

## Lecture 2: Sample size calculations for common Phase III study designs

- Superiority design for binary responses
- Noninferiority design for binary responses

**Context:** Ronald A. Fisher

- “Father of modern statistical analysis of experiments”
- defined the rules of inference that are used to evaluate RCTs
- methods were developed in the 1920s, applied to problems in agriculture

### Superiority trial

#### 1. Hypotheses refer to the parameter value under $H_0$ :

- Choose a parameter with which to express the contrast between groups (e.g., difference or ratio)
- State  $H_0$  in words: There is no difference between the outcomes of patients treated with the experimental and control interventions.
- Express via a (population) parameter:

$$H_0 : \delta = 0 \text{ versus } H_A : \delta \neq 0,$$

where  $\delta$  is the difference between Experimental and Control group means

\* Hypotheses can be 1-sided or 2-sided

## 2. Select an outcome variable and find its distribution:

e.g., Gaussian:

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad \text{where } x = 0 \text{ for Control and } x = 1 \text{ for Experimental patients}$$

	Parameter	Sample
Control	$\mu_C = \beta_0$	$\bar{Y}_C$
Experimental	$\mu_E = \beta_0 + \beta_1$	$\bar{Y}_E$
Difference	$\delta = \mu_E - \mu_C = \beta_1$	$\bar{Y}_E - \bar{Y}_C$

Does the observed difference reflect a true difference or a chance finding?

\* Treatment effect,  $\delta$ , is relative to an assumed “starting point,”  $\mu_C$ :

- $\mu_C$ , the Control group mean, comes from prior research; secular trends?
- $\delta$  is the smallest clinically relevant effect of interest under  $H_A$

## 3. Plan to base analysis on 2-sample $t$ -test for continuous responses:

$$t_{n_E+n_C-2} = \frac{|\bar{Y}_E - \bar{Y}_C| - \delta}{s_p \sqrt{\frac{1}{n_C} \left(1 + \frac{1}{\gamma}\right)}}$$

where  $\gamma = n_E/n_C$  is the allocation ratio and  $s_p$  is the pooled standard deviation

Note: Effects of  $|\bar{Y}_E - \bar{Y}_C|$ ,  $s_p$ , and  $n_C + n_E$  on chance of rejecting  $H_0$ .

Note: When testing,  $\delta$  and  $s_p$  are defined under  $H_0$ . Note: Trials with  $\gamma = 1$  are referred to as “balanced”:  $n_E = n_C$ .

4. Allow for errors (perfection requires infinite information):

- $\alpha$  = type I error rate (e.g., if  $\alpha = 0.05$  then  $z_{1-\alpha/2} = 1.96$  and  $z_{1-\alpha} = 1.645$ ) . . . interpretation?
- $\beta$  = type II error rate (e.g., if  $\beta = 0.10$  then  $z_{1-\beta} = 1.2816$ ) . . . consequence?

	Sample	Parameter	
		$\delta = 0$	$\delta \neq 0$
Do not reject $H_0$ :	$ \bar{Y}_E - \bar{Y}_C /se < z_{1-\alpha/2}$	TN	FN
Reject $H_0$ :	$ \bar{Y}_E - \bar{Y}_C /se \geq z_{1-\alpha/2}$	FP	TP

\* Re: Figure below:

**A.  $H_0$  true:**

Notice the 1:1 relationship between:

- $\alpha$  (an area)
- Points along the x-axis (“critical values”)

A test can be conducted either by:

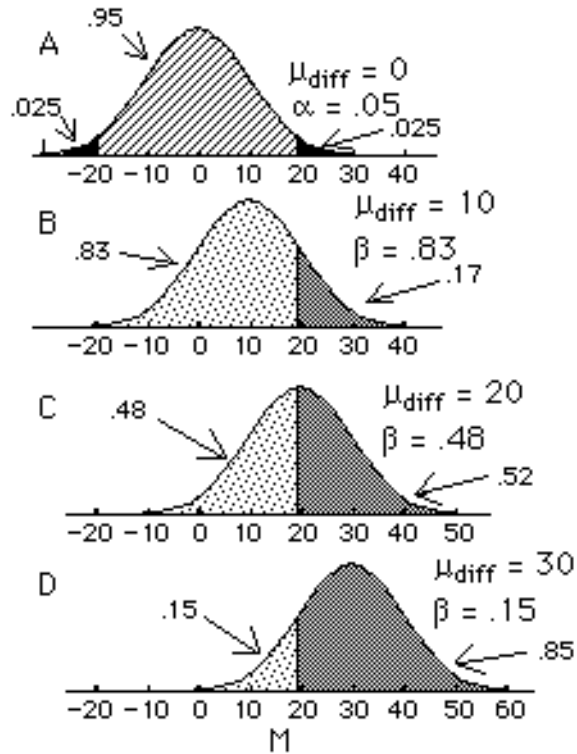
- Comparing a  $P$ -value to a pre-specified  $\alpha$  level (e.g.,  $\alpha/2 = 0.025$  or  $\alpha = 0.05$ )
- Comparing the value of the test statistic to a “critical value”
  - standardized critical value:  $z_{1-\alpha/2} = 1.96$
  - unstandardized (not adjusted for std dev) critical value:  $c_{1-\alpha/2} = 20$

**B,C,D.  $H_A$  true:** The critical value, defined under  $H_0$ , remains fixed ( $c_{1-\alpha/2} = 20$ ) across all figures.

- Under  $H_0$ , the area beyond the critical value(s) is  $\alpha$ .
- Under  $H_A$ , the area beyond the critical value is the power of the test. For constant  $N$ , power increases (A, .05, B. 0.17, C. 0.52, D. 0.85) as  $\delta = \mu_E - \mu_C$  increases (A 0, B 10, C 20, C 30).

Figure 1: Distribution of the test statistic under the *superiority* null hypothesis (A,  $\delta = 0$ ) and three alternative hypotheses (B,  $\delta = 10$ ; C,  $\delta = 20$ ; D,  $\delta = 30$ ). (N.B. Their  $\mu_{\text{diff}}$  is my  $\delta$ .)

x-axes:  $[\bar{Y}_E - \bar{Y}_C]$  (numerator of test statistic); y-axes =  $\Pr\{[\bar{Y}_E - \bar{Y}_C] - \delta > c\}$



\* Note assumption of equal variance under  $H_0$  and  $H_A$ . Note effect of  $\delta$  on power. Note that large values lead to rejecting  $H_0$ .

**5. Calculate sample sizes, assuming equal variance under  $H_0$  and  $H_A$ :**

$$n_C = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (1 + \frac{1}{\gamma}) \sigma^2}{(|\mu_E - \mu_C| - \delta)^2}$$
$$N = (1 + \gamma)n_C$$

- $\sigma$ , the standard deviation of the outcome  $Y$ , assumed to be the same in both groups, comes from prior research; secular trends?
- Note: When calculating power or sample size (e.g., via the formula above),  $\mu_E - \mu_C$  is defined under  $H_A$  and  $\delta$  is defined under  $H_0$ . For superiority trials, the latter typically drops out and is omitted from the notation. Also, then we often just see  $\delta$  in the denominator of this formula, in place of  $\mu_E - \mu_C$ .

Clearer, more general notation for the denominator might be  $\delta_A - \delta_0$ .

## Noninferiority Trial

Shiffman ML, Suter F, Bacon BR, et al; ACCELERATE Investigators. Peginterferon alfa-2a and ribavirin for 16 or 24 weeks in HCV genotype 2 or 3. *N Engl J Med* 2007; **357**: 124-34.

### Study design (from abstract):

**Patients:** HCV genotype 2 or 3

**Intervention:** Treatment duration of  $C$ , 24 weeks;  $E$ , 16 weeks.

**Comparison:** Noninferiority trial (efficacy)

**Outcome:** *Sustained* virologic response, defined as undetectable HCV RNA 24 weeks after EOT  
⇒ binary outcome variable; response rate in control group,  $\pi_C = .80$ ; odds ratio,  $\psi = ?$

### Study design (from paper):

**Patients:** Strata: HCV genotype 2 or 3 and country (132!). Planned  $n_C = n_E = 700 \rightarrow \gamma = n_E/n_C = 1.0$

**Intervention:** 180  $\mu\text{g}$  peginterferon weekly ⇒ at least 90  $\mu\text{g}$  per week ... EO treatment, if stopped  
800 mg ribavirin daily ⇒ at least 600 mg per day ... but could be stopped (???)  
other named med's: permitted but discouraged

**Comparison:** Noninferiority trial using "per protocol" sample

$\alpha = .05$ ,  $\beta = .20$ ; two-sided

Conflicting information!!  $\delta = .06$  or  $\psi = .70$

Conflicting information!!  $\{\pi_C = .80, \pi_E = .74\}$  or  $\{\pi_C = .80, \pi_E = .80\}$

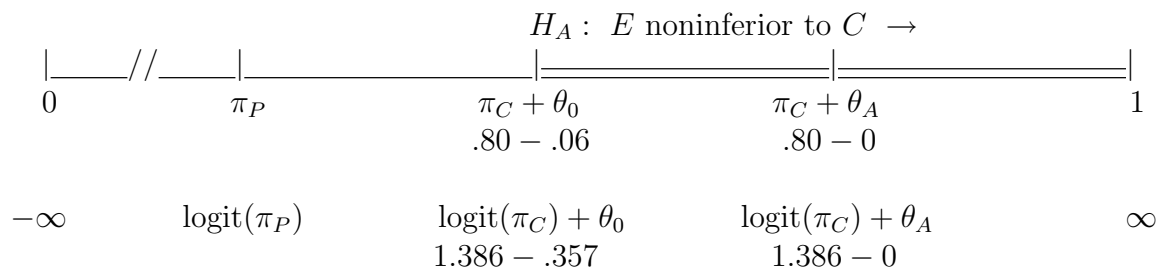
→ (*Trying to comment on  $H_A$ ?*)

**Outcome:** Also *rapid* response

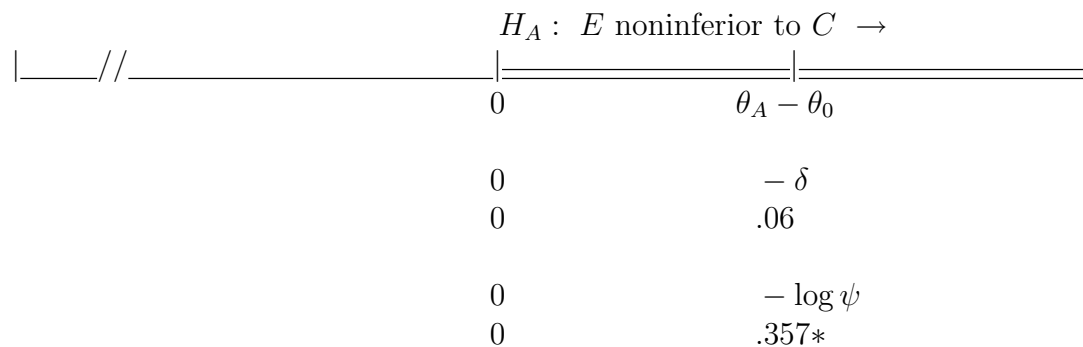
Also *virologic relapse*

Figure 2: (a) Group-specific success rates and log odds of success, and (b) centered differences in relation to null (single line) and alternative (double line) noninferiority hypotheses.

(a) Within-groups scale: Ranges of  $\pi_E$  and  $\text{logit}(\pi_E)$ , respectively:



(b) Between-groups scale, “centered”: Ranges of  $\theta - \theta_0$  when  $\theta = \delta$  and  $\psi$ , respectively:



Note: NI trials are always one-sided. *In this example*, large values provide evidence in support of  $H_A$ .

Note: NI trials are typically defined to have power against the specific alternative of equal groups.

Note: What is role of  $\pi_P$  in NI trials?

**1. Hypotheses refer to the parameter value under  $H_0$ :**

$$H_0 : \delta \leq -.06 \text{ versus } H_A : \delta > -.06, \text{ when } \delta = \pi_E - \pi_C$$

became . . .

$$H_0 : \psi \leq .70 \text{ versus } H_A : \psi > .70, \text{ when } \psi = [\pi_E/(1 - \pi_E)]/[\pi_C/(1 - \pi_C)]$$

In NI trials,  $\theta_0$  (i.e.,  $\delta$  or  $\log \psi$ ) is known as the “noninferiority margin” rather than as the “treatment effect.”

Note: NI trials are one-sided!!

**2. Select an outcome variable and find its distribution:**

e.g., Binomial,  $Y = 1$  if success:

$$\log[\Pr\{Y_i\}/(1-\Pr\{Y_i\})] = \beta_0 + \beta_1 x_i + e_i, \text{ where } x = 0 \text{ for Control and } x = 1 \text{ for Experimental patients}$$

	Parameter	Sample
Control	$\text{logit}(\pi_C) = \beta_0$	$\log[\bar{Y}_C/(n_C - \bar{Y}_C)]$
Experimental	$\text{logit}(\pi_E) = \beta_0 + \beta_1$	$\log[\bar{Y}_E/(n_E - \bar{Y}_E)]$
Difference, log-odds scale	$\log \psi = \text{logit}(\pi_E) - \text{logit}(\pi_C) = \beta_1$	$\log\{[\bar{Y}_E * (n_C - \bar{Y}_C)]/[\bar{Y}_C * (n_E - \bar{Y}_E)]\}$

Why do we work on the log-odds instead of the OR scale?

\* NI margin,  $\psi$  (or  $\delta$ ), is relative to an assumed “starting point,”  $\pi_C$ :

- $\pi_C$ , the Control group mean, comes from prior research; secular trends?
- $\psi$  is the largest loss of efficacy that is clinically allowable

### 3. Plan to base analysis on analog of 2-sample $t$ -test for binomial responses:

$$t_{n_E+n_C-2} = \frac{\log\{[\bar{Y}_E * (n_C - \bar{Y}_C)]/[\bar{Y}_C * (n_E - \bar{Y}_E)]\} - \log \psi}{se},$$

where  $\gamma = n_E/n_C$  is the allocation ratio. No longer use  $s_p$ , the pooled standard deviation!!!

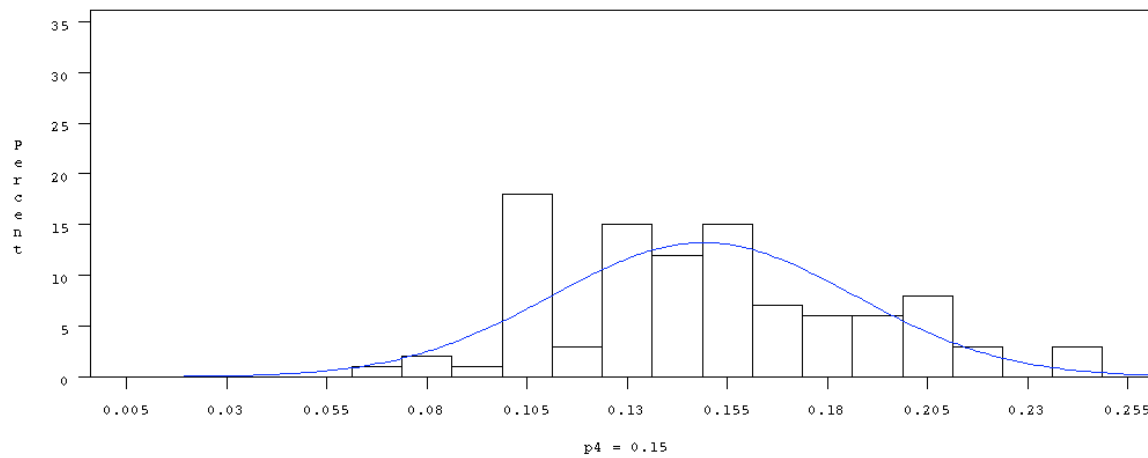
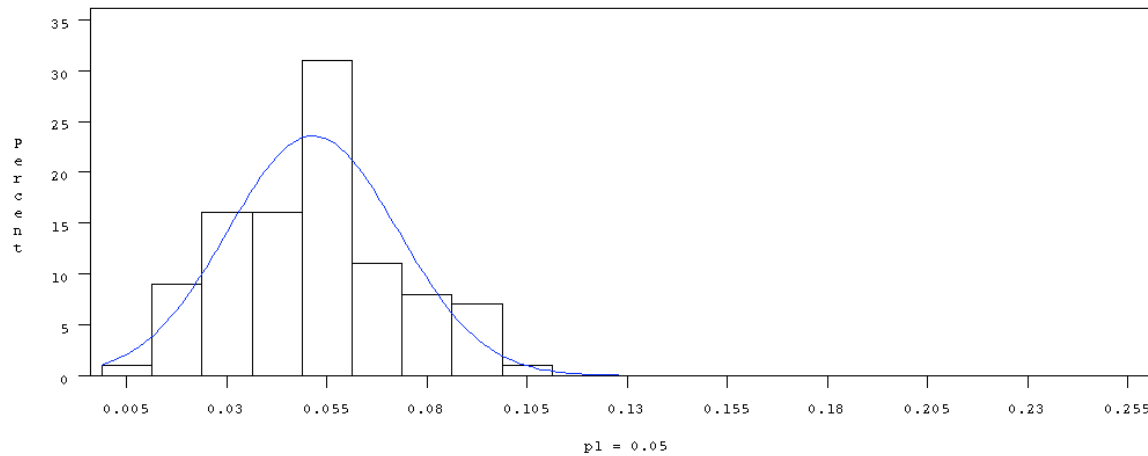
\* Note: For testing,  $se$  is defined under  $H_0$ . In NI trials, group means differ under  $H_0$ !

\* When responses are binary, variances depend on the mean response rate. Thus different variances occur:

- under  $H_0$  and  $H_A$
- by scale of the parameter:  $\delta$  and  $\log \psi$
- by trial type: Superiority or Noninferiority

These differences must be accounted for in sample-size calculations for binomial outcomes.

Figure 3: Random samples of  $N = 100$  independent binomial random variables with mean  $\pi = 0.05$  (*upper*) and mean  $\pi = 0.15$  (*lower*). The variance of the distribution increases as  $\pi$  approaches 0.50. These distributions are approximately normal as  $N$  increases and  $\pi$  approaches 0.50.



**4. Allow for errors (perfection requires infinite information):**

- $\alpha$  = type I error rate (e.g., if  $\alpha = 0.025$  then  $z_{1-\alpha} = 1.96$ ) . . . Why select this value?
- $\beta$  = type II error rate (e.g., if  $\beta = 0.20$  then  $z_{1-\beta} = 0.8416$ ) . . . Why select this value?

**5. Calculate sample sizes, allowing UNEQUAL variances under  $H_0$  and  $H_A$ :**

$$n_C = \frac{(z_{1-\alpha} \hat{\sigma}_0 + z_{1-\beta} \sigma_A)^2}{(\theta_A - \theta_0)^2}, \quad \text{and} \quad N = (1 + \gamma)n_C.$$

In the Shiffman et al (2007) NI trial example,

- on the  $\delta$  scale:  $\theta_0 = -0.06$  and  $\theta_A = 0$ ,
- on the  $\log \psi$  scale:  $\theta_0 = -0.357$  and  $\theta_A = 0$ .

***Noninferiority Trial***

For binary response data,  $\sigma_0$  and  $\sigma_A$  – standard deviations under  $H_0$  and  $H_A$ , respectively – typically differ, and these terms differ between the  $\delta$  and  $\log \psi$  scales.

Data	Parameters	
	$H_A : \delta = 0$	$H_0 : \delta \neq 0$
	Both	$C$ $E$
$Y = 1$	$\pi$	$\pi_C$ $\pi_C + \delta$
$Y = 0$	$1 - \pi$	$1 - \pi_C$ $1 - (\pi_C + \delta)$

$$\begin{aligned} \text{Alternative hypothesis: } \sigma_A(\hat{\delta}) &= \left\{ \left(1 + \frac{1}{\gamma}\right) \pi(1 - \pi) \right\}^{0.5} \\ \text{Null hypothesis: } \hat{\sigma}_0(\hat{\delta}) &= [\hat{\pi}_C(1 - \hat{\pi}_C) + (\hat{\pi}_C + \delta)\{1 - (\hat{\pi}_C + \delta)\}/\gamma]^{0.5} \end{aligned}$$

- How are the overall response rate,  $\pi$ , and the group-specific response rates,  $\{\pi_C, \pi_E\}$ , related?

By starting with the “likelihood equation” under  $H_0$  for a noninferiority trial,

$$L_{H_0}(\delta) = [\pi_C^{x_C}(1 - \pi_C)^{n_C - x_C}] [(\pi_C + \delta)^{x_E}\{1 - (\pi_C + \delta)\}^{n_E - x_E}]$$

it can be shown that  $\pi$  and  $\delta$  (and  $\sigma_A$ ) are fixed, known parameters, whereas  $\pi_C$  is unknown and must be estimated (because  $\hat{\sigma}_0$  is needed in the sample-size equation).

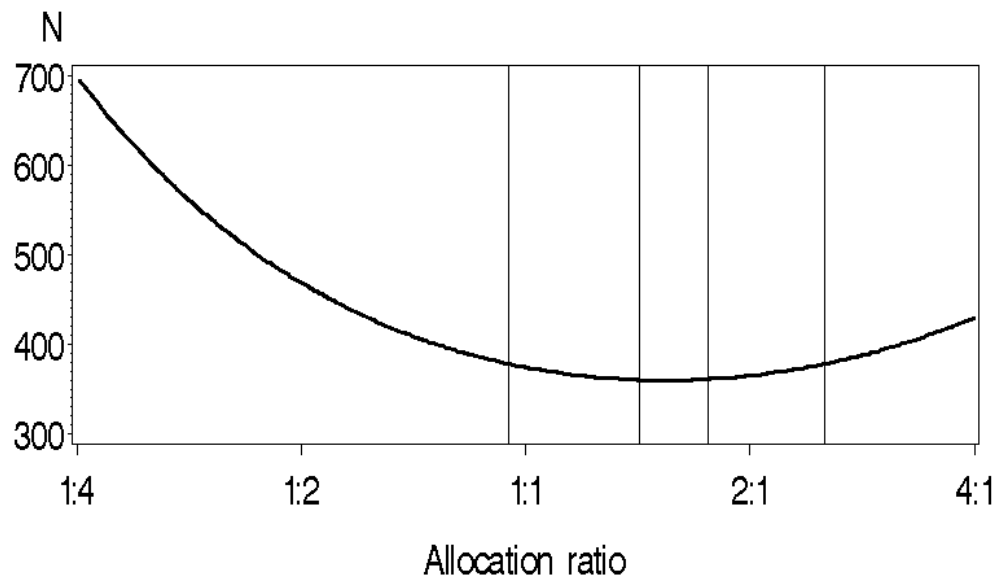
The estimate of  $\pi_C$  depends on the allocation ratio,  $\gamma = n_E/n_C$ .

$$\begin{aligned} \pi &= (1 - \omega)\hat{\pi}_C + \omega(\hat{\pi}_C + \delta), \\ \text{where } \omega &= \frac{\gamma}{\gamma + \frac{(\hat{\pi}_C + \delta)\{1 - (\hat{\pi}_C + \delta)\}}{\hat{\pi}_C(1 - \hat{\pi}_C)}} \end{aligned} \tag{1}$$

The specific value of  $\gamma$  that minimizes  $N = n_E + n_C$  can be calculated iteratively (special software needed); typically  $\gamma \neq 1$ .

- Investigators must specify  $\delta$  (minimum clinically acceptable difference) and  $\pi$  (from where?), and  $\gamma$  (from where?).

Figure 4: For test statistic  $T(\hat{\delta})$  and  $\alpha = \beta = 0.05$ , relationship of overall sample size ( $N$ ) to allocation ratio ( $\gamma$ ) for noninferiority design  $\{\pi = 0.087, \delta = 0.10\}$ .  $N$  is at its minimum between the inner vertical reference lines and is within 5% of its minimum between the outer lines; optimal  $\gamma \in (1.42, 1.67)$ .



**Superiority Trial:**

Data	Parameters		
	$H_A : \delta \neq 0$		$H_0 : \delta = 0$
	$C$	$E$	Both
$Y = 1$	$\pi_C$	$\pi_C + \delta$	$\pi$
$Y = 0$	$1 - \pi_C$	$1 - (\pi_C + \delta)$	$1 - \pi$

- Alternative hypothesis:  $\sigma_A(\hat{\delta}) = [\pi_C(1 - \pi_C) + (\pi_C + \delta)\{1 - (\pi_C + \delta)\}/\gamma]^{0.5}$
- Null hypothesis (compare with pooled standard deviation on p.2 of notes):  $\hat{\sigma}_0(\hat{\delta}) = \left[ \left(1 + \frac{1}{\gamma}\right) \hat{\pi}(1 - \hat{\pi}) \right]^{0.5}$
- How are the overall response rate and the group-specific response rates related?

By starting with the “likelihood equation” under  $H_0$  for a superiority trial,

$$L_{H_0}(\pi) = [\pi^{x_E}(1 - \pi)^{n_E - x_E}][\pi^{x_C}(1 - \pi)^{n_C - x_C}],$$

it can be shown that  $\pi_C$  and  $\delta$  are fixed, known parameters, whereas  $\pi$  is unknown and must be estimated (because  $\hat{\sigma}_0$  is needed for sample-size equation).

The estimate of  $\pi$  depends on the allocation ratio,  $\gamma = n_E/n_C$ :

$$\hat{\pi} = (1 - \omega)\pi_C + \omega(\pi_C + \delta), \quad \text{where } \omega = \frac{\gamma}{\gamma + 1}.$$

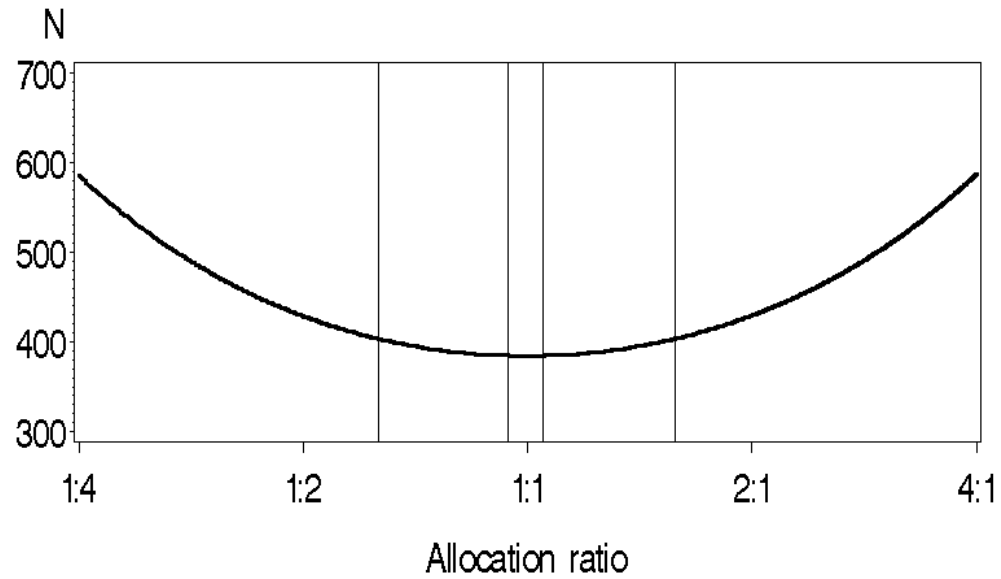
For the superiority design,  $\gamma = 1$  typically minimizes  $N = n_E + n_C$ . Then:

$$\hat{\pi} = \frac{\pi_E + \pi_C}{2}$$

$\Rightarrow \pi$  lies exactly halfway between  $\pi_E$  and  $\pi_C$ .

- Investigators must specify  $\pi_C$  (from past research) and  $\delta$  (minimum difference of clinical interest).

Figure 5: For test statistic  $T(\hat{\delta})$  and  $\alpha = \beta = 0.05$ , relationship of overall sample size ( $N$ ) to allocation ratio ( $\gamma$ ) for superiority design  $\{\pi_C = 0.15, \delta = -0.10\}$ . The overall sample size is at its minimum between the inner vertical reference lines and is within 5% of its minimum between the outer lines; optimal  $\gamma = 1.0$ .

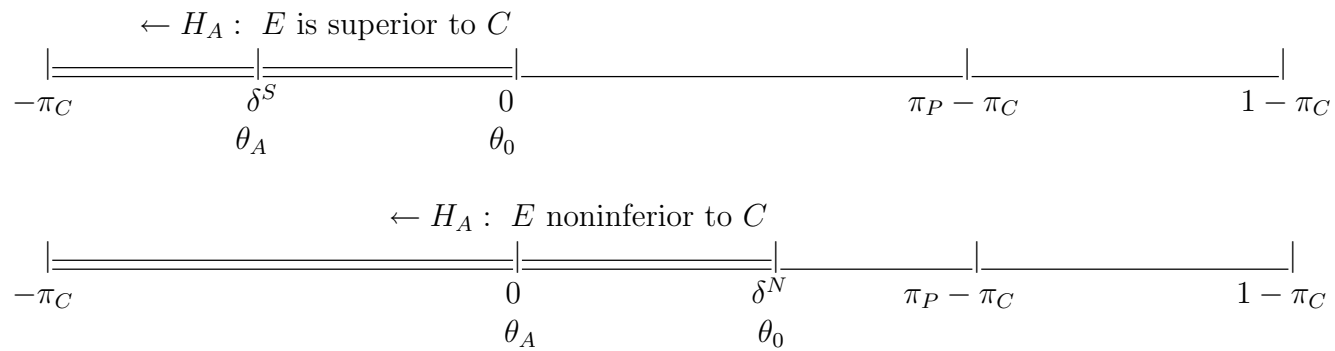


Examples / Software: Design Algorithms

Setting where test is based on  $\hat{\delta} = \hat{\pi}_E - \hat{\pi}_C$  and small  $\hat{\Delta}$  is desired:

Figure 6: Null (single line) and alternative (double line) hypotheses of superiority and noninferiority designs in relation to the difference between group-specific response rates,  $\Delta = \pi_E - \pi_C$ .

$E$  = Experimental,  $C$  = Control, and  $P$  = Placebo group;  $\pi$  = response rate.



## Validity of Statistical Test:

### *Superiority Trial:*

Study Claim	True State	
	$H_A : \Delta < 0$	$H_0 : \Delta \geq 0$
$\hat{\Delta} < 0$ , $E$ superior to $C$	$TP$ / Power	$FP$ / Type 1
$\hat{\Delta} \geq 0$ , $E$ not superior to $C$	$FN$ / Type 2	$TN$

### *Noninferiority Trial:*

Study Claim	True State	
	$H_A : \Delta < \delta^N$	$H_0 : \Delta \geq \delta^N$
$\hat{\Delta} < \delta^N$ , $E$ noninferior to $C$	$TP$ / Power	$FP$ / Type 1
$\hat{\Delta} \geq \delta^N$ , $E$ inferior to $C$	$FN$ / Type 2	$TN$

The following hold for both designs because the hypotheses are in the same directions:

$$\text{Power} = \text{Sensitivity} = TP / (TP + FN)$$

Type 2 error rate =  $\Pr\{\text{Do not Reject } H_0 \text{ when } H_0 \text{ is false}\} = 1 - \text{Power};$

Risk: Withhold the better treatment from future patients

Type 1 error rate =  $\Pr\{\text{Reject } H_0 \text{ when } H_0 \text{ is true}\} = FP / (FP + TN) = 1 - \text{Specificity};$

Risk: Expose future patients to a less effective treatment + side effects

ROC curve is defined by x-axis=  $TP$  rate ( $\alpha$ ), y-axis=  $FP$  rate ( $1 - \beta$ )

- diagonal ref line [ $\alpha = 1 - \beta$ ]: lower left to upper right
- upper left corner of plot is optimal: low  $\alpha$ , high  $1 - \beta \Rightarrow$  both errors are small

## Rationale for noninferiority study design:

The experimental regimen has the *disadvantage* of lower effectiveness.

It has the *advantage* of a shorter course, which may result in

- fewer treatment-related adverse events due to shorter exposure
- if regimen more toxic, routine monitoring of patients for hepatotoxicity should prevent long-term harm
- greater compliance and less development of treatment resistance
- (especially in developing countries) less reliance on health care personnel and resources

This research question calls naturally for a randomized trial with a one-sided noninferiority design.

### Concerns:

- Has the effectiveness of the Control been established and quantified?
- Is the answer above known for the eligible patient population?

When designing noninferiority trials, must keep  $\pi_P$  ( $P =$  Placebo group) in mind:

- Real interest is in effect of  $E$  relative to  $P$ ; Expect to infer:  $\pi_E \in (\pi_P, \pi_C)$ .
- Assumes past evidence that  $\pi_C > \pi_P$  (i.e.,  $C$  is safe & effective). If  $C$  has become standard of care, cannot ethically withhold it from patients.
- How does estimate of  $\pi_C$  in current trial compare with past evidence? If too different, then cannot be sure that  $\pi_E > \pi_P$ .
  - Best if evidence of  $\pi_C > \pi_P$  was gathered under identical circumstances (patient population, trialist expertise).
  - To allow for trial-to-trial variability,  $\pi_C + \delta$  should not be too close to  $\pi_P$  (e.g., no more than halfway between  $\pi_C$  and  $\pi_P$ ).