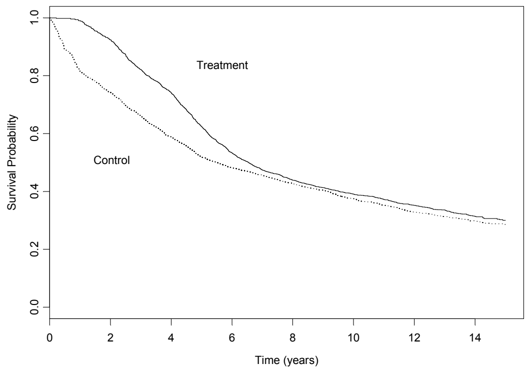
- (NOTE: This example does not pertain to RCT, but does illustrate an important point)
- Consider survival with a particular treatment used in renal dialysis patients
  - Extract data from registry of dialysis patients
    - To ensure quality, only use data after 1995
      - Incident cases in 1995: Follow-up 1995 – 2002 (8 years)
      - Prevalent cases in 1995: Data from 1995 - 2002
        - » Incident in 1994: Information about 2nd – 9th year
        - » Incident in 1993: Information about 3rd – 10th year
        - » …
        - » Incident in 1988: Information about 8th – 15th year

Hypothetical Example: Analysis

- Methods to account for censoring/truncation
  - Descriptive statistics using Kaplan-Meier
  - Options for inference
    - Parametric models
      - Weibull, lognormal, etc.
    - Semiparametric models
      - Proportional hazards, etc.
    - Nonparametric
      - Weighted rank tests: logrank, Wilcoxon, etc.
      - Comparison of Kaplan-Meier estimates

Hypothetical Example: KM Curves

Kaplan-Meier Curves for Simulated Data (n=5623)



Who Wants To Be A Millionaire?

- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  A: 2.07 (logrank P = .0018)

  B: 1.13 (logrank P = .0018)

  C: 0.87 (logrank P = .0018)

  D: 0.48 (logrank P = .0018)

  - Lifelines:
    - 50-50? Ask the audience? Call a friend?

Who Wants To Be A Millionaire?

- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  B: 1.13 (logrank P = .0018)
  C: 0.87 (logrank P = .0018)

  - Lifelines:
    - 50-50? Ask the audience? Call a friend?

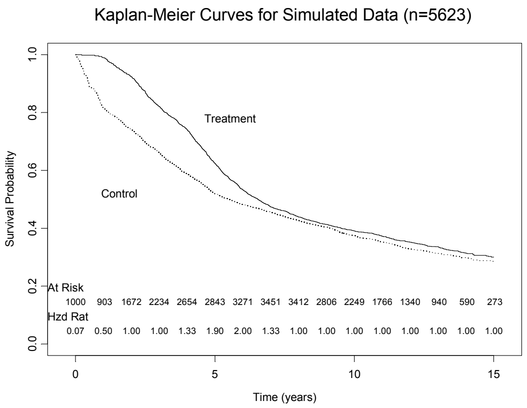
69

Who Wants To Be A Millionaire?

- How could you have known this?
  - In PH, the standard error of log hazard ratio estimates is approximately 2 divided by the square root of the number of events.
    - A P value of .0018 corresponds to | Z | = 3.13
    - log(2.07) = -log(0.48) is approximately 0.7
    - 3 x 2 / .7 is about 8.4
    - Number of deaths would be about 72
    - We had 5000+ subjects with survival estimated down to 30%

70

Hypothetical Example: KM Curves



71

Who Wants To Be A Millionaire?

- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  B: 1.13 (logrank P = .0018)

- The weighting using the risk sets made no scientific sense in this non-RCT, non-PH setting
  - Statistical precision to estimate a meaningless quantity is meaningless
  - Also: In the setting of non proportional hazards, all weighted logrank tests can be intransitive

72

Alternatives to Hazard Ratio

- More scientifically useful summary measures
  - Proportion surviving to a fixed point in time
    - Distribution free estimates from Kaplan Meier
  - Quantiles of the distribution
    - Distribution free estimates from Kaplan Meier
  - Restricted mean to a fixed point in time
    - “Months of life saved during the first 3 years”
    - Distribution free estimates from area under Kaplan Meier curve

- In all cases need to consider precision, because KM curve extremely variable on right hand side

Comments

- In any case, the decision regarding which parameter to use as the basis for inference sould be made prior to performing any analysis directly related to the question of interest
  - Basing decisions regarding choice of analysis method on the observed data will tend to inflate the type I error
    - Decrease our confidence in our statistical conclusions

Statistical Analysis Plan

Common Statistical Analysis Models

Where am I going?

- The scientific question posed by a clinical trial is typically translated into a statistical comparison of probability distributions
  - Unadjusted or adjusted comparison of summary measures

Summary Measures

- The measures commonly used to summarize and compare distributions vary according to the types of data
  - Means: binary; quantitative
  - Medians: ordered; quantitative; censored
  - Proportions: binary; nominal
  - Odds: binary; nominal
  - Hazards: censored
    - hazard = instantaneous rate of failure

General Regression

- General notation for variables and parameter

Y_i	Response measured on the ith subject
X_i	Value of the POI for the ith subject
$W_{1i}, W_{2i}, \ldots$	Value of adjustment variables for the ith subject
θ_i	Parameter of distribution of Y_i

  - The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

Multiple Regression

- General notation for simple regression model

$$g(\theta_i) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \cdots$$

$g(\ )$	"link" function used for modeling
β_0	"Intercept"
β_1	"Slope for Pred of Interest X"
β_j	"Slope for covariate W_{j-1}"

  - The link function is usually either none (means) or log (geom mean, odds, hazard)

Regression Models

- According to the parameter compared across groups
  - Means → Linear regression
  - Geom Means → Linear regression on logs
  - Odds → Logistic regression
  - Rates → Poisson regression
  - Hazards → Proportional Hazards regr
  - Quantiles → Parametric survival regr

"Everything is Regression"

- The most commonly used two sample tests are special cases of regression
  - Regression with a binary predictor
    - Linear → t test
    - Logistic → chi square (score test)
    - Proportional hazards → logrank (score test)

Interpretation of Slopes

- Difference in interpretation of slopes

$$\text{Unadjusted Model}: \quad g\left[\theta \,\middle|\, X_i\right] = \beta_0 + \beta_1 \times X_i$$

  - β_1 = Compares θ for groups differing by 1 unit in X
    - (The distribution of W might differ across groups being compared)

$$\text{Adjusted Model}: \; g\left[\theta \,\middle|\, X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

  - γ_1 = Compares θ for groups differing by 1 unit in X, but agreeing in their values of W

Points Meriting Repeated Emphasis

- Common regression models allow us to consider both adjusted and unadjusted analyses

- Generally reasonable distribution-free inference
  - Linear regression
    - Extremely robust
  - Logistic regression
    - Some issues with model mis-specification
  - Proportional hazards model
    - Dependence on censoring distribution
    - Robust if log hazard linear in log time over support of censoring distribution

Prespecification of Analysis

- To avoid multiple comparison problems, must prespecify
  - Type of regression model
  - How treatment effect will be modeled (usually binary)
  - Exactly which covariates will be included
    - Not acceptable to specify stepwise procedure
  - Exactly how covariates will be modeled (binary vs continuous)
  - Exact form of test statistic
    - Regression: Wald (based on parameter estimate), score, or likelihood ratio
    - Others: Continuity corrections, variance assumptions etc.

Sample Size Considerations

Issues of Precision

Where am I going?

- At the end of the study, we analyze our data in order to be able to make an informed decision about the effectiveness of a new treatment

- We choose a sample size for our study in order to have sufficient precision to make such inference

The Enemy

- "Let's start at the very beginning, a very good place to start…"

- Maria von Trapp

(as quoted by Rodgers and Hammerstein)

Scientific Experimentation

- At the end of the experiment, we want to present results that are convincing to the scientific community
  - The limitations of the experiment must be kept in mind

"Statistics means never having to say you are certain."
-ASA T-shirt

  - This also holds more generally for science
    - Distinguish results from conclusions
      - Dirac's sheep

Reporting Inference

- At the end of the study analyze the data
- Report three measures (four numbers)
  - Point estimate
  - Interval estimate
  - Quantification of confidence / belief in hypotheses

Reporting Frequentist Inference

- Three measures (four numbers)
  - Consider whether the observed data might reasonably be expected to be obtained under particular hypotheses
    - Point estimate: minimal bias? MSE?
    - Confidence interval: all hypotheses for which the data might reasonably be observed
    - P value: probability such extreme data would have been obtained under the null hypothesis
      - Binary decision: Reject or do not reject the null according to whether the P value is low

Reporting Bayesian Inference

- Three measures (four numbers)
  - Consider the probability distribution of the parameter conditional on the observed data
    - Point estimate: Posterior mean, median, mode
    - Credible interval: The “central” 95% of the posterior distribution
    - Posterior probability: probability of a particular hypothesis conditional on the data
      - Binary decision: Reject or do not reject the null according to whether the posterior probability is low

Parallels Between Tests, CIs

- If the null hypothesis not in CI, reject null
    - (Using same level of confidence)
- Relative advantages
  - Test only requires sampling distn under null
  - CI requires sampling distn under alternatives
  - CI provides interpretation when null is not rejected

Scientific Information

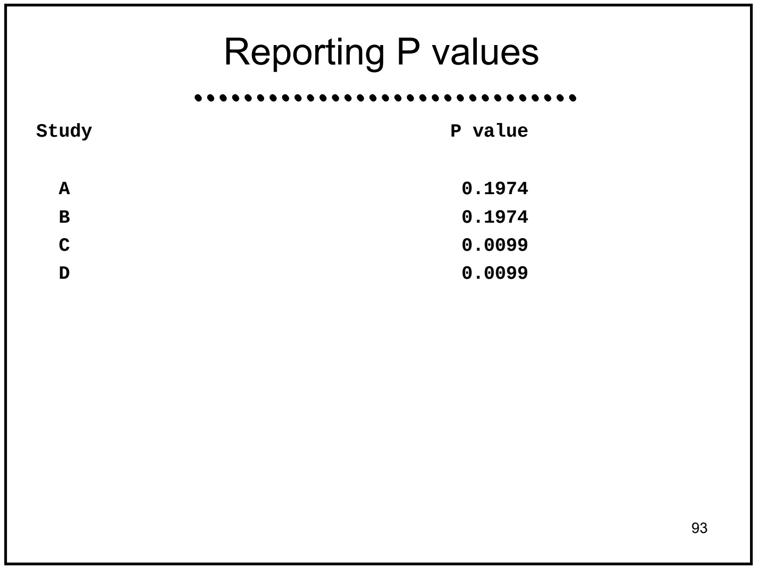
- “Rejection” uses a single level of significance
  - Different settings might demand different criteria

- P value communicates statistical evidence, not scientific importance

- Only confidence interval allows you to interpret failure to reject the null:
  - Distinguish between
    - Inadequate precision (sample size)
    - Strong evidence for null

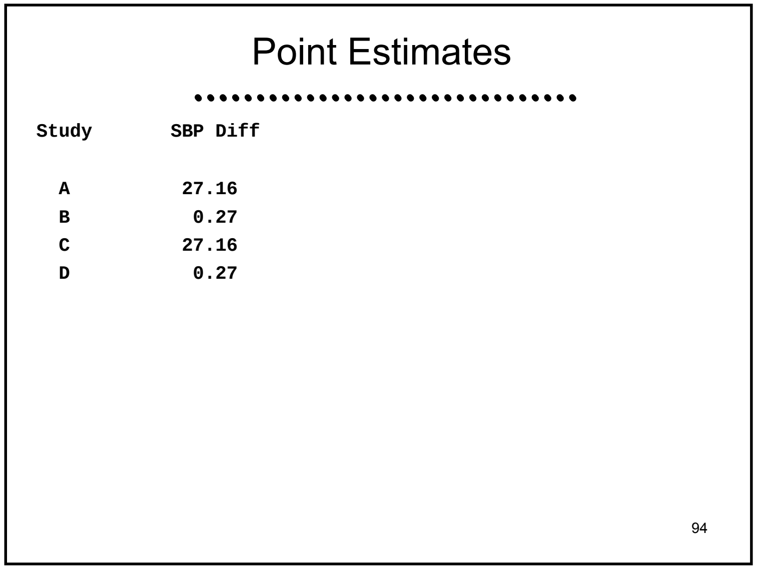
Hypothetical Example

- Clinical trials of treatments for hypertension
  - Screening trials for four candidate drugs
    - Measure of treatment effect is the difference in average SBP at the end of six months treatment
    - Drugs may differ in
      - Treatment effect (goal is to find best)
      - Variability of blood pressure
    - Clinical trials may differ in conditions
      - Sample size, etc.



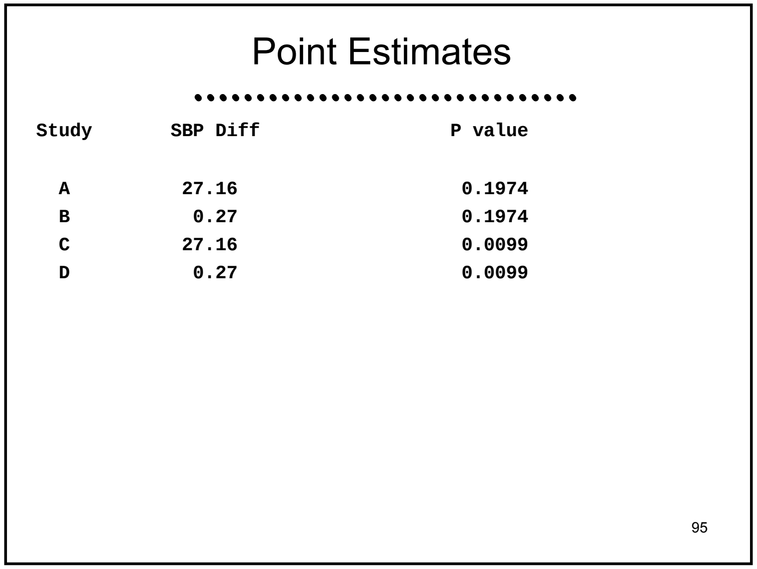
Reporting P values

Study	P value
A	0.1974
B	0.1974
C	0.0099
D	0.0099



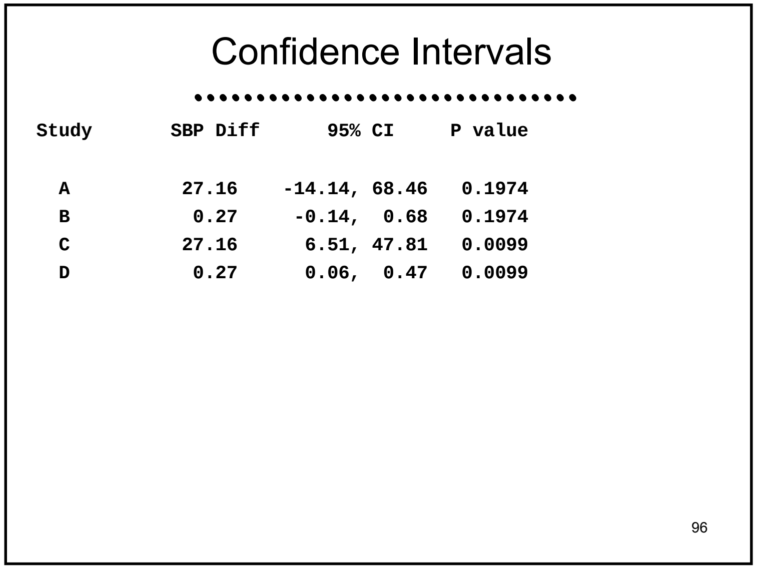
Point Estimates

Study	SBP Diff
A	27.16
B	0.27
C	27.16
D	0.27



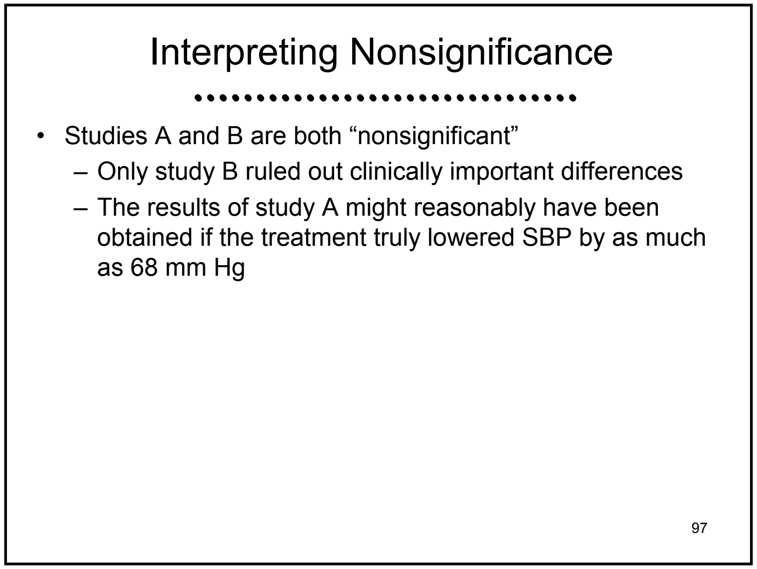
Point Estimates

Study	SBP Diff	P value
A	27.16	0.1974
B	0.27	0.1974
C	27.16	0.0099
D	0.27	0.0099



Confidence Intervals

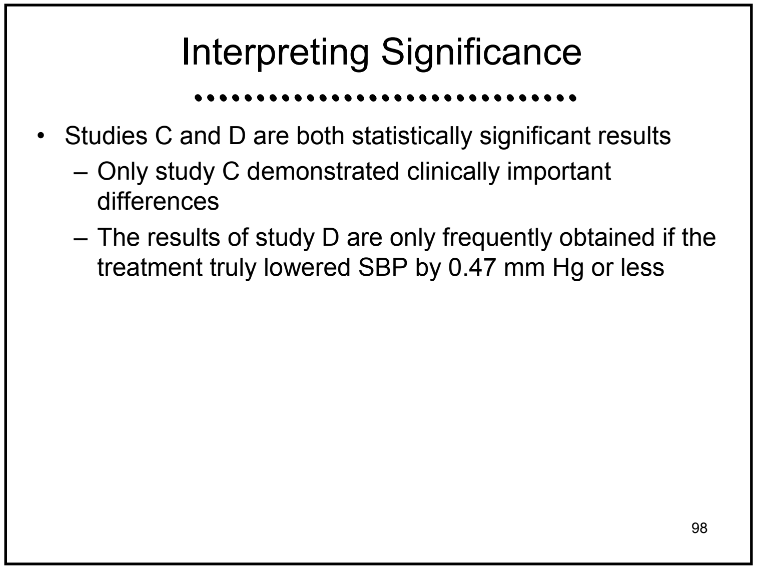
Study	SBP Diff	95% CI	P value
A	27.16	-14.14, 68.46	0.1974
B	0.27	-0.14, 0.68	0.1974
C	27.16	6.51, 47.81	0.0099
D	0.27	0.06, 0.47	0.0099



Interpreting Nonsignificance

- Studies A and B are both “nonsignificant”
  - Only study B ruled out clinically important differences
  - The results of study A might reasonably have been obtained if the treatment truly lowered SBP by as much as 68 mm Hg

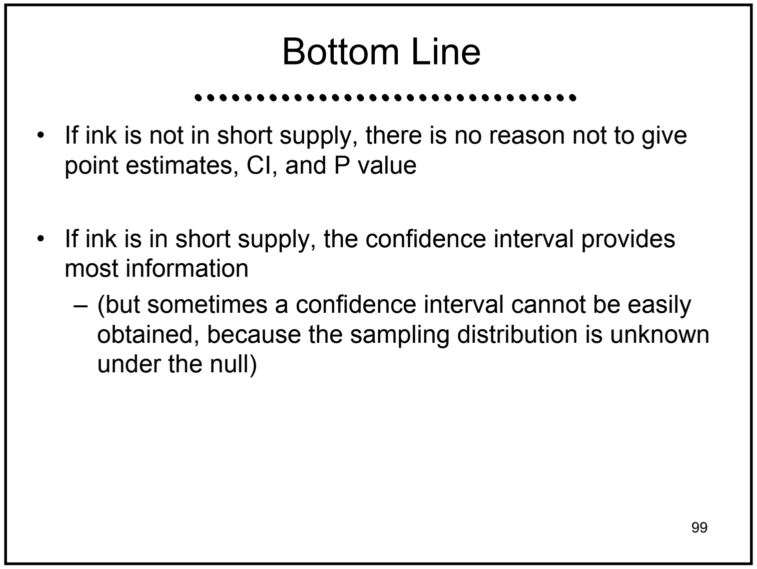
97



Interpreting Significance

- Studies C and D are both statistically significant results
  - Only study C demonstrated clinically important differences
  - The results of study D are only frequently obtained if the treatment truly lowered SBP by 0.47 mm Hg or less

98

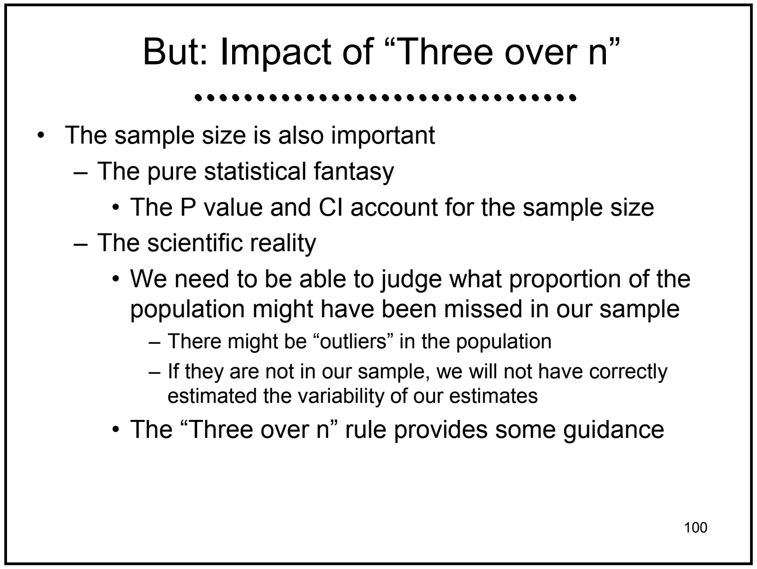


Bottom Line

- If ink is not in short supply, there is no reason not to give point estimates, CI, and P value

- If ink is in short supply, the confidence interval provides most information
  - (but sometimes a confidence interval cannot be easily obtained, because the sampling distribution is unknown under the null)

99



But: Impact of “Three over n”

- The sample size is also important
  - The pure statistical fantasy
    - The P value and CI account for the sample size
  - The scientific reality
    - We need to be able to judge what proportion of the population might have been missed in our sample
      - There might be “outliers” in the population
      - If they are not in our sample, we will not have correctly estimated the variability of our estimates
    - The “Three over n” rule provides some guidance

100

Real World Example

- Consider the following data:

  0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 7

- Do we throw out the outlier?
  - What would we have said after the first 24 observations?

101

Elevator Stats: 0 events in n trials

- Two-sided confidence intervals fail in the case where there are either 0 or n events observed in n Bernoulli trials
  - If Y=0, there is no lower confidence bound
  - If Y=n, there is no upper confidence bound

– We can, however, derive one-sided confidence bounds in that case

102

Upper Conf Bnd for 0 Events

- Exact upper confidence bound when all observations are 0

Suppose $Y \sim B(n, p)$ and $Y = 0$ is observed

Exact $100(1\text{-}\alpha)\%$ upper confidence bound for p is $\hat{p}_U$

$$\Pr\left[Y = 0; \hat{p}_U\right] = \left(1 - \hat{p}_U\right)^n = \alpha$$

$$\Downarrow$$

$$\hat{p}_U = 1 - \alpha^{1/n}$$

103

Large Sample Approximation

$$\left(1 - \hat{p}_U\right)^n = \alpha \quad \Rightarrow \quad n\log\left(1 - \hat{p}_U\right) = \log(\alpha)$$

For small $\hat{p}_U$ $\quad \log\left(1 - \hat{p}_U\right) \approx -\hat{p}_U$

so for large n $\quad \Rightarrow \quad \hat{p}_U \approx -\dfrac{\log(\alpha)}{n}$

104

Elevator Stats: 0 Events in n trials

- “Three over n rule”
  - log (.05) = -2.9957
  - In large samples, when 0 events observed, the 95% upper confidence bound for p is approximately 3 / n

- 99% upper confidence bound
  - log (.01) = -4.605
  - Use 4.6 / n as 99% upper confidence bound

Elevator Stats vs Exact

- When X=0 events observed in n Bernoulli trials

	95% bound		99% bound	
n	Exact	3/n	Exact	4.6/n
2	.7764	1.50	.9000	2.3000
5	.4507	.60	.6019	.9200
10	.2589	.30	.3690	.4600
20	.1391	.15	.2057	.2300
30	.0950	.10	.1423	.1533
50	.0582	.06	.0880	.0920
100	.0295	.03	.0450	.0460

Real World Example

- How many people die on a space shuttle launch:

- Data as of January 28, 1986:

  0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 7

- Do we throw out the outlier?
  - What would we have said after the first 24 observations?
    - 95% upper bound on failure rate ≈ 3/24 = 12.5%

Full Report of Analysis

Study	n	SBP Diff	95% CI	P value
A	20	27.16	-14.14, 68.46	0.1974
B	20	0.27	-0.14, 0.68	0.1974
C	80	27.16	6.51, 47.81	0.0099
D	80	0.27	0.06, 0.47	0.0099

Interpreting a “Negative Study”

- This then highlights issues related to the interpretation of a study in which no statistically significant difference between groups was found
  - – We have to consider the “differential diagnosis” of possible situations in which we might observe nonsignificance

General approach

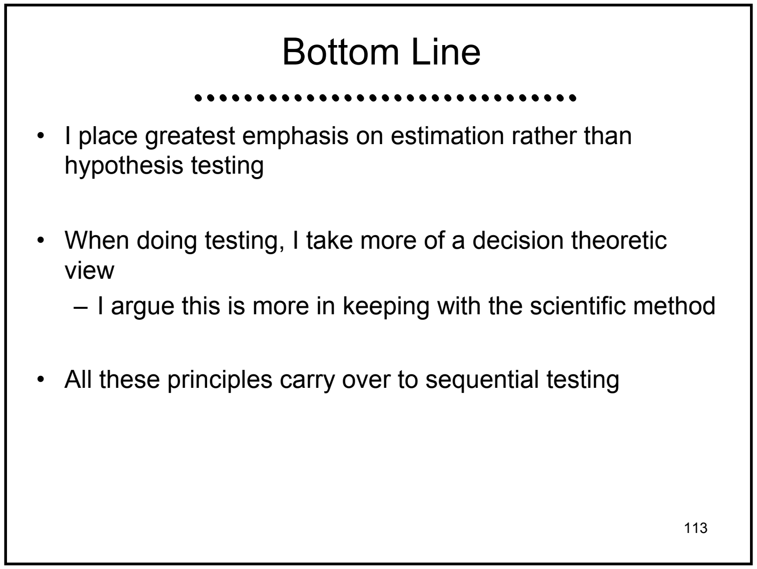
- Refined scientific question
  - – We compare the distribution of some response variable differs across groups
    - E.g., looking for an association between smoking and blood pressure by comparing distribution of SBP between smokers and nonsmokers
  - – We base our decisions on a scientifically appropriate summary measure θ
    - E.g., difference of means, ratio of medians, …

Interpreting a “Negative Study”

- Possible explanations for no statistically significant difference in estimate of θ
  - – There is no true difference in the distribution of response across groups
  - – There is a difference in the distribution of response across groups, but the value of θ is the same for both groups
    - (i.e., the distributions differ in some other way)
  - – There is a difference in the value of θ between the groups, but our study was not precise enough
    - A “type II error” from low “statistical power”

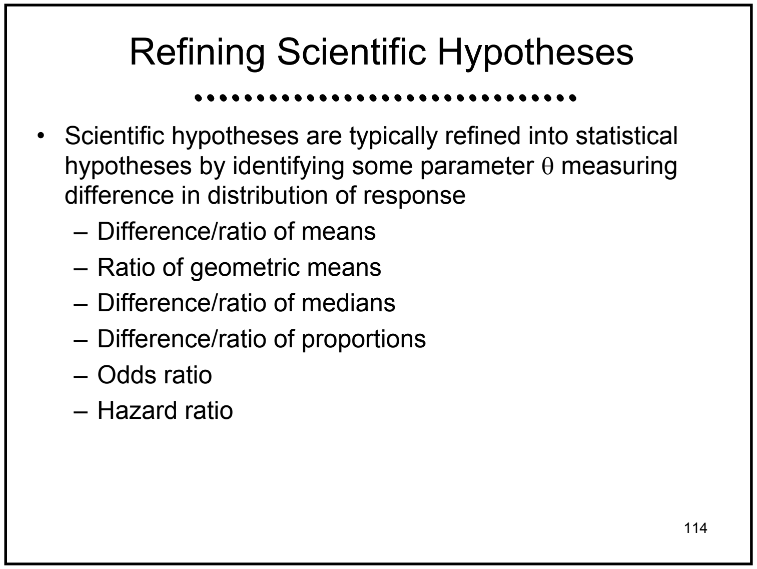
Interpreting a “Positive Study”

- Analogous interpretations when we do find a statistically significant difference in estimate of θ
  - – There is a true difference in the value of θ
  - – There is no true difference in θ, but we were unlucky and observed spuriously high or low results
    - Random chance leading to a “type I error”
      - – The p value tells us how unlucky we would have had to have been
    - (Used a statistic that allows other differences in the distn to be misinterpreted as a difference in θ
      - – E.g., different variances causing significant t test)



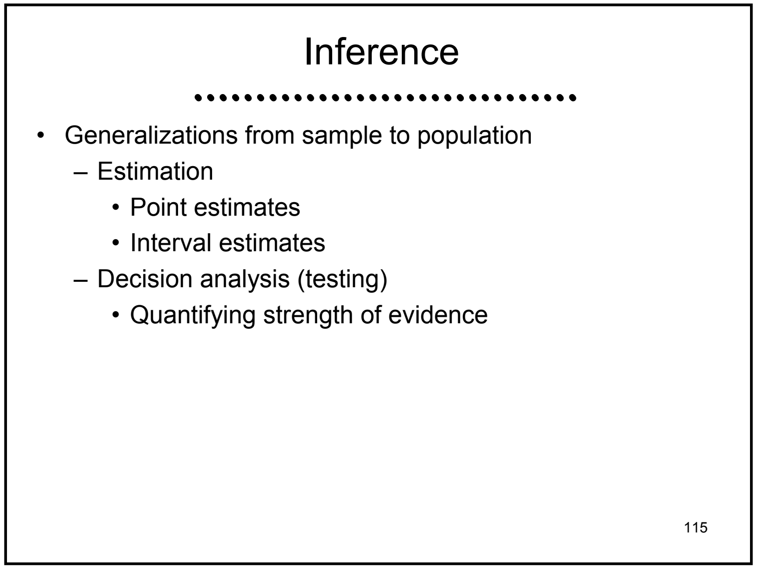
Bottom Line

- I place greatest emphasis on estimation rather than hypothesis testing
- When doing testing, I take more of a decision theoretic view
  - I argue this is more in keeping with the scientific method
- All these principles carry over to sequential testing



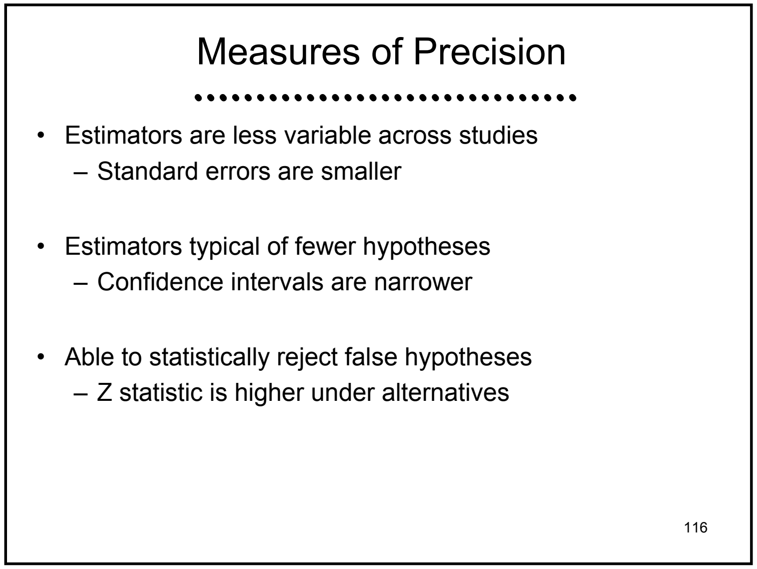
Refining Scientific Hypotheses

- Scientific hypotheses are typically refined into statistical hypotheses by identifying some parameter θ measuring difference in distribution of response
  - Difference/ratio of means
  - Ratio of geometric means
  - Difference/ratio of medians
  - Difference/ratio of proportions
  - Odds ratio
  - Hazard ratio



Inference

- Generalizations from sample to population
  - Estimation
    - Point estimates
    - Interval estimates
  - Decision analysis (testing)
    - Quantifying strength of evidence



Measures of Precision

- Estimators are less variable across studies
  - Standard errors are smaller
- Estimators typical of fewer hypotheses
  - Confidence intervals are narrower
- Able to statistically reject false hypotheses
  - Z statistic is higher under alternatives

Criteria for Precision

- Standard error
- Width of confidence interval
- Statistical power
  - Probability of rejecting the null hypothesis
    - Select “design alternative”
    - Select desired power

Statistics to Address Variability

- At the end of the study:
  - Frequentist and/or Bayesian data analysis to assess the credibility of clinical trial results
    - Estimate of the treatment effect
      - Single best estimate
      - Precision of estimates
    - Decision for or against hypotheses
      - Binary decision
      - Quantification of strength of evidence

Sample Size Determination

- Based on sampling plan, statistical analysis plan, and estimates of variability, compute
  - Sample size that discriminates hypotheses with desired power, or
  - Hypothesis that is discriminated from null with desired power when sample size is as specified, or
  - Power to detect the specific alternative when sample size is as specified

Sample Size Computation

Standardized level α test (n = 1): $\delta_{\alpha\beta}$ detected with power β

Level of significance α when $\theta = \theta_0$

Design alternative $\theta = \theta_1$

Variability V within 1 sampling unit

Required sampling units: $$n = \frac{\left(\delta_{\alpha\beta}\right)^2 V}{\left(\theta_1 - \theta_0\right)^2}$$

(Fixed sample test: $\delta_{\alpha\beta} = z_{1-\alpha/2} + z_{\beta}$)

When Sample Size Constrained

- Often (usually?) logistical constraints impose a maximal sample size
  - Compute power to detect specified alternative

Find β such that $\quad \delta_{\alpha\beta} = \sqrt{\frac{n}{V}}\left(\theta_1 - \theta_0\right)$

  - Compute alternative detected with high power

$$\theta_1 = \theta_0 + \delta_{\alpha\beta}\sqrt{\frac{V}{n}}$$

General Comments

- What alternative to use?
  - Minimal clinically important difference (MCID)
    - To detect? (use in sample size formula)
    - To declare significant? (look at critical value)
- What level of significance?
  - “Standard”: one-sided 0.025, two-sided 0.05
  - “Pivotal”: one-sided 0.005?
    - Do we want to be extremely confident of an effect, or confident of an extreme effect
- What power?
  - Science: 97.5% (unless MCID for significance→ ~50%)
  - Subterfuge: 80% or 90%

Role of Secondary Analyses

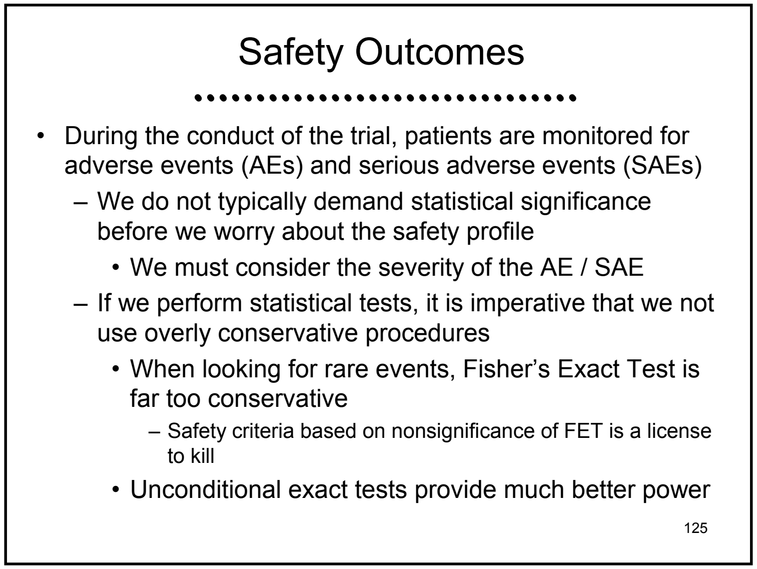
- We choose a primary outcome to avoid multiple comparison problems
  - That primary outcome may be a composite of several clinical outcomes, but there will only be one CI, test

- We select a few secondary outcomes to provide supporting evidence or confirmation of mechanisms
  - Those secondary outcomes may be
    - alternative clinical measures and/or
    - different summary measures of the primary clinical endpoint

Secondary Analysis Models

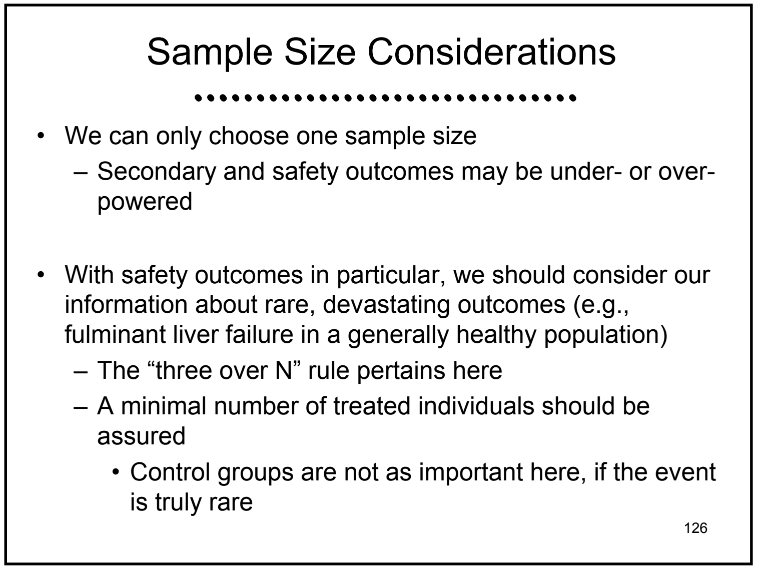
- Selection of statistical models for secondary analyses should generally adhere to same principles as for primary outcome, including intent to treat

- Some exceptions:
  - Exploratory analyses based on dose actually taken may be undertaken to generate hypotheses about dose response
  - Exploratory cause specific time to event analyses may be used to investigate hypothesized mechanisms



Safety Outcomes

- During the conduct of the trial, patients are monitored for adverse events (AEs) and serious adverse events (SAEs)
  - We do not typically demand statistical significance before we worry about the safety profile
    - We must consider the severity of the AE / SAE
  - If we perform statistical tests, it is imperative that we not use overly conservative procedures
    - When looking for rare events, Fisher's Exact Test is far too conservative
      - Safety criteria based on nonsignificance of FET is a license to kill
    - Unconditional exact tests provide much better power



Sample Size Considerations

- We can only choose one sample size
  - Secondary and safety outcomes may be under- or over-powered

- With safety outcomes in particular, we should consider our information about rare, devastating outcomes (e.g., fulminant liver failure in a generally healthy population)
  - The "three over N" rule pertains here
  - A minimal number of treated individuals should be assured
    - Control groups are not as important here, if the event is truly rare