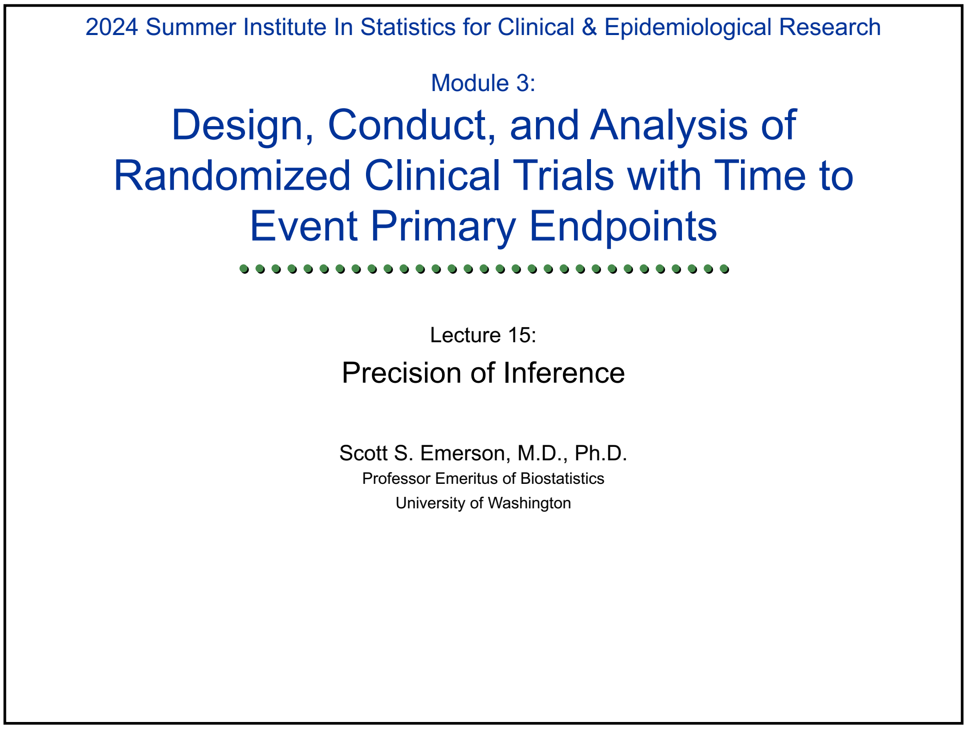


2024 Summer Institute In Statistics for Clinical & Epidemiological Research

Module 3:

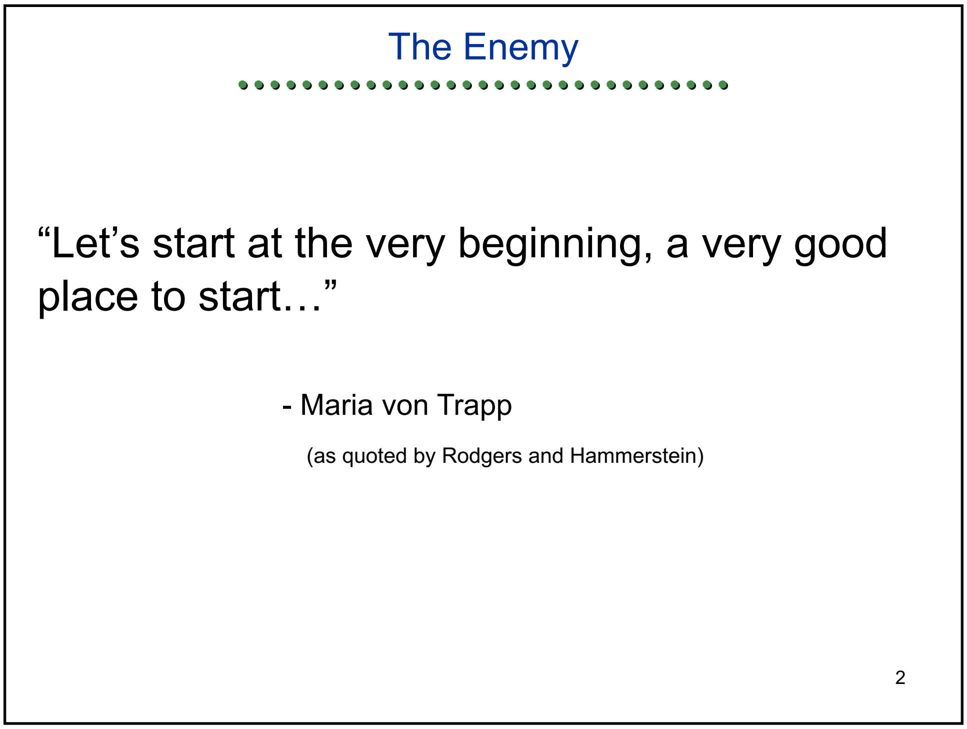
Design, Conduct, and Analysis of Randomized Clinical Trials with Time to Event Primary Endpoints

Lecture 15:

Precision of Inference

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

1



The Enemy

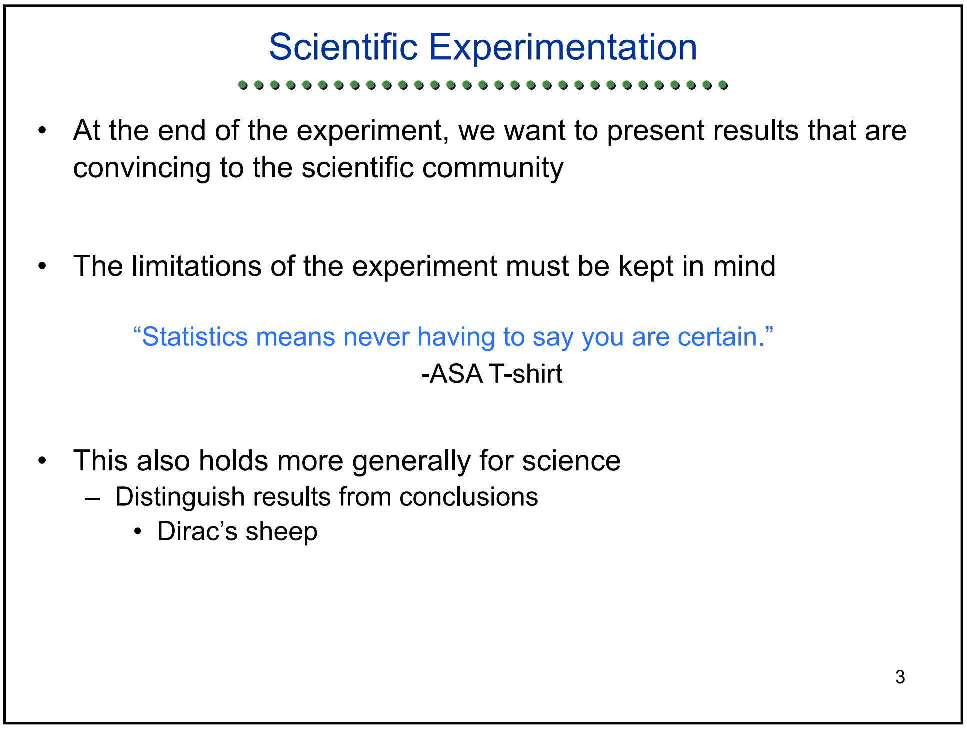
“Let’s start at the very beginning, a very good place to start…”

- Maria von Trapp

(as quoted by Rodgers and Hammerstein)

2

2

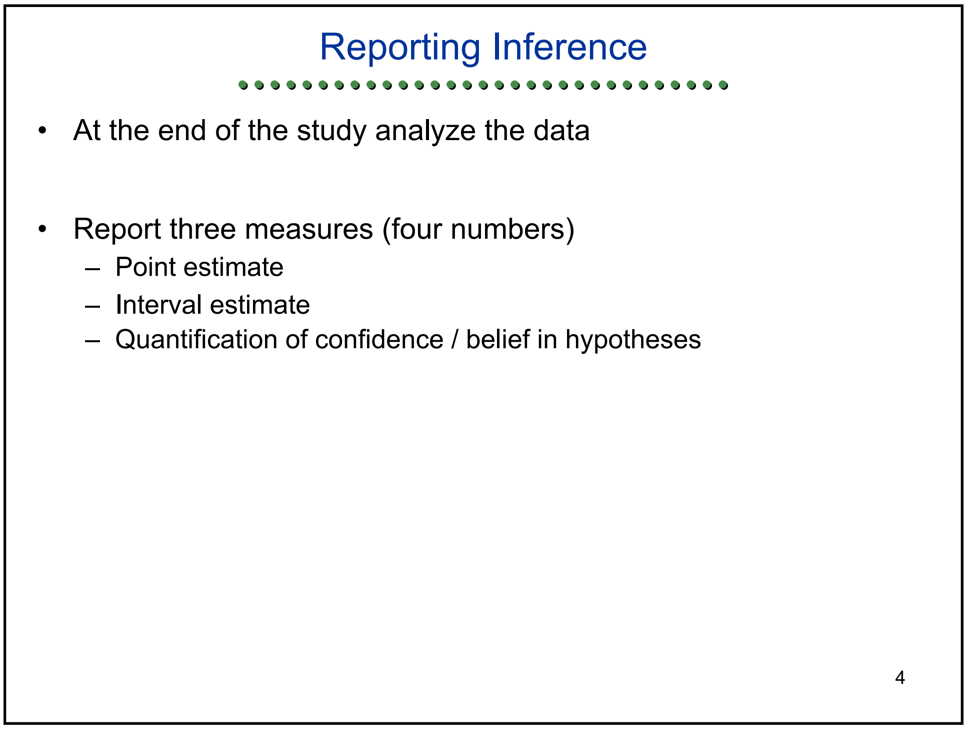


Scientific Experimentation

- 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

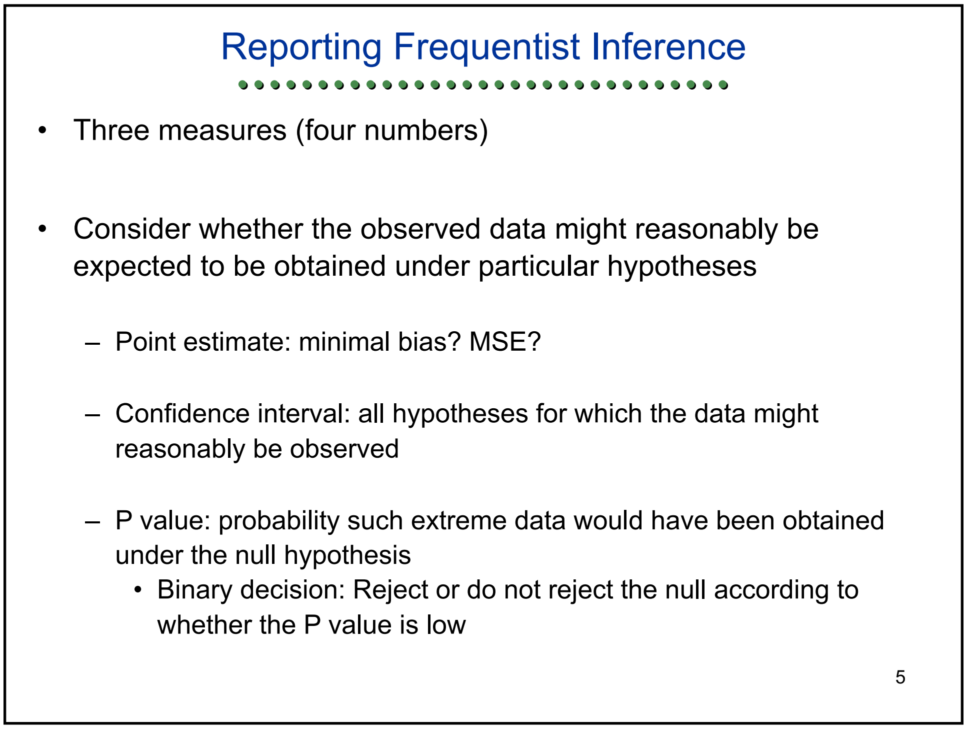
> "Statistics means never having to say you are certain."
> -ASA T-shirt

- 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



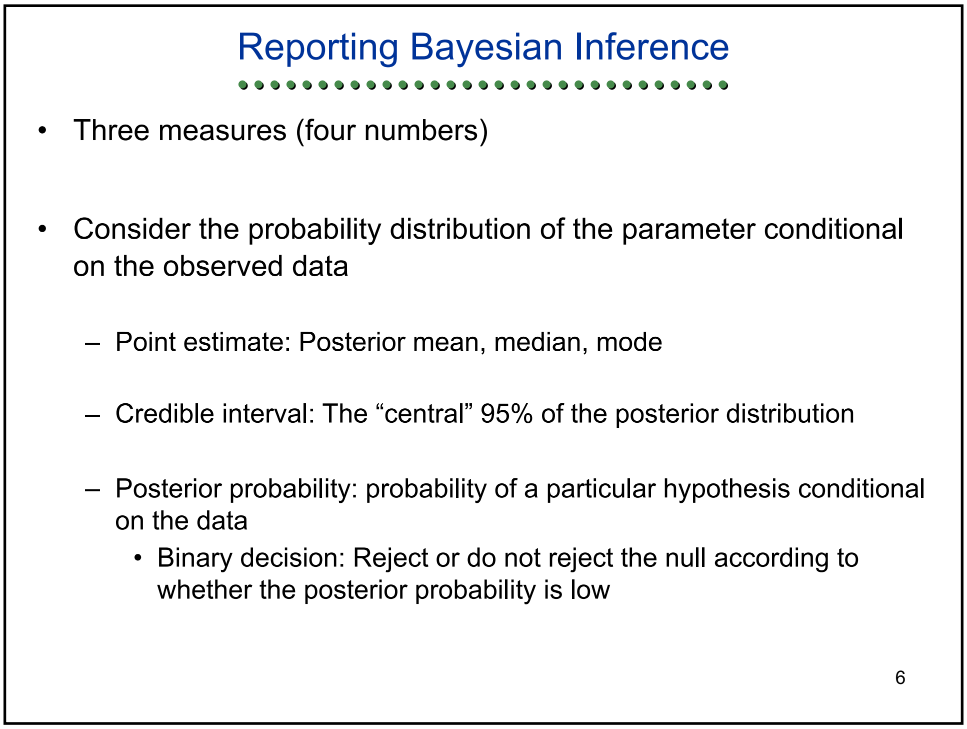
Reporting Frequentist Inference

- 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

5

5



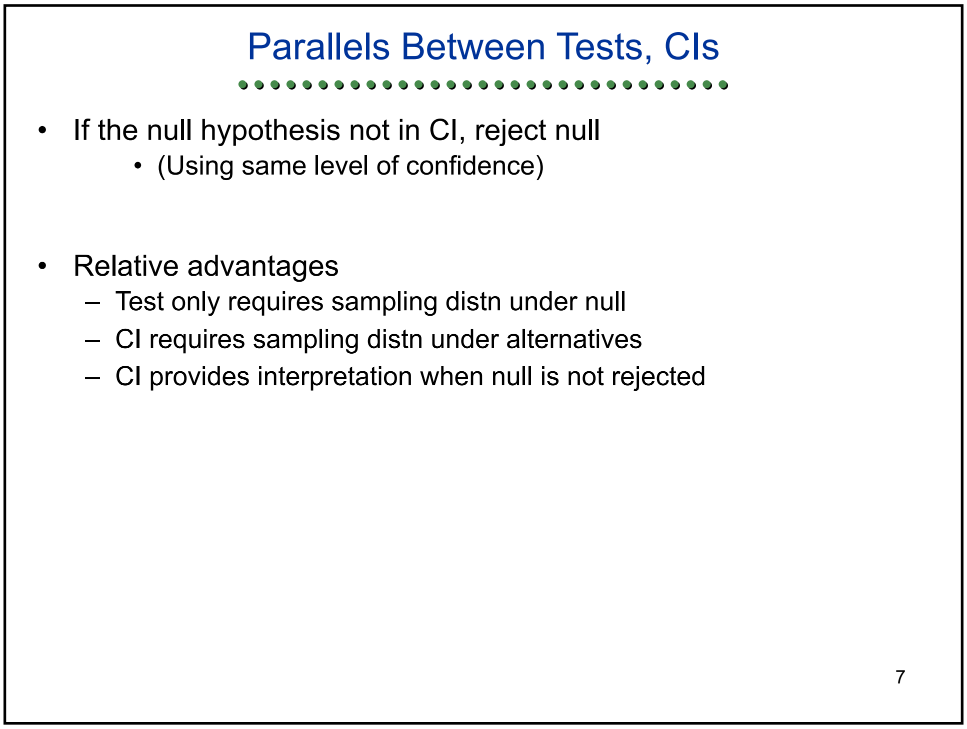
Reporting Bayesian Inference

- 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

6

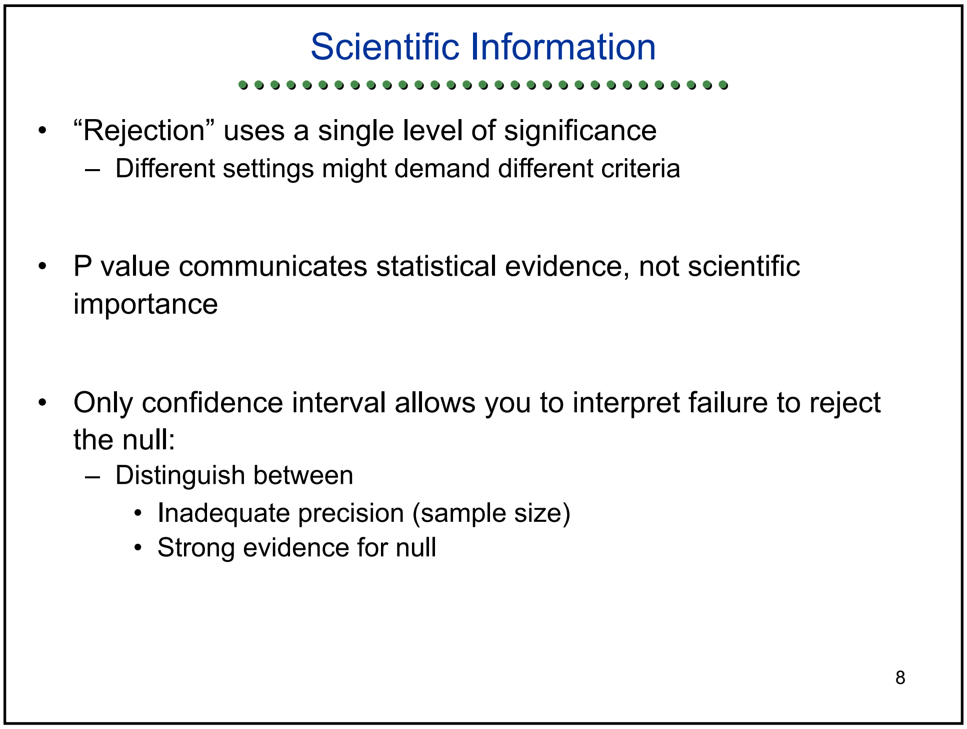
6



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

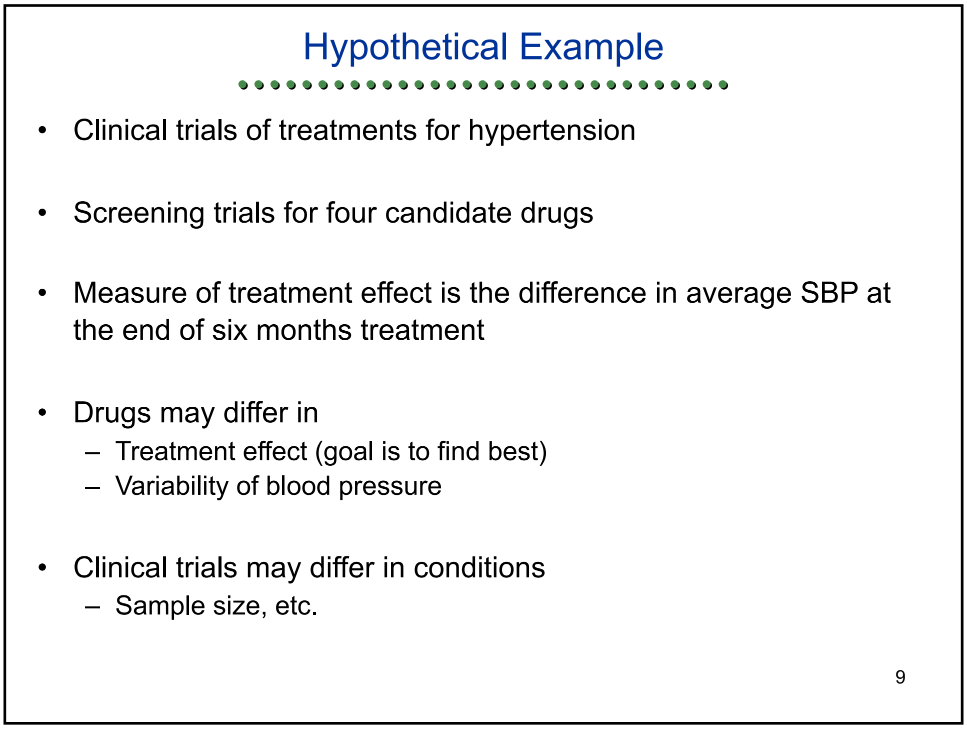
7



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

8

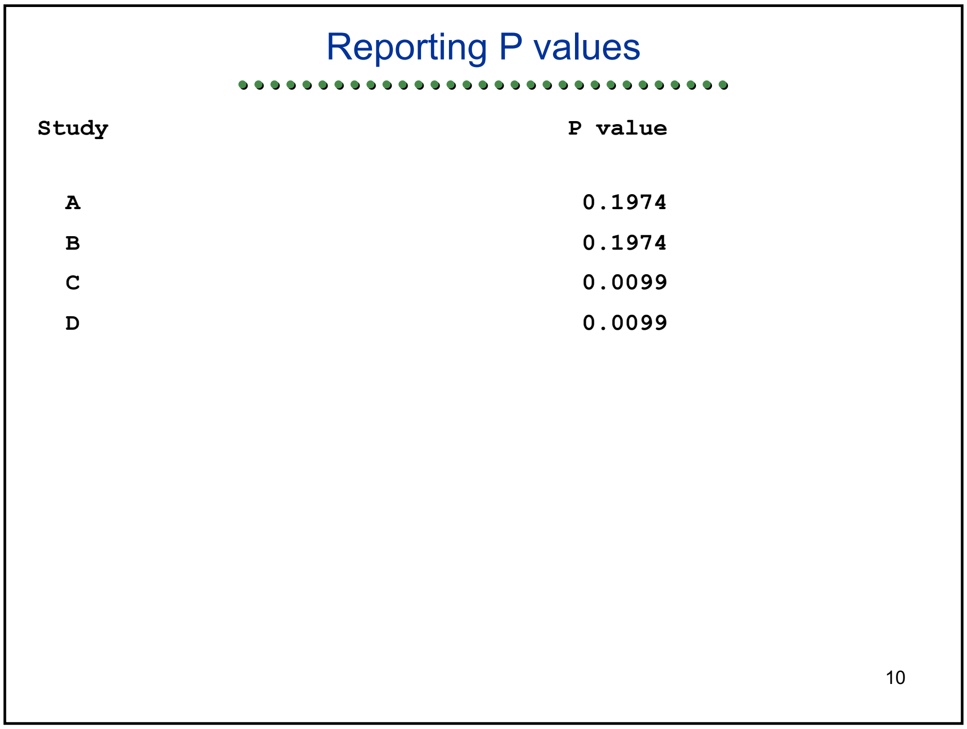


Hypothetical Example

- 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.

9

9

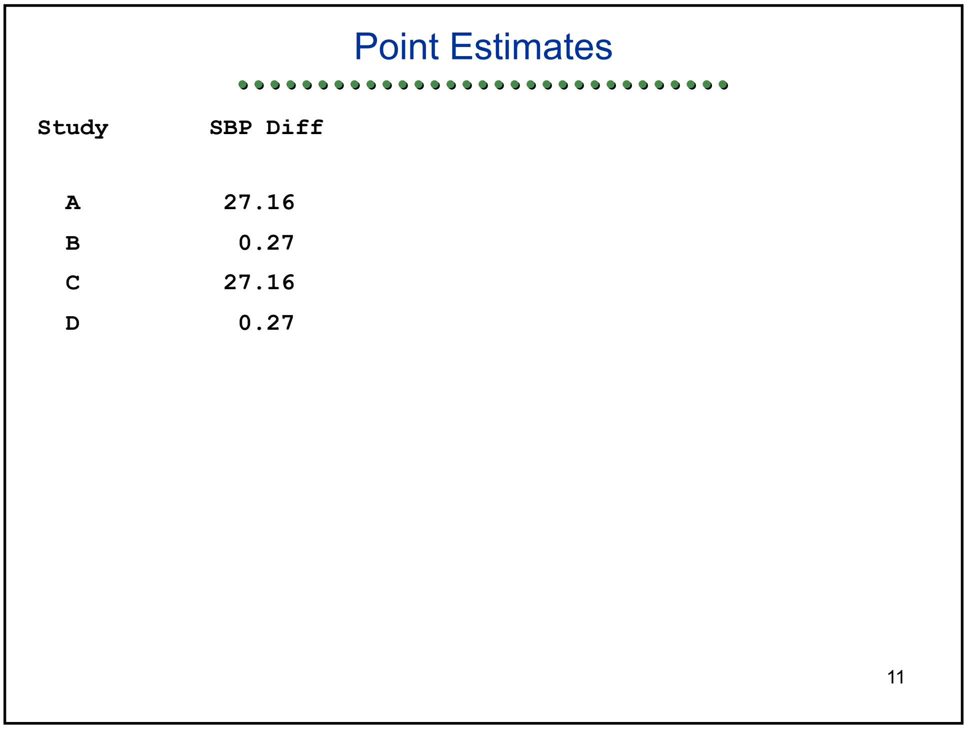


Reporting P values

Study	P value
A	0.1974
B	0.1974
C	0.0099
D	0.0099

10

10

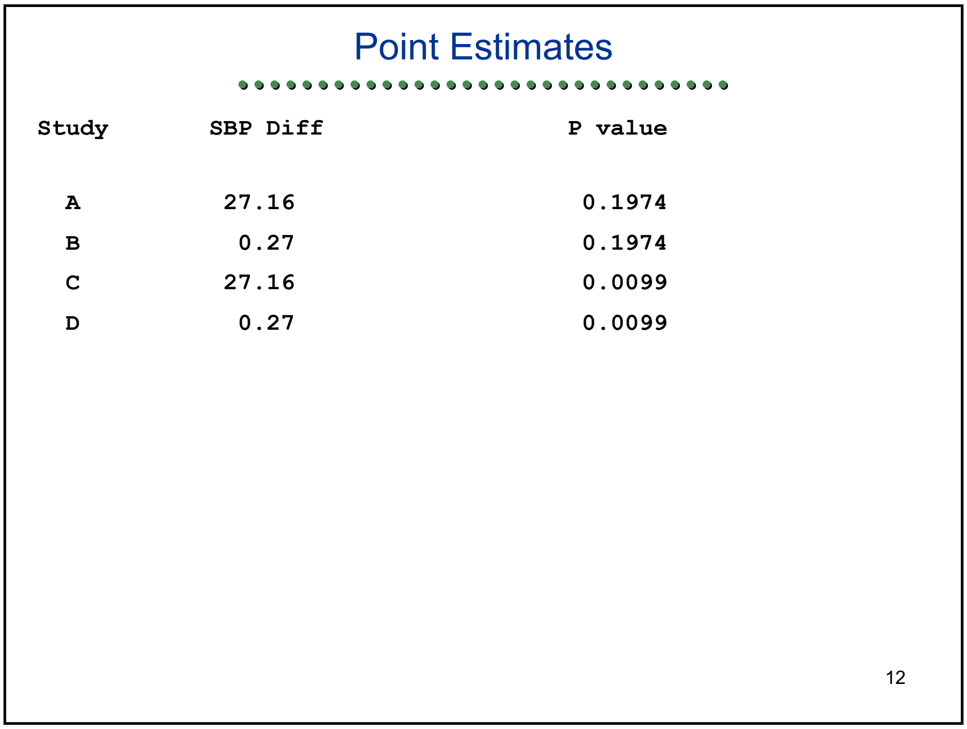


Point Estimates

Study	SBP Diff
A	27.16
B	0.27
C	27.16
D	0.27

11

11

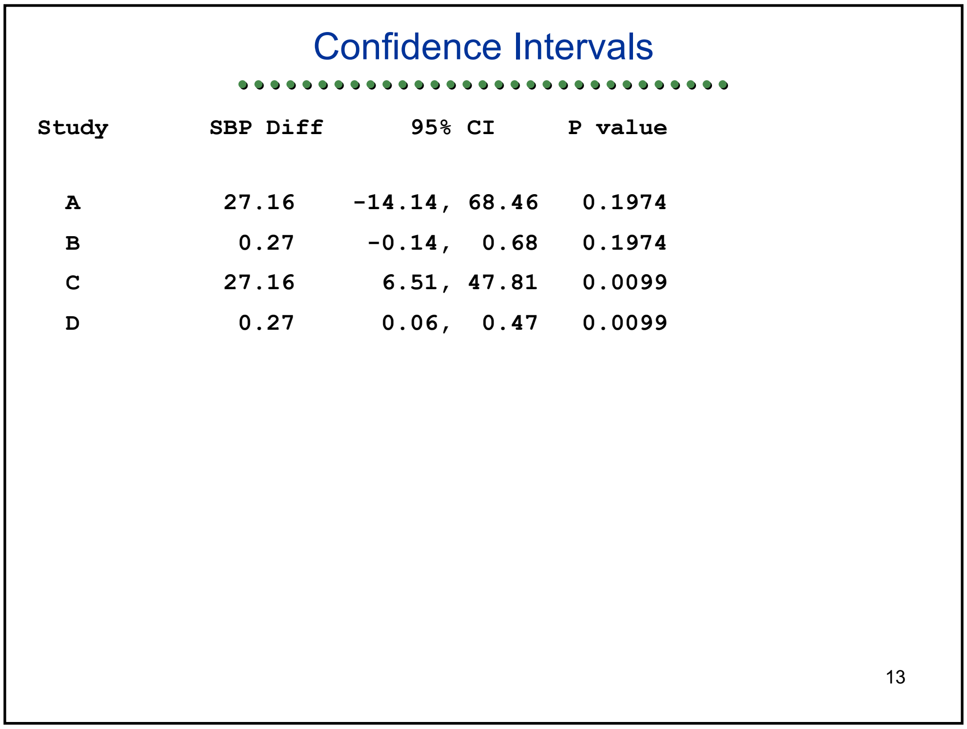


Point Estimates

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

12

12

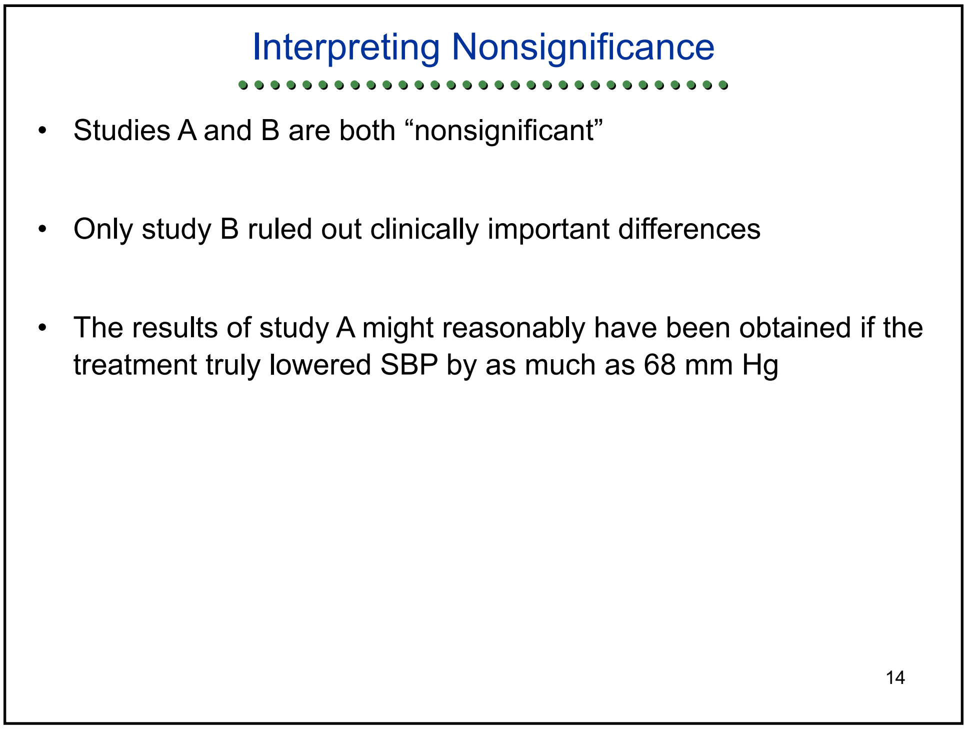


Confidence Intervals

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

13

13

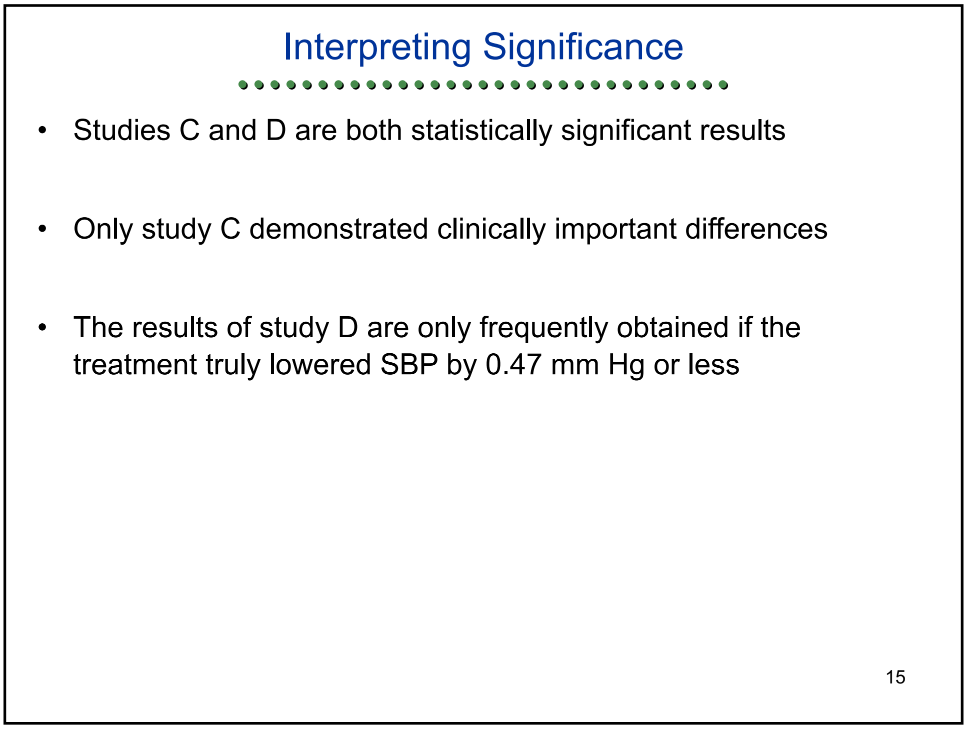


Interpreting Nonsignificance

- 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

14

14

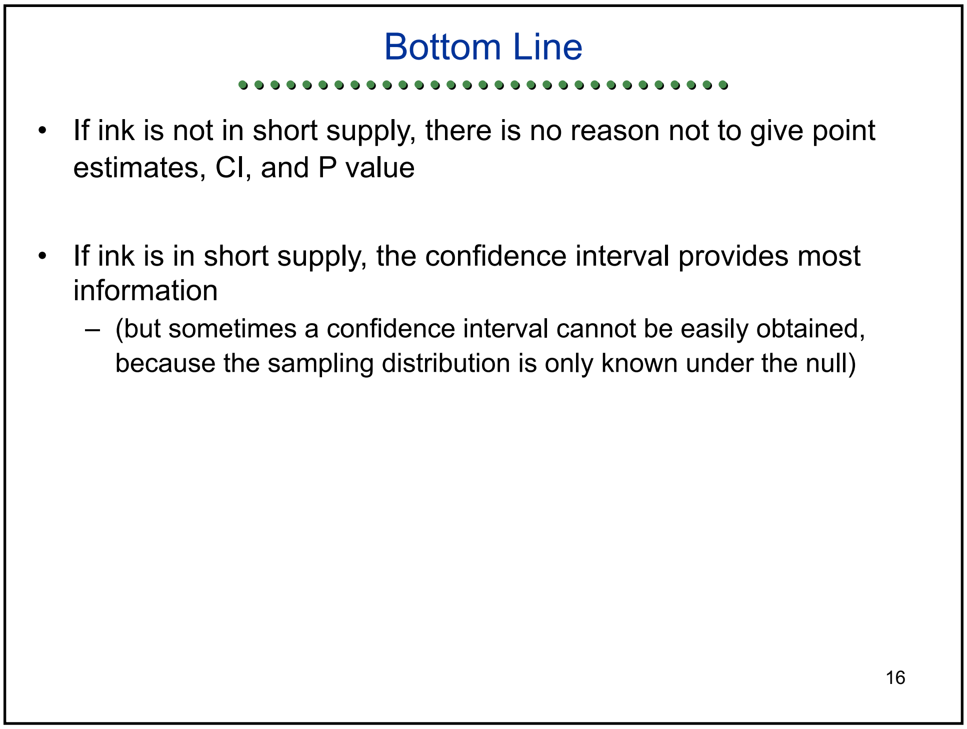


Interpreting Significance

- 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

15

15

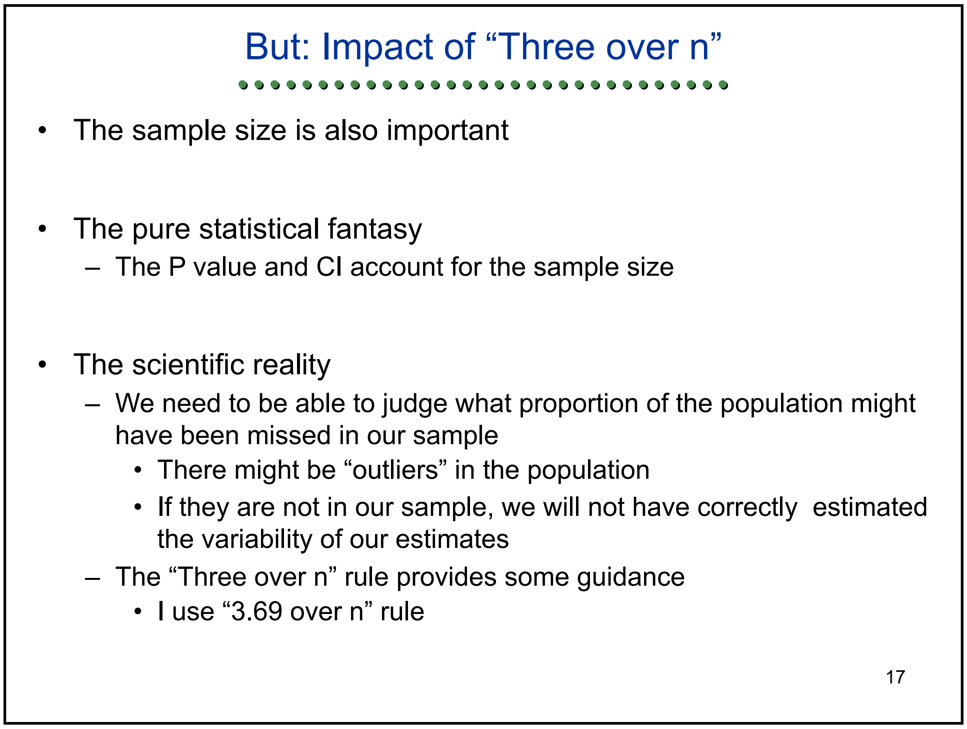


Bottom Line

- 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 only known under the null)

16

16

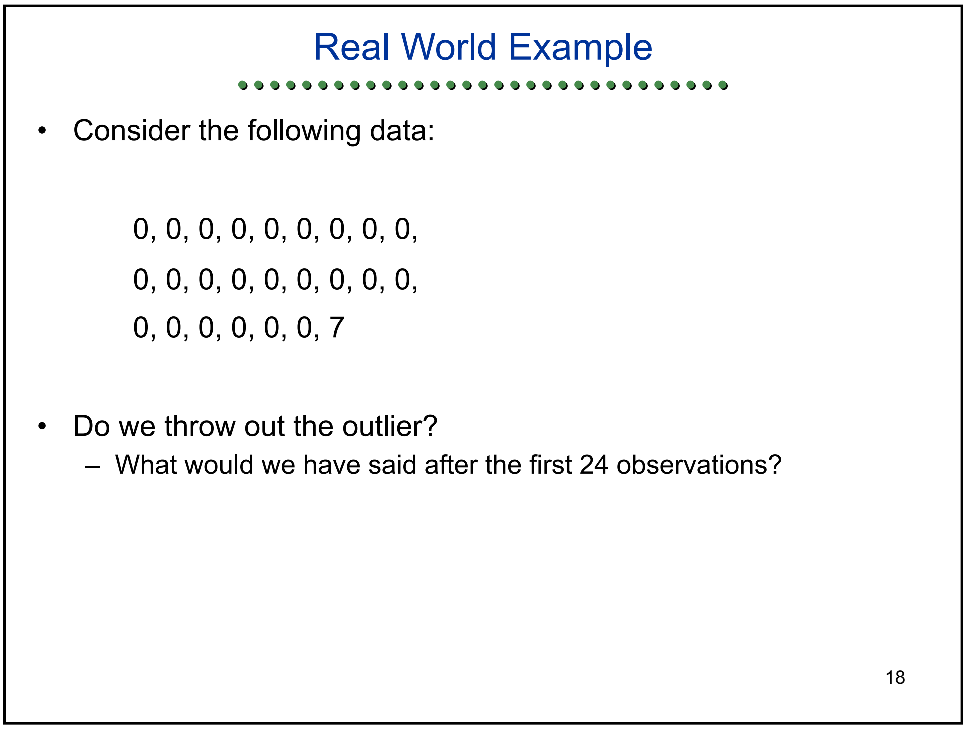


But: Impact of “Three over n”

- 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
    - I use “3.69 over n” rule

17

17



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?

18

18

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

Upper Conf Bnd for 0 Events

- Exact upper confidence bound when all observations are 0

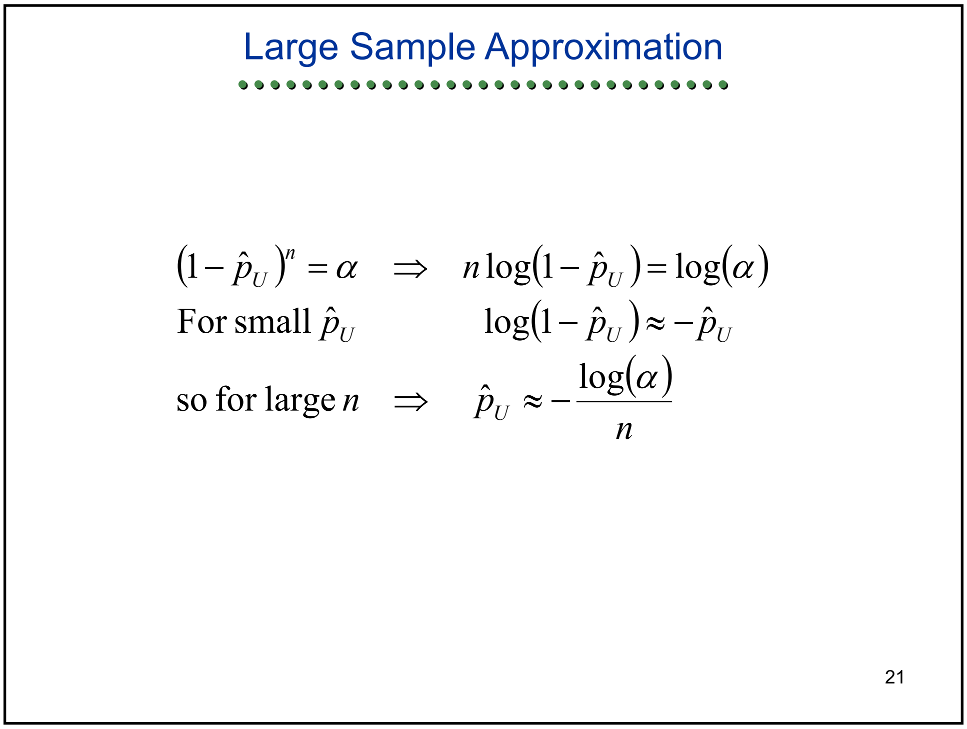
Suppose $Y \sim B(n, p)$ and $Y = 0$ is observed

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

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

$$\Downarrow$$

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



Large Sample Approximation

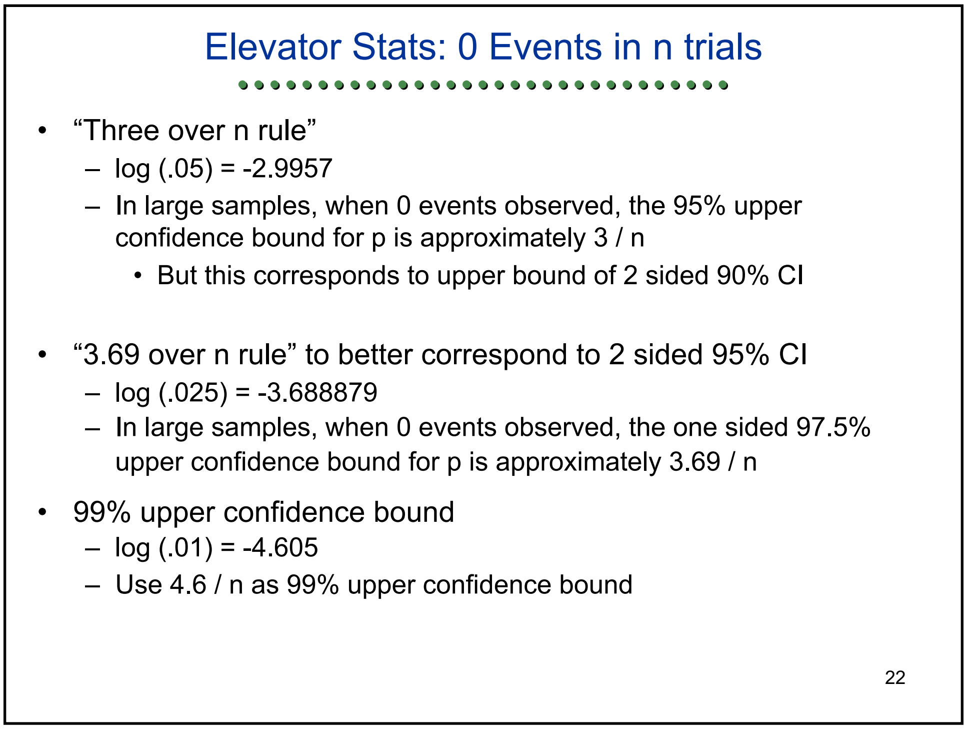
$$\left(1-\hat{p}_U\right)^n = \alpha \quad \Rightarrow \quad n\log\left(1-\hat{p}_U\right) = \log(\alpha)$$

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

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

21

21

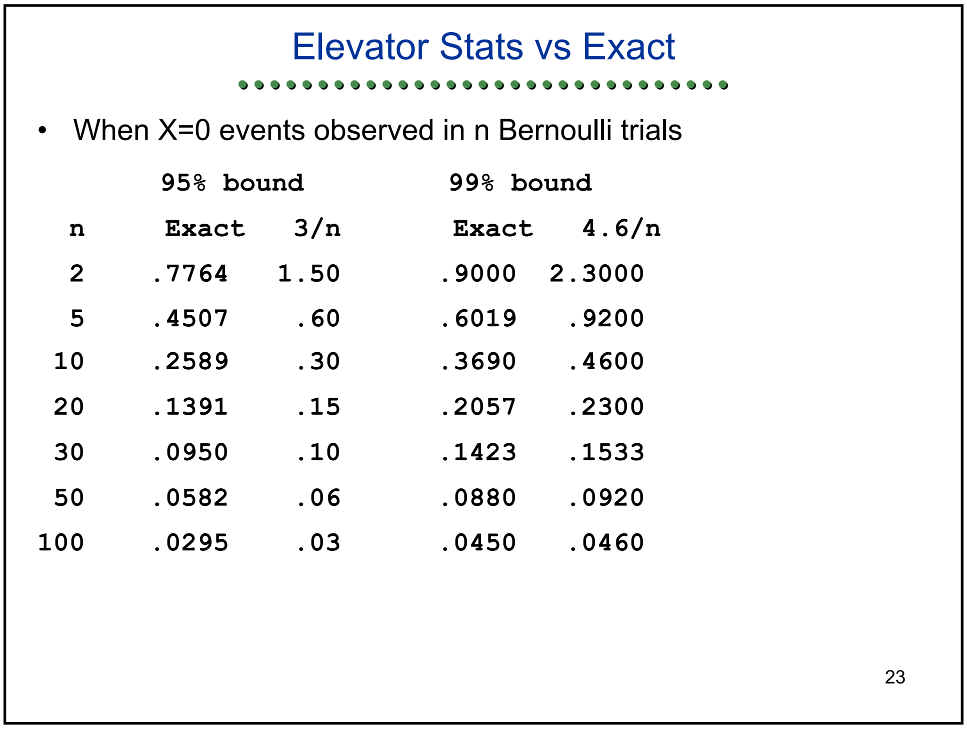


Elevator Stats: 0 Events in n trials

- "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
    - But this corresponds to upper bound of 2 sided 90% CI
- "3.69 over n rule" to better correspond to 2 sided 95% CI
  - log (.025) = -3.688879
  - In large samples, when 0 events observed, the one sided 97.5% upper confidence bound for p is approximately 3.69 / n
- 99% upper confidence bound
  - log (.01) = -4.605
  - Use 4.6 / n as 99% upper confidence bound

22

22



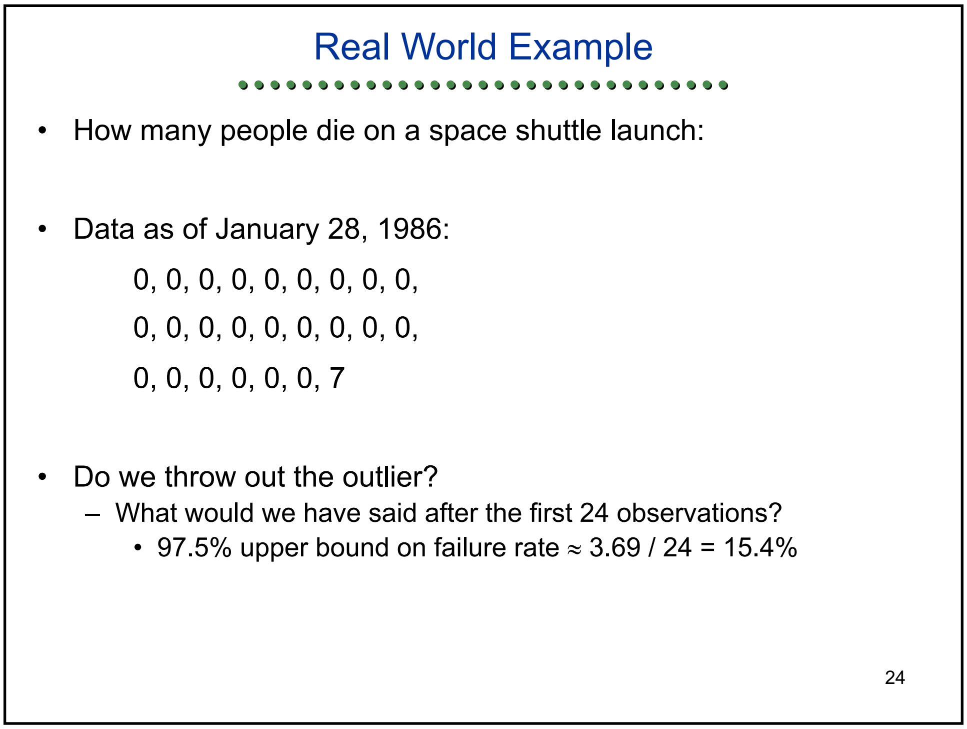
Elevator Stats vs Exact

- 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

23

23



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,

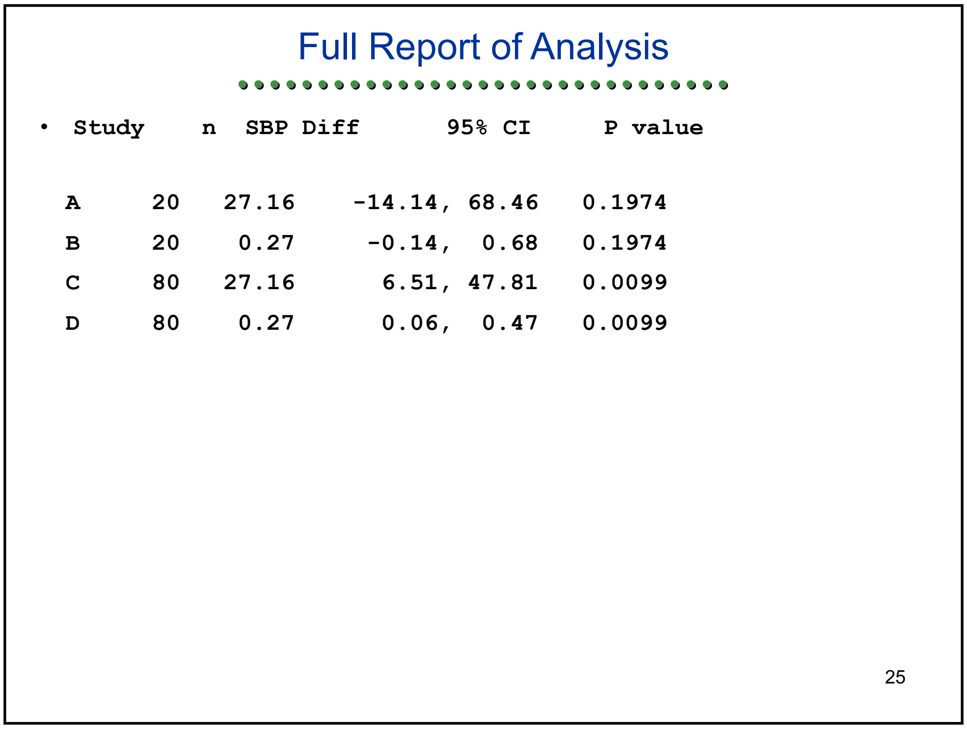
  0, 0, 0, 0, 0, 0, 0, 0, 0,

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

- Do we throw out the outlier?
  - What would we have said after the first 24 observations?
    - 97.5% upper bound on failure rate ≈ 3.69 / 24 = 15.4%

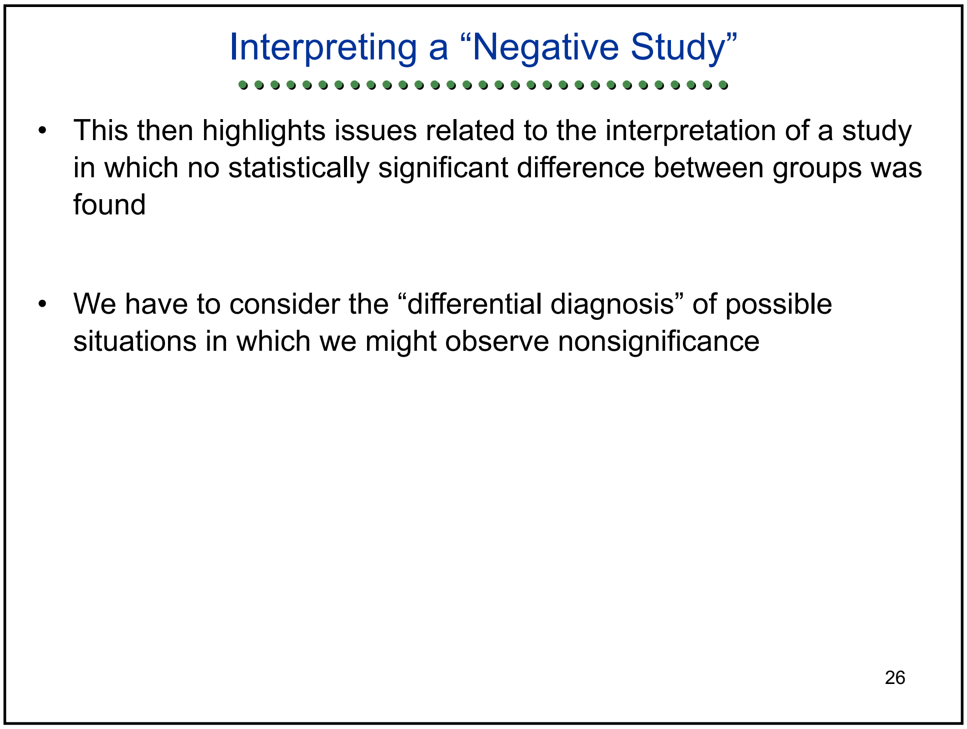
24

24



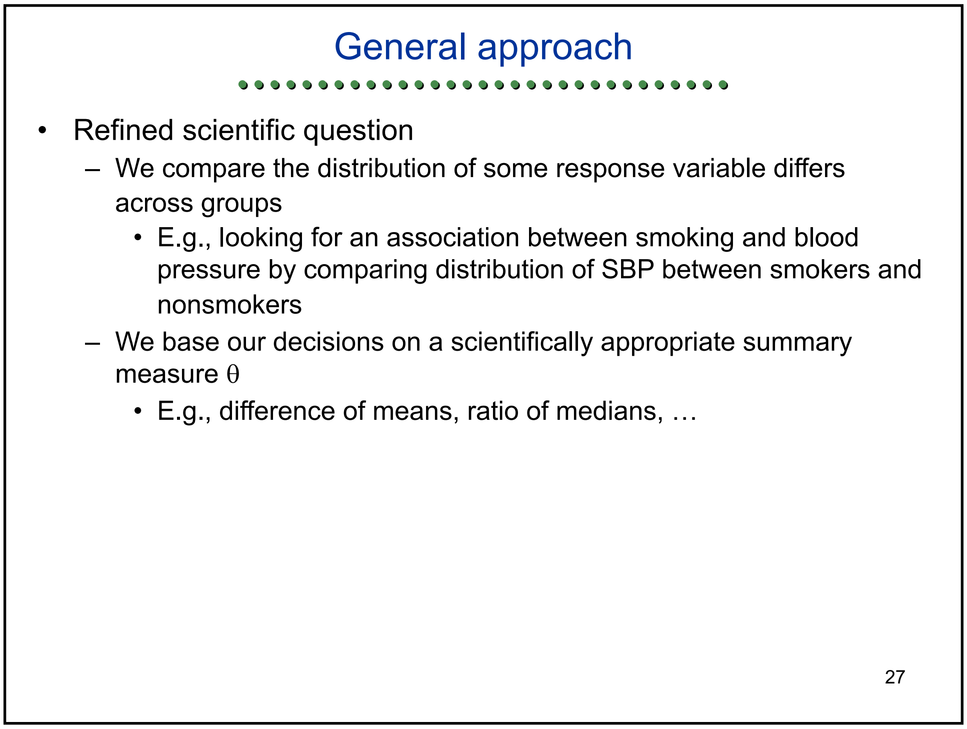
Full Report of Analysis

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



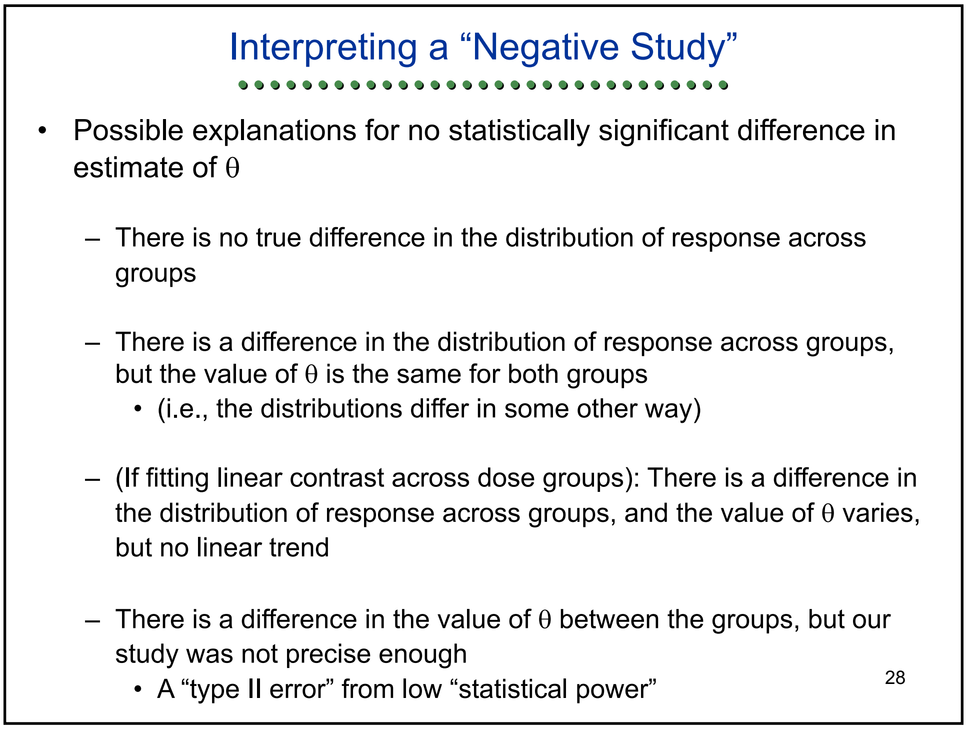
Interpreting a “Negative Study”

- 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



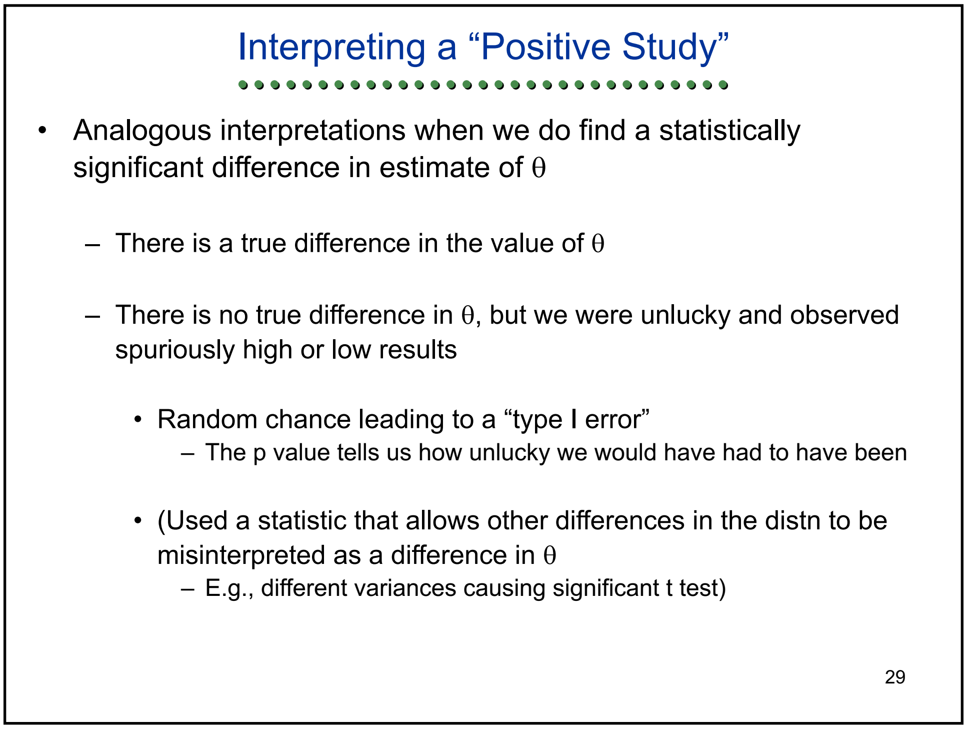
General approach

- 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)
  - (If fitting linear contrast across dose groups): There is a difference in the distribution of response across groups, and the value of θ varies, but no linear trend
  - 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”

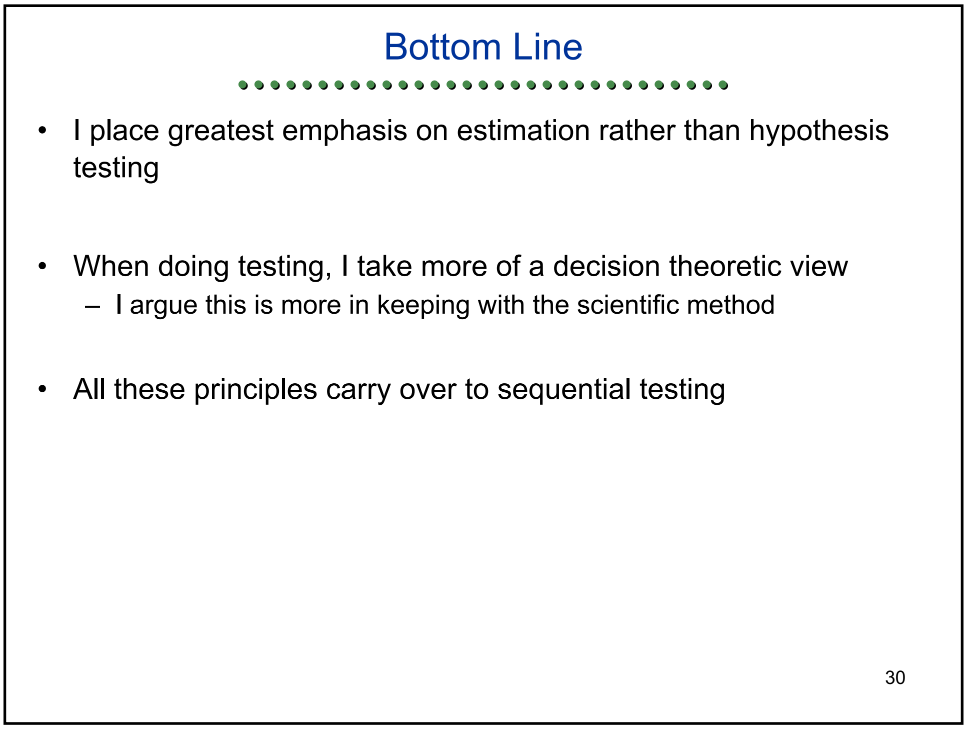


Interpreting a “Positive Study”

- 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)

29

29

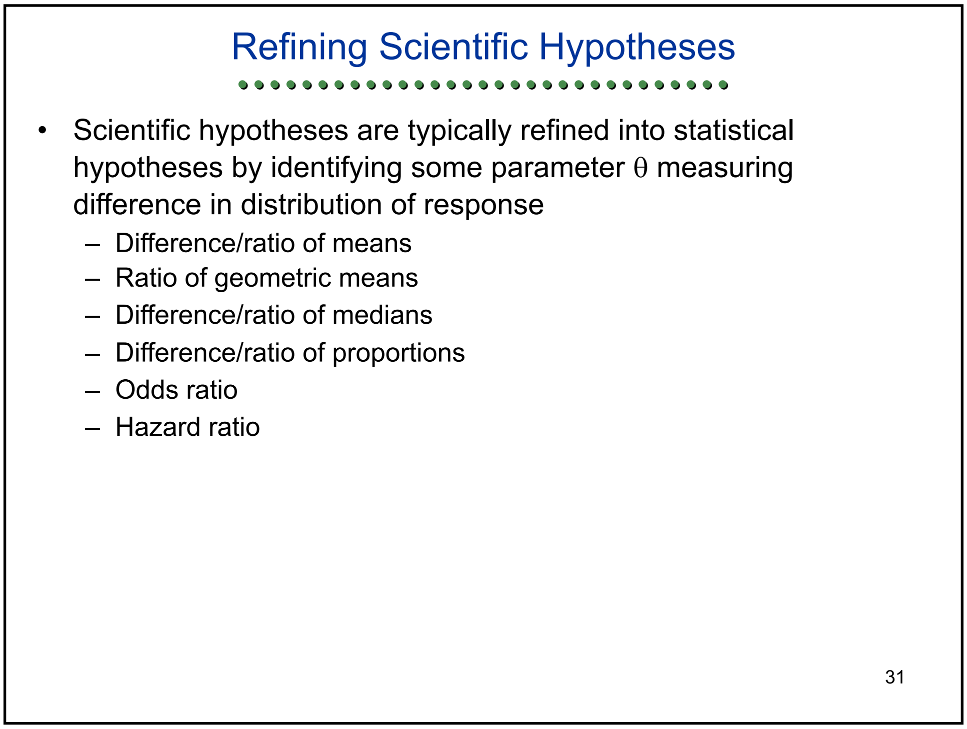


Bottom Line

- 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

30

30

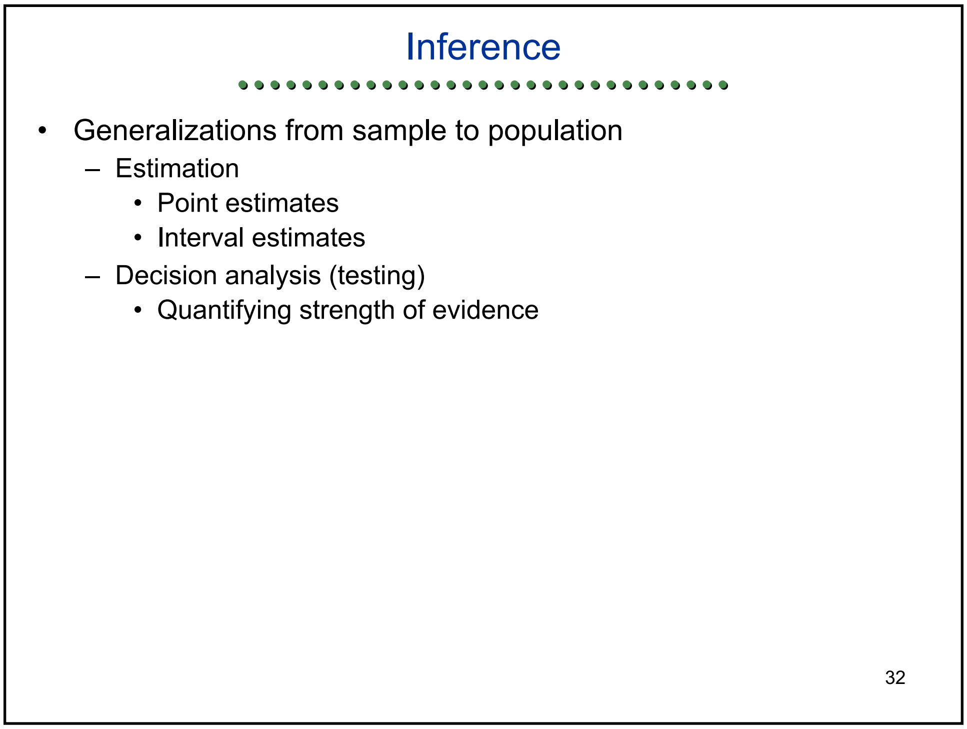


Refining Scientific Hypotheses

- 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

31

31

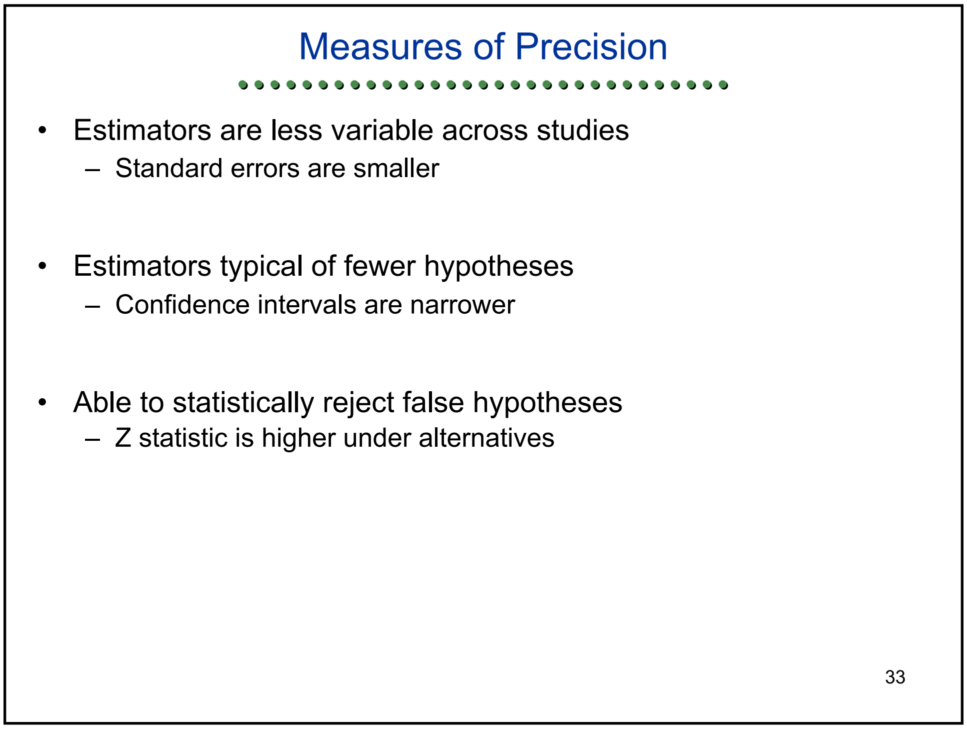


Inference

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

32

32

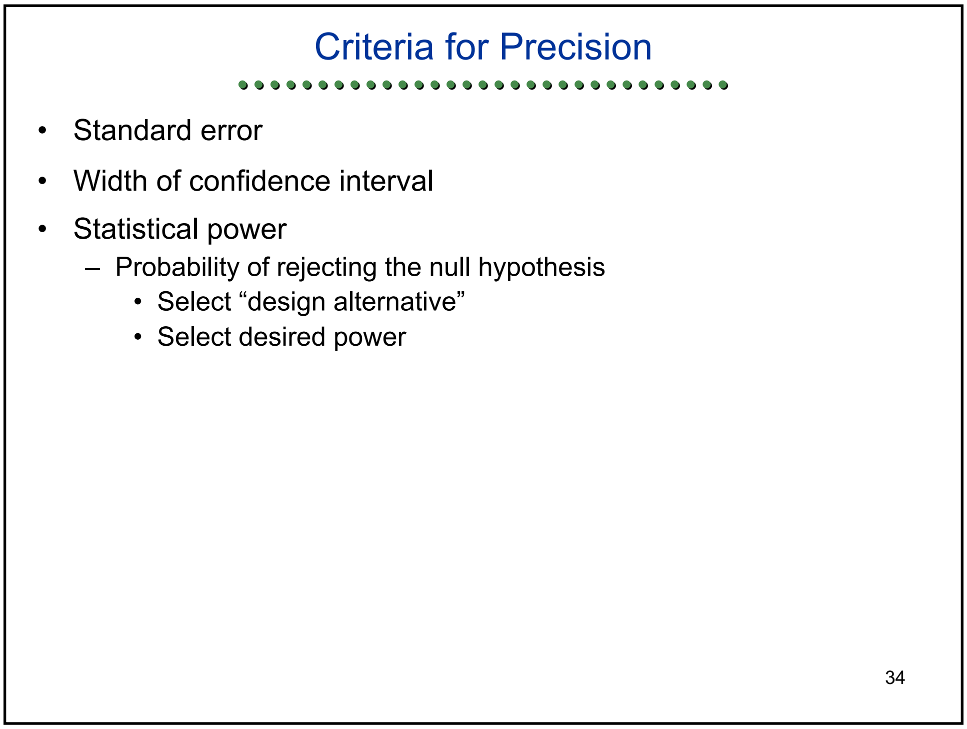


Measures of Precision

- 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

33

33

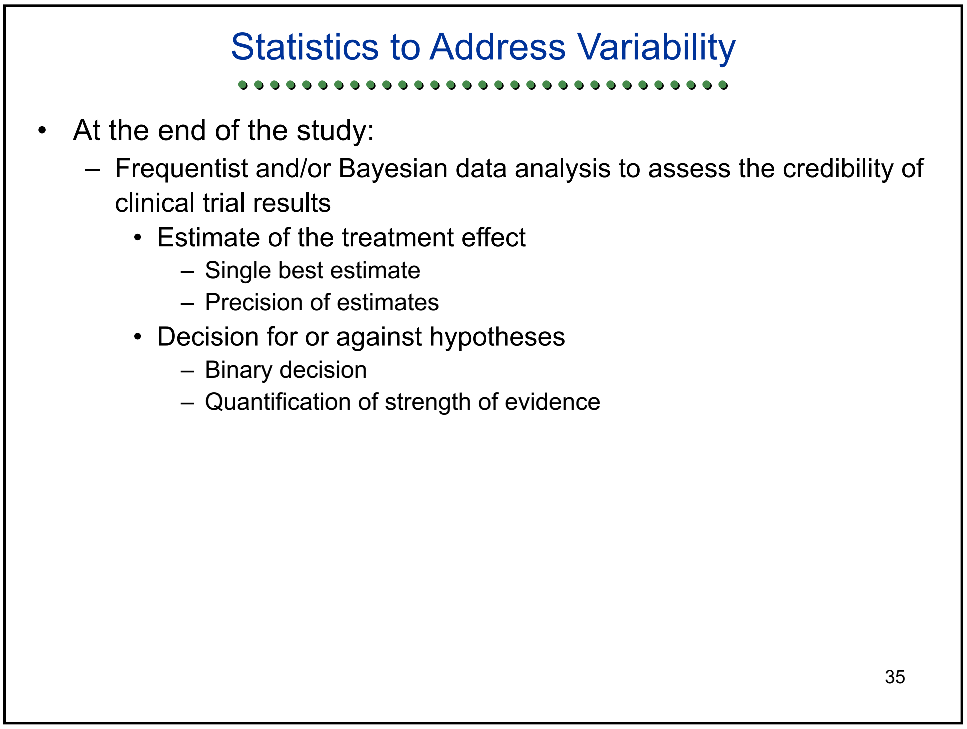


Criteria for Precision

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

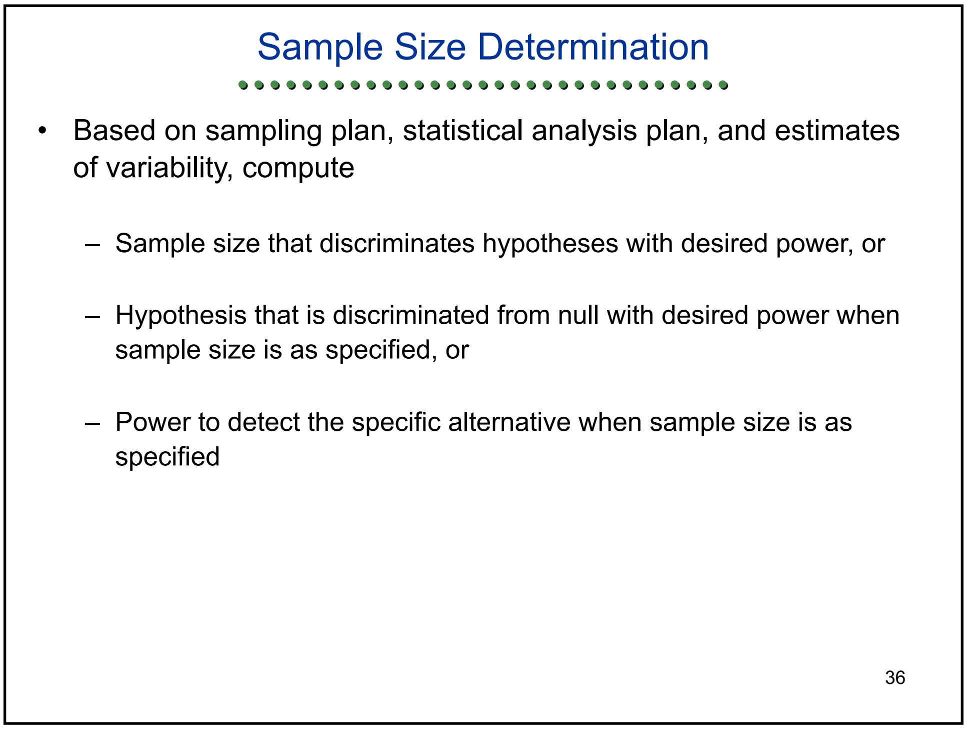
34

34



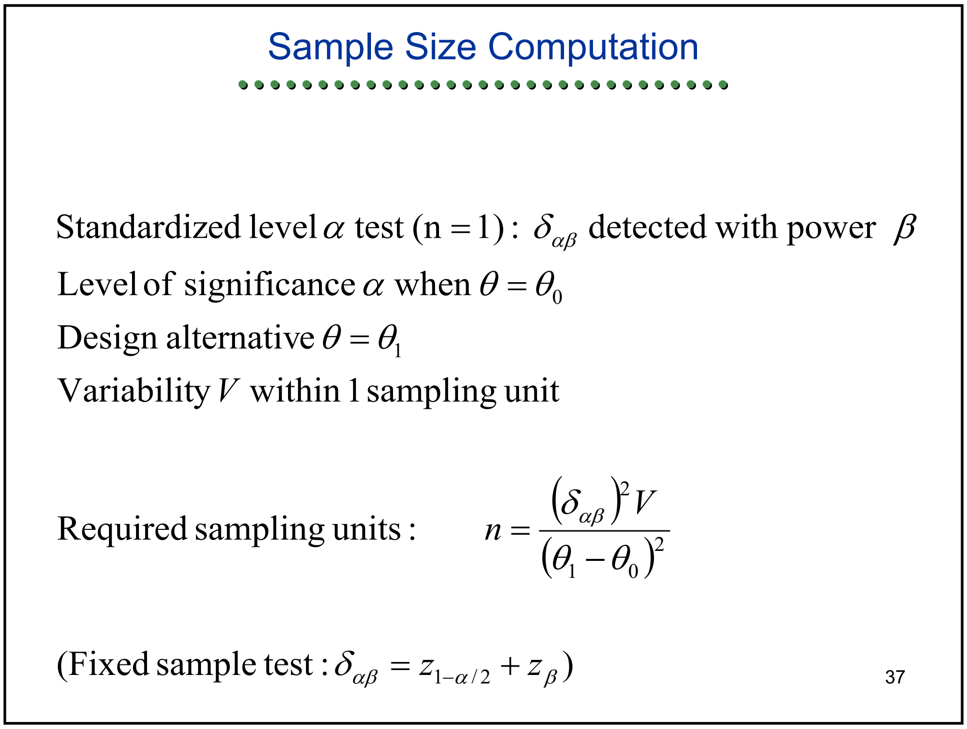
Statistics to Address Variability

- 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 $\theta = \theta_1$

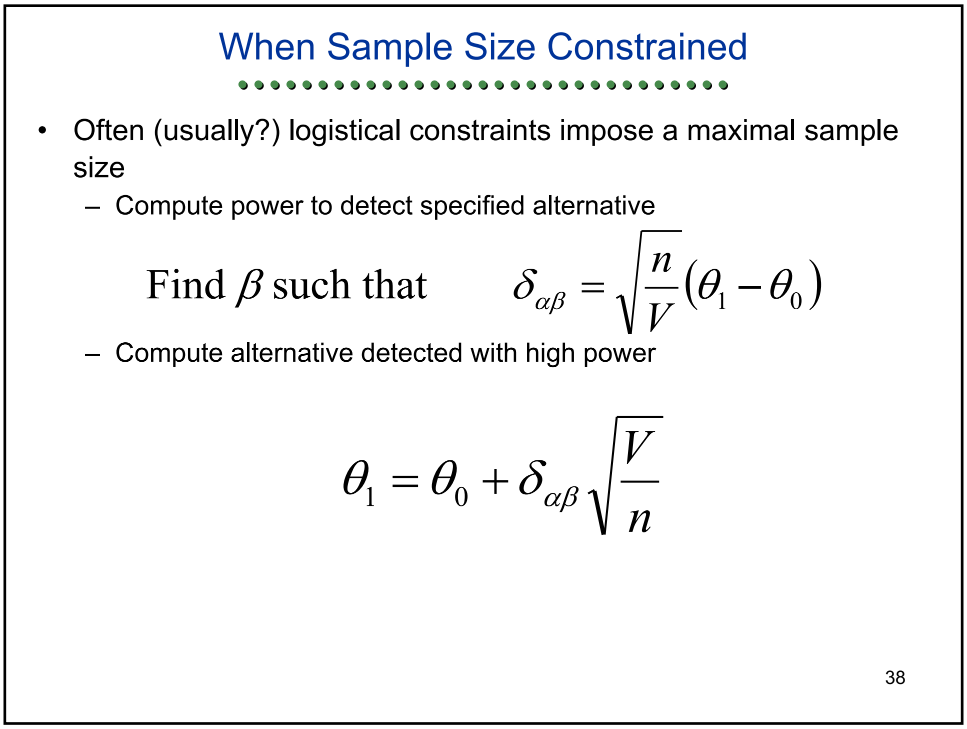
Variability V within 1 sampling unit

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

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

37

37



When Sample Size Constrained

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

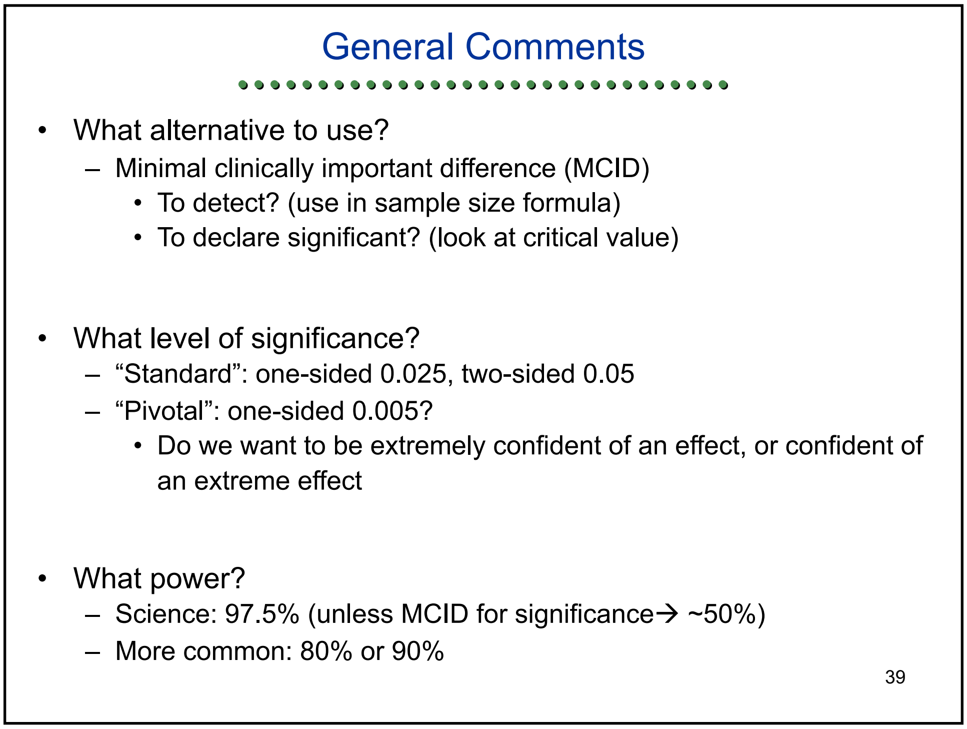
Find β such that $$\delta_{\alpha\beta} = \sqrt{\frac{n}{V}}\left(\theta_1 - \theta_0\right)$$

  - Compute alternative detected with high power

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

38

38

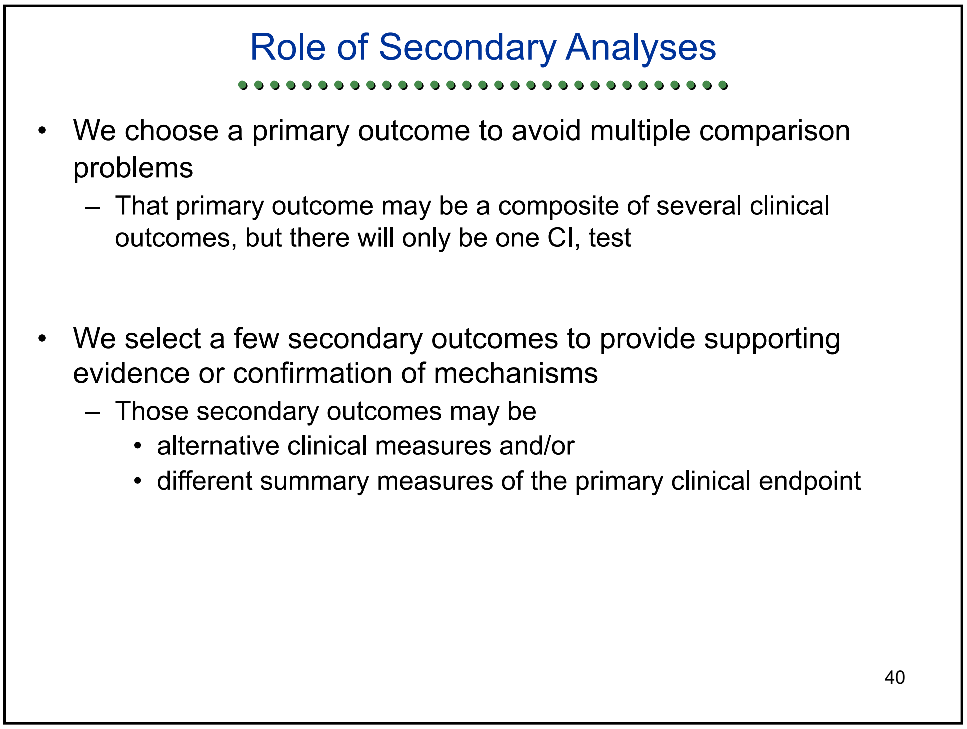


General Comments

- 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%)
  - More common: 80% or 90%

39

39

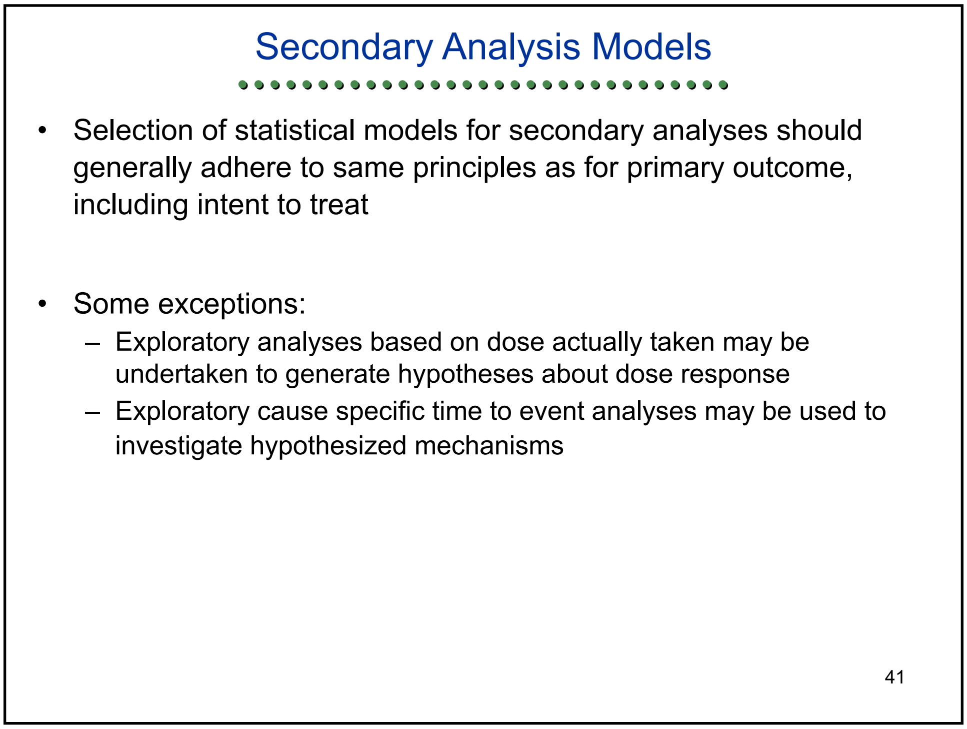


Role of Secondary Analyses

- 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

40

40

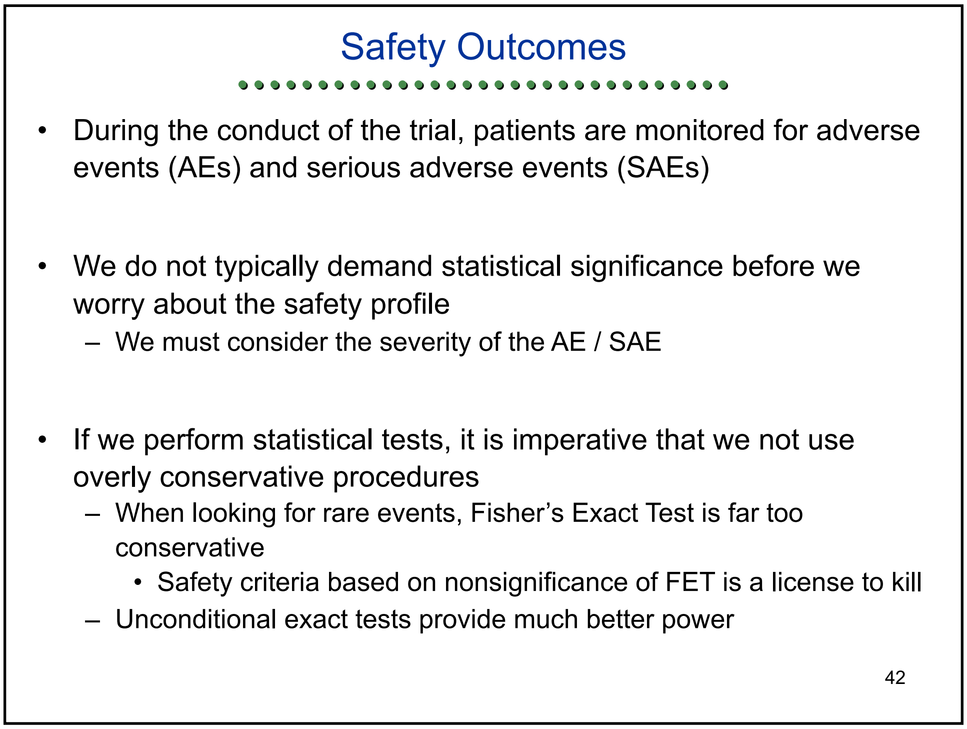


Secondary Analysis Models

- 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

41

41

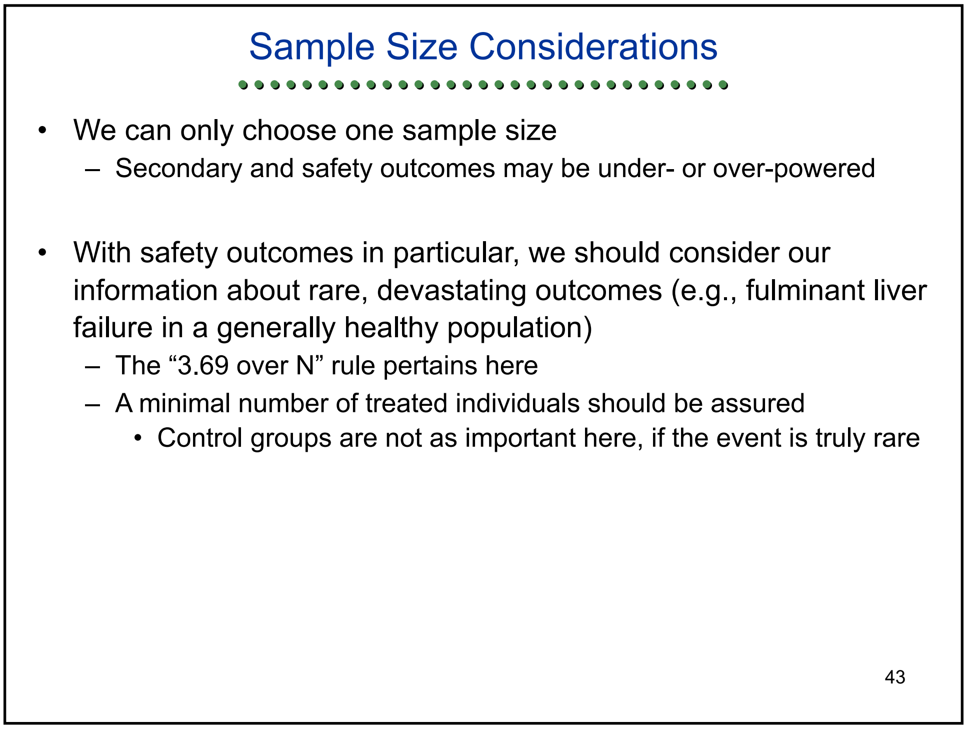


Safety Outcomes

- 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

42

42



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 “3.69 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

43

43