

Foundations of Basic Sample Size Calculations

Lance Ford, PhD

Assistant Professor of Research, Biostatistics

Department of Biostatistics and Epidemiology

Hudson College of Public Health

Lance-Ford@ouhsc.edu

Invited COBRE Lecture

October 28, 2022

Objectives

1. Select the correct type of hypothesis test for your research question (superiority, non-inferiority, or equivalency)
2. Describe the required components for basic power and sample size calculations
 - a) Continuous outcomes
 - b) Dichotomous outcomes
3. Explain the relationships between type I error, power, sample size, and effect size

Types of Hypothesis Tests

Hypotheses

- Null hypothesis: H_0
 - Typically, a statement of no treatment effect
 - Assumed true until evidence suggests otherwise
 - Example: H_0 : No difference in mean diastolic blood pressure (DBP) between treatment groups

Hypotheses

- Alternative hypothesis: H_A
 - Reject null hypothesis in favor of alternative hypothesis
 - Different types:
 - Superiority
 - Non-Inferiority
 - Equivalence

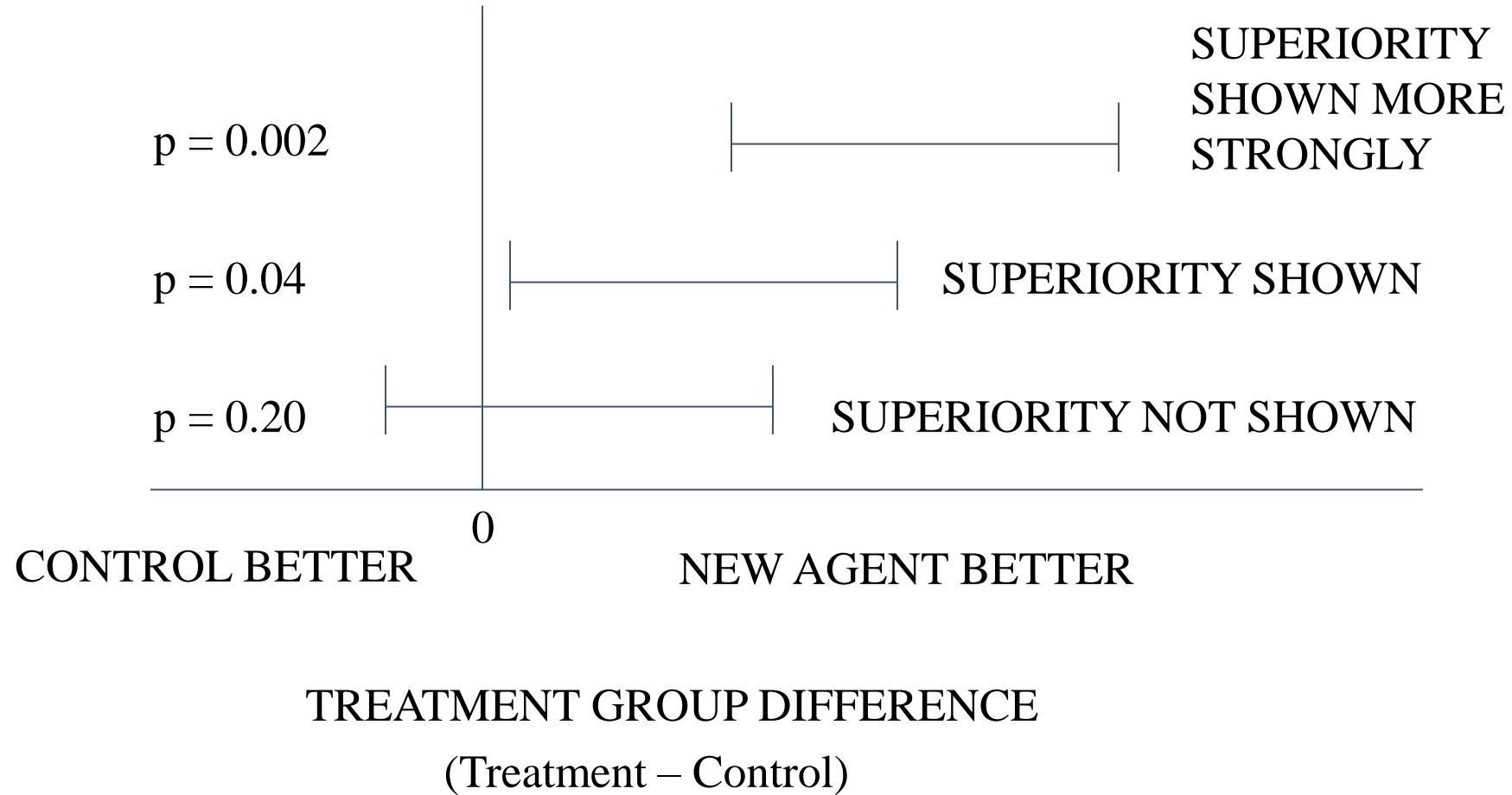
Hypothesis of Superiority

- Example: A trial with the primary objective of showing that response to the investigational product is superior to a comparative agent (active or placebo control)
- Is treatment A better than treatment B?
- Are these two groups different regarding response time?

Superiority

- Designed to detect a difference between treatments
- Test of statistical significance
- Observed difference:
 - Point estimate = $\mu_A - \mu_B$ or $p_A - p_B$
 - Statistical significance: 95% CI should not include 0
 - Clinically relevant?

Superiority



Relationship between significance tests and CIs

Test Your Knowledge!

- Example:

Set-up:

We perform a superiority study and fail to find statistical significance. Can we conclude groups are equivalent by default?

Response:

Failure to show a significant difference is not the same as proving equivalence!

Hypothesis of Non-Inferiority

- Example: A trial with the primary objective of showing that the response to the investigational product is not clinically inferior to a comparative agent (active or placebo control)
- Tests whether a new treatment is not worse than an active treatment it is being compared to
- May be safer and easier to take or cause fewer side effects.
- Non-inferiority trials helpful when a placebo cannot be used

Non-inferiority Example

- Example:

The primary goal for this study was to determine whether fluconazole would be as effective (or nearly as effective) as amphotericin B in preventing the relapse of cryptococcal meningitis in patients with AIDS.

It was thought that the reduced toxicity and oral administration of fluconazole might give it an advantage over amphotericin B, even if fluconazole was slightly less effective.

Fluconazole vs amphotericin B in prevention of relapse of cryptococcal meningitis (Powderly, Saaf et al, NEJM (1992))

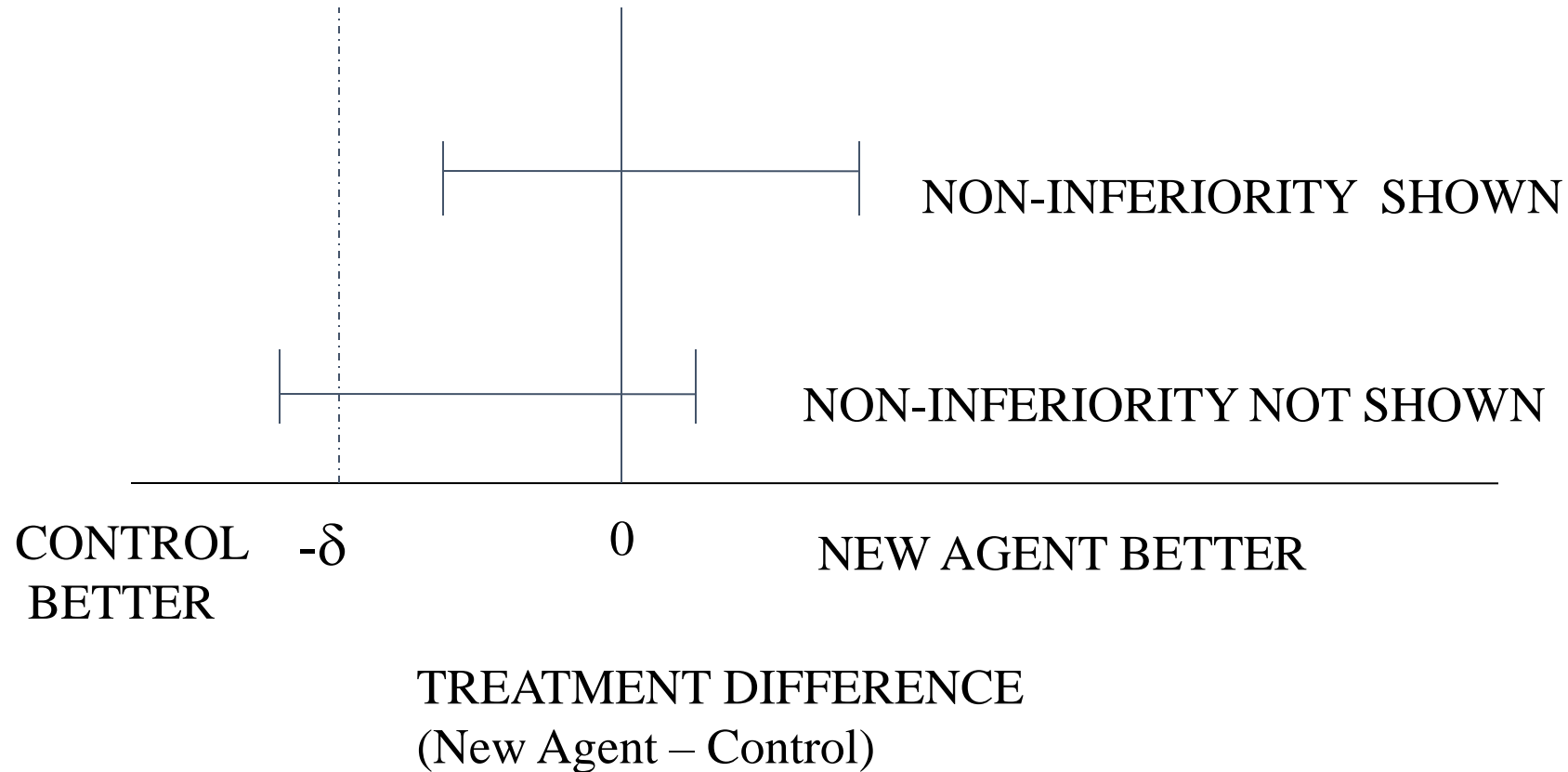
Equivalence Margin (δ)

- This margin is the largest difference between groups that can be judged as still being clinically non-inferior
- The margin should be specified in the protocol
 - Needs to be specified in sample size calculations
- The choice of the equivalence margin should be justified clinically

Non-Inferiority Trial

- Statistical analysis is generally based on the use of confidence intervals. A one-sided confidence interval should be used for non-inferiority trials.
- 1-sided 97.5% CI should lie entirely to the right of the value $-\delta$

Non-Inferiority



Confidence interval approach to analysis of non-inferiority trial

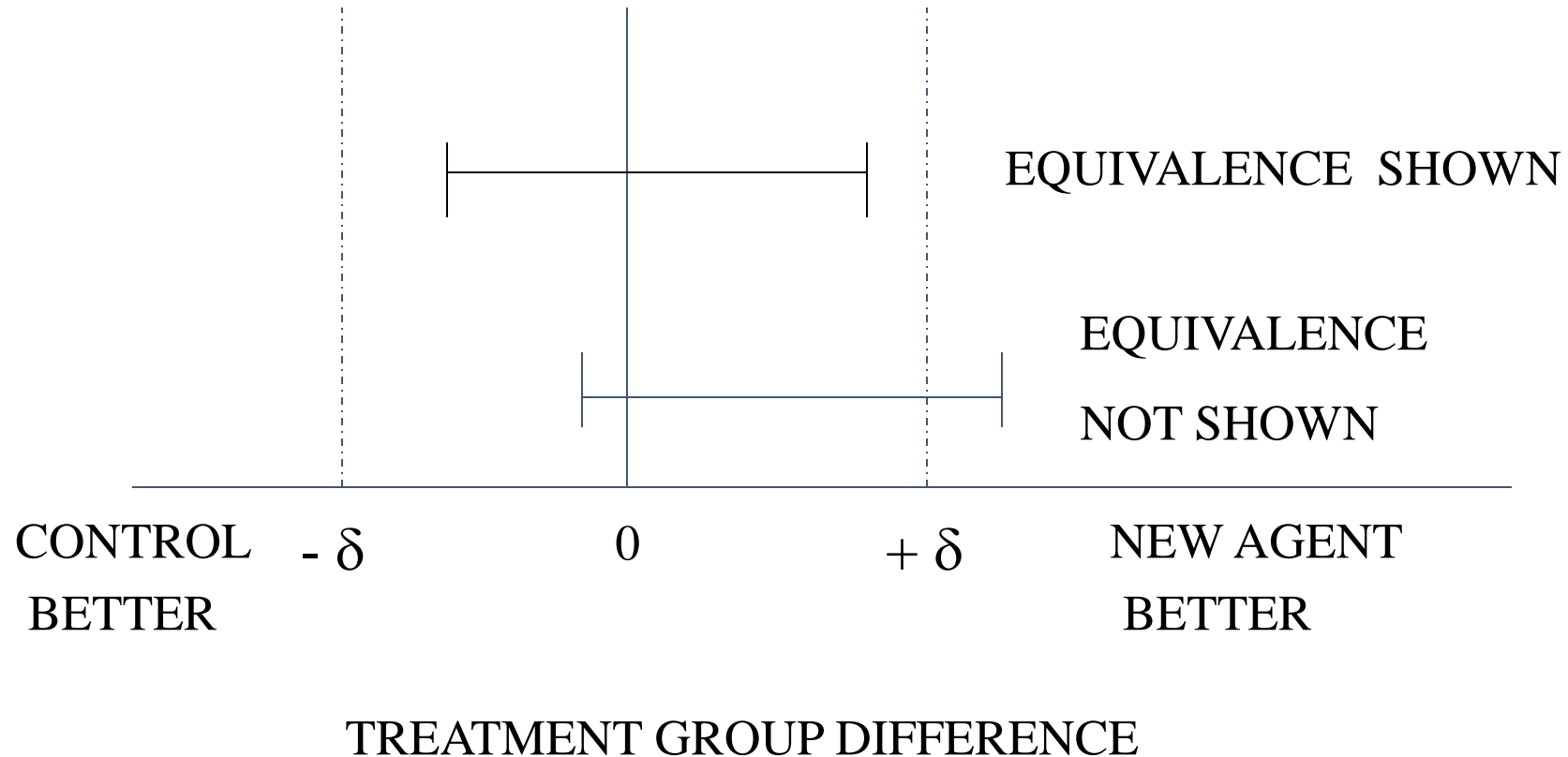
Hypothesis of Equivalence

- Example: A trial with the primary objective of showing that the response to two or more treatments differs by an amount which is clinically unimportant.
- This is usually demonstrated by showing that the true treatment difference is likely to lie between a lower and an upper equivalence margin (δ) of clinically acceptable differences.
- Equivalence trials aim to show the new treatment is no better and no worse.
 - Non-inferiority trials aim to show that the new drug is no worse than standard treatment.

Equivalence Trial

- Statistical analysis is generally based on the use of confidence intervals (CI).
- For equivalence trials, two-sided confidence intervals should be used.
- Equivalence is inferred when the entire confidence interval falls within the equivalence margins ($-\delta$ to $+\delta$).

Equivalence Trial



Confidence interval approach to analysis of equivalence trial

Summary

- Choose the correct alternative hypothesis test for your research question
- Superiority: Treatment A better than treatment B?
- Non-Inferiority: Treatment A no worse than treatment B
- Equivalence: Treatment A no better and no worse than treatment B
- Impacts your sample size calculations!

Required Components For Basic Power And Sample Size Calculations

Why do sample sizes matter?

- Needs to be adequate to detect clinically meaningful effect - but also consider:
 - Does the sample size reflect proper use of resources
 - Does the sample size hold up ethically and not expose more people than necessary to potential harm (never use too big a sample size!)
- Need to justify the sample size early in planning

Sample sizes depend on...

- Study Design:
 - Observational study - need larger sample size
 - Experimental study - depends on number of groups, randomization
- Type of outcome
 - Continuous outcome (weight, blood pressure)
 - Dichotomous outcome (presence and absence)
- Study population (diverse or homogeneous)
- Budget

Type I and Type II Errors

- Errors associated with hypothesis testing:

Truth

Concluded *Association* *No Association*

Association

Correct!

True positive

Power ($1 - \beta$)

False positive

Type I error

Alpha (α)

No Association

False negative

Type II error

Beta (β)

Correct!

True negative

$1 - \alpha$

Example: Type I and Type II Errors

Truth




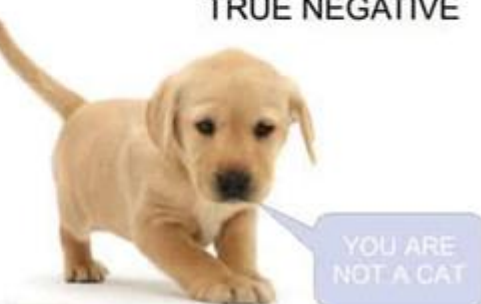
Concluded

Cat

Not a Cat

Cat

Not a Cat

<p>TRUE POSITIVE</p>  <p>YOU ARE A CAT</p>	<p>FALSE POSITIVE</p>  <p>YOU ARE A CAT</p> <p>TYPE I ERROR</p>
<p>FALSE NEGATIVE</p>  <p>YOU ARE A DOG</p> <p>TYPE II ERROR</p>	<p>TRUE NEGATIVE</p>  <p>YOU ARE NOT A CAT</p>

Significance Level

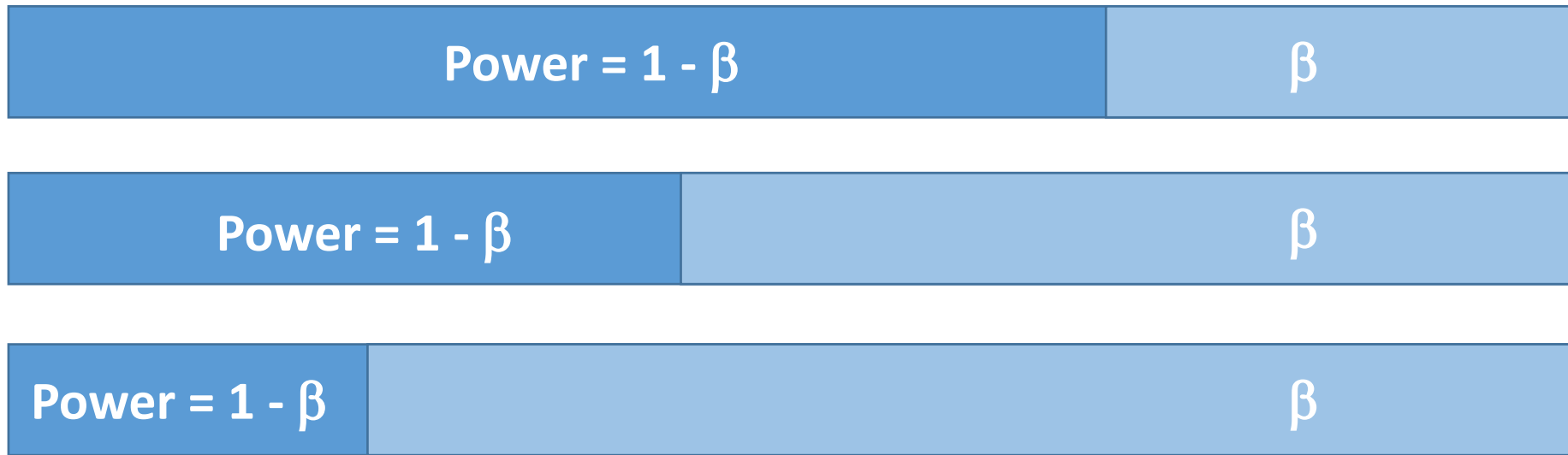
- Significance level: α
 - Probability of a Type I error
 - Probability of a false positive
 - Example: If the effect on DBP of the treatments do not differ (in truth), what is the probability of incorrectly concluding that there is a difference between the treatments?
 - Typically chosen to be 5%, or 0.05

Statistical Power

- Power: $1 - \beta$
 - Probability of detecting a true treatment effect
 - Power = (1 - probability of a false negative)
 - = (1 - probability of Type II error)
 - = $(1 - \beta)$ = probability of a true positive
 - Example: If the effects of the treatments do differ, what is the probability of detecting such a difference?
 - Typically chosen to be 80-99%

Power

- Power = $1 - \beta$



Sample Size Calculations for Continuous Outcomes

Treatment Effect

- What is the minimal, clinically significant difference in treatments we would like to detect?
- Pilot studies may indicate magnitude
- Example: The authors felt that a 3 mm Hg difference in DBP between the treatment groups was clinically significant
- Denoted by δ

Variability in Response

- To estimate sample size, we need an estimate of the variability of the response in the population
- Estimate variability from pilot or previous, related study
- Example: The authors estimate that the standard deviation of DBP is 7 mm Hg.
- Denoted by σ^2

Sample Size Calculation

- Continuous response from 2 independent samples, 2-sided alternative

$$N = \frac{4(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2}{\delta^2}$$

N = total sample size

$Z_{1-\alpha/2}$ = critical value corresponding to 2-sided, Type I error rate

$Z_{1-\beta}$ = critical value corresponding to Type II error rate

δ = effect size

σ = response standard deviation

Sample Size Calculation Example

- Estimate the sample size required to detect a 3 mm Hg difference in blood pressure level with 90% power at the two-sided, 0.05 significance level where the standard deviation in has been estimated to be 7 mm Hg.

$$\text{Total Sample Size} = \frac{4(1.96 + 1.282)^2 7^2}{3^2} = 229$$

So, 229 patients would be required, or 115 per group patients in each group

Standardized Effect Size

- When considering continuous data, the sample size calculations depend on the standardized effect size, defined as:

$$SES = \frac{\delta}{\sigma}$$

where δ is the effect size (difference in means) and σ is the standard deviation.

Standardized Effect Size

- When estimates of the standard deviation are not available, there are general guidelines which quantify the standardized effect size (Cohen, J. A Power Primer. *Psychological Bulletin*. 1992. 112(1)155-159)
- Small Effect: $SES = 0.2$
 - “Noticeably smaller than medium but not so small as to be trivial”
- Medium Effect: $SES = 0.5$
 - “An effect likely to be visible to the naked eye of a careful observer”
 - “The average size of observed effects in various fields”
- Large Effect: $SES = 0.80$
 - “Same distance above medium as small is below it”

- Guidelines for standard effect sizes are subjective
- It is best to find estimates of the standard deviation through pilot studies or related studies of the same response

Sample Size Calculation

- Sample size calculation for a continuous response from 2 independent samples, 2-sided alternative in terms of standardized effect size

$$N = \frac{4(Z_{1-\alpha/2} + Z_{1-\beta})^2}{SES^2}$$

$$N = \frac{4(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2}{\delta^2}$$

N = total sample size

$Z_{1-\alpha/2}$ = critical value corresponding to Type I error

$Z_{1-\beta}$ = critical value corresponding to Type II error

SES = Standardized Effect Size

What Do I Need?

Tests for One Mean	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
σ	An estimate of the standard deviation (must be positive).
δ_0	A value for the population mean under the NULL hypothesis. This is the baseline mean.
δ_1	A value (or range of values) for the population mean under the ALTERNATIVE hypothesis. In the power calculation, the important quantity is not the value of the mean itself, but the size of the difference between the two means ($\delta_0 - \delta_1$).

What Do I Need?

Test for Paired Means	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
σ	Standard Deviation of Paired Differences: An estimate of the standard deviation of paired differences (must be positive).
δ	Mean of Paired Differences (Alternative): the value(s) for the mean of paired differences under the ALTERNATIVE hypothesis. This value indicates the minimum detectable difference for the corresponding power and sample size.

What Do I Need?

Two-Sample T-Tests (assuming equal variance)	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
σ	The standard deviation entered here is the assumed standard deviation for both the Group 1 population and the Group 2 population.
μ_1	A value for the assumed mean of Group 1.
μ_2	A value for the assumed mean of Group 2.

What Do I Need?

Two-Sample T-Tests (assuming unequal variance)	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
σ_1	The standard deviation entered here is the assumed standard deviation for the Group 1 population.
σ_2	The standard deviation entered here is the assumed standard deviation for the Group 2 population.
μ_1	A value for the assumed mean of Group 1.
μ_2	A value for the assumed mean of Group 2.

What Do I Need?

One-Way Analysis of Variance (ANOVA)	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
σ	This is the standard deviation within a group. It is assumed to be the same for all groups. As a standard deviation, the number(s) must be greater than zero.
k	This is the number of groups being compared. Thus, it is the number of means. It must be greater than or equal to two.
$\mu_{1...k}$	A set of hypothesized means, one for each group. These means represent the group centers under the alternative hypothesis (the null hypothesis is that they are equal).

Sample Size Calculations for Dichotomous Outcomes

Sample Size for a Dichotomous Response

- The total number of subjects needed is given by

$$N = \frac{4(Z_{1-\alpha/2} + Z_{1-\beta})^2 [\bar{p}(1 - \bar{p})]}{(p_C - p_I)^2}$$

using the estimated probability of response in the control (p_C) and intervention (p_I) groups as well as the average response

$$\bar{p} = \frac{(p_I + p_C)}{2}$$

Example

- A randomized study was designed to investigate the efficacy of laser therapy for regeneration of nerves after an oral injury.
- It was hypothesized that **75% of the subjects** in the treated group would experience increased sensation in their mouths (based on a particular scale) while only **15% of the control group subjects** would experience an improvement.
- Calculate the number of subjects required for the randomized study if we would like to have **80% power** to detect the hypothesized difference with a 2-sided **5% Type I error**.

$$\begin{aligned}\text{Total Sample Size} &= \frac{4(1.96 + 0.84)^2 [.45 * .55]}{(.75 - .15)^2} \\ &= 22\end{aligned}$$

What Do I Need?

Tests for One Proportion using Proportions	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
p_0	This is the value of the proportion under the null (H0) hypothesis. The proportion estimated from the data will be compared to this value by the statistical test.
p_1	This is the value of the proportion (P1) under the alternative hypothesis. The power calculations assume that this is the actual value of the proportion.

What Do I Need?

Tests for Two Proportion using Proportions	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
p_1	This is the value of P1 under the alternative hypothesis, H1. The power calculations assume that this is the actual value of the proportion.
p_2	A value for the proportion in the control (baseline, standard, or reference) group, P2.

What Do I Need?

Tests for Two Correlated Proportions (McNemar)	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
$P_{10} - P_{01}$	Difference ($P_{10}-P_{01}$). In a matched-pairs study, the hypothesis being tested is whether $P_t=P_s$ which is equivalent to $P_{10}=P_{01}$. This parameter sets the value of this difference.
$P_{10} + P_{01}$	Proportion Discordant ($P_{10}+P_{01}$). In a matched-pairs study, this is the proportion of pairs for which the response differed. It is referred to as the proportion of discordant pairs.

		STANDARD			
		Yes	No	Total	
TREATMENT	Yes	P_{11}	P_{10}	P_t	$D = P_t - P_s = (P_{11}+P_{10}) - (P_{11}+P_{01}) = P_{10} - P_{01}$
	No	P_{01}	P_{00}	$1-P_t$	
Total		P_s	$1-P_s$	1	

What Do I Need?

Chi-Square Tests	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
df	Specify the number of degrees of freedom in the Chi-square test. For a contingency table, $DF = (Rows-1)(Columns-1)$. For example, for a 3-by-4 table, $DF = (3-1)(4-1) = 6$.
W	W (Effect Size). Specify one or more values for W , the effect size of the Chi-square test. It must be greater than zero. In a contingency table, $W = \text{SQRT}[(\text{ChiSquare}) / N]$. Cohen suggests using $W = 0.1$ as a small value and $W = 0.5$ as a large value.

What Do I Need?

Kappa Test for Agreement Between Two Raters	
α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
p	Two or more marginal classification frequencies (proportions). These are the relative proportions of subjects in each category as assigned by the two raters or judges.
κ_0	This is the value of Kappa under the null hypothesis. Values must be between -1 and 1.
κ_1	This is the value of Kappa under the alternative hypothesis. Values must be between -1 and 1.

What Do I Need?

Tests for One-Sample Sensitivity and Specificity

α	Alpha is the probability of obtaining a false positive with the statistical test. That is, it is the probability of rejecting a true null hypothesis.
$1 - \beta$	Power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 - Beta, so specifying power implicitly specifies beta.
p	P (Prevalence): the anticipated proportion of the population of interest that has the disease. Because this is a proportion all values must be between zero and one
Sn_0	This is the value of the sensitivity under the null (H0) hypothesis.
Sn_1	This is the value of the sensitivity under the alternative (H1) hypothesis.
Sp_0	This is the value of the specificity under the null (H0) hypothesis.
Sp_1	This is the value of the specificity under the alternative (H1) hypothesis.

Relationships Between Type I Error, Power, Sample Size, And Effect Size

Factors Influencing Power

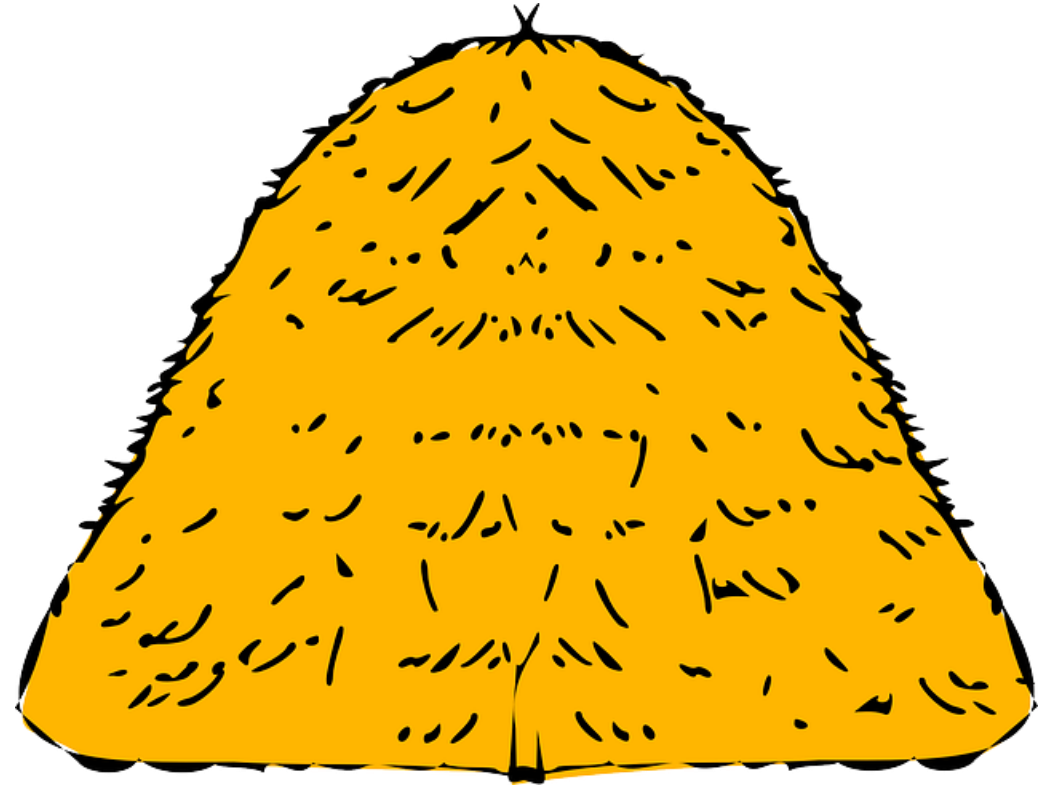
Assuming all other factors fixed, power decreases when the following changes occur:

- \downarrow significance level \Rightarrow \downarrow power
- \downarrow effect size \Rightarrow \downarrow power
- \uparrow variability in response \Rightarrow \downarrow power
- \downarrow sample size \Rightarrow \downarrow power

Effect Size and Power

- Assuming all other factors remain constant
- \downarrow effect size \Rightarrow \downarrow power

Effect Size



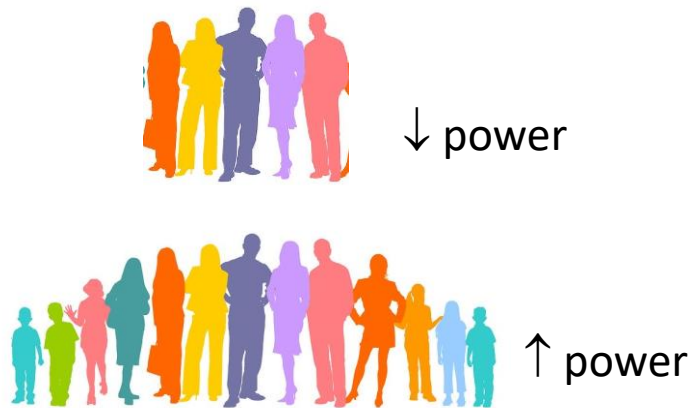
Factors Influencing Sample Size

Assuming all other factors fixed, required sample size increases when the following changes occur:

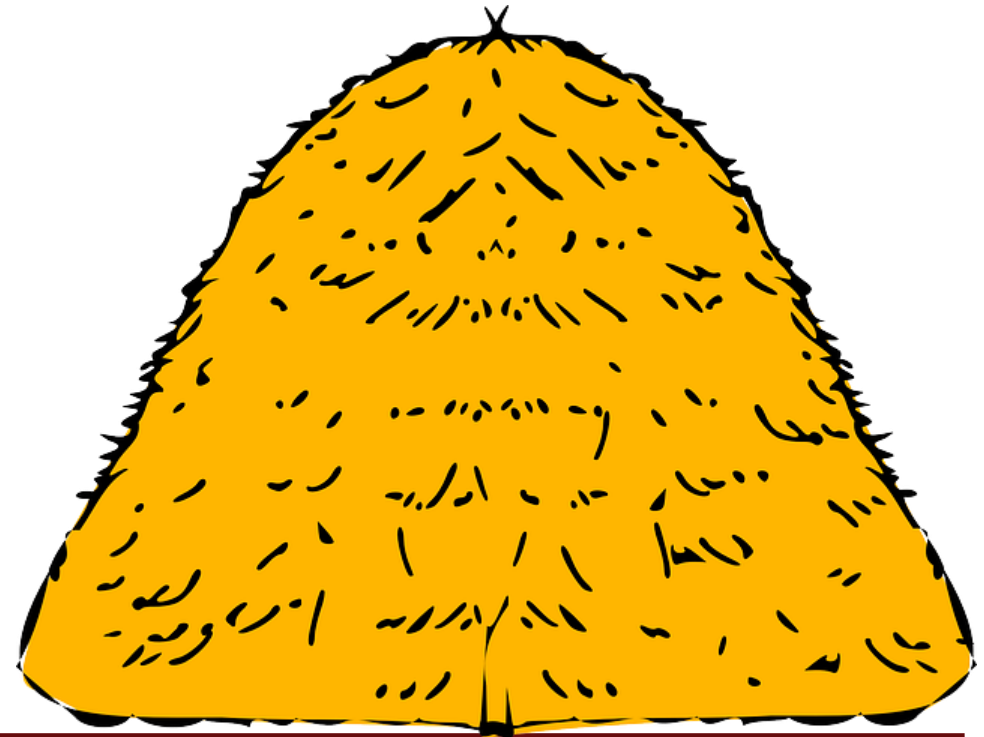
- \uparrow power \Rightarrow \uparrow sample size
- \downarrow significance level \Rightarrow \uparrow sample size
- \uparrow variability in response \Rightarrow \uparrow sample size
- \downarrow effect size \Rightarrow \uparrow sample size

Sample Size and Power

- One way to increase the power of your study is to increase the sample size
- \uparrow sample size \Rightarrow \uparrow power \Rightarrow \downarrow beta (false negative probability)
- \uparrow power \Rightarrow \uparrow sample size



Effect Size



Other Sample Size Considerations

Other Sample Size Considerations

- Unequal group size adjustment
- Adjustment for drop-out
- Multiple comparisons
- Repeated measures / correlated data designs

Unequal Group Size Adjustment

- An adjustment to the total number of required subjects (N) can be made as follows:

$$N^* = \frac{N(1+k)^2}{4k}$$

Where k equals the ratio of the subgroup sizes

Adjustment for Drop-out

- Dropouts will decrease the effective sample size
 - Adjust the sample size for dropouts
 - Example: Simple adjustment to calculated sample size for dropout rate is
Adjusted total size =
Calculated size / (1-dropout rate)

Multiple Comparisons

- Multiple comparisons result in an increased probability of detecting a difference in treatments by chance alone, i.e. the false positive rate, or probability of a Type I error
- Bonferroni adjustment:
 - Adjust the significance level, α , by dividing it by the number of multiple comparisons

Repeated Measures

- Simple case of baseline and 1 follow-up measure
 - For continuous measures, sample size calculations will be based on comparisons of mean change between groups
 - Mean change for each group
 - Standard deviation of the change values
 - For dichotomous measures, sample size calculations will be based on proportion of patients who “improve”, for example

Correlated Data Designs

- Longitudinal studies
 - Multiple time points for each individual
- Family studies
 - Heritable traits from parent to child
- Cluster-randomized studies
 - Groups clustered together more similar than other groups
- Let your statistician know!

Foundations of Basic Sample Size Calculations

Lance Ford, PhD (Lance-Ford@ouhsc.edu)
Assistant Professor of Research, Biostatistics

Sara Vesely, PhD (Sara-Vesely@ouhsc.edu)
Associate Dean of Academic Affairs
David Ross Boyd Professor, Biostatistics

Kai Ding, PhD (Kai-Ding@ouhsc.edu)
Associate Professor, Biostatistics

Chao Xu, PhD (Chao-Xu@ouhsc.edu)
Assistant Professor, Biostatistics

