OHSU

Power and Sample Size 101

Biostatistics, Epidemiology, & Research Design (BERD) Seminar

OCTRI Research Forum

DATE: Feb 4, 2021
PRESENTED BY: Meike Niederhausen, PhD

OREGON CLINICAL & TRANSLATIONAL RESEARCH INSTITUTE

Goals

- Understand power, sample size, effect size, and alpha level, and how they are related
- Understand the information required for power and sample size (PSS) calculations
- Know what grant reviewers are looking for in a PSS section
- Perform simple PSS calculations
  - Paired t-test
  - Independent samples t-test
  - Chi-square test

Prerequisites

An introductory statistics class

- Know what a hypothesis test is
  - Null and alternative hypotheses
  - Test statistic
  - Significance level
- Test types
  - T-test (one and two-sample)
  - Chi-squared test

Hypothetical Intervention Study

Hypothetical Intervention Study

- Suppose you've developed a new spinal surgery intervention
- Will this new intervention decrease (at 1 month post operation)
  - disability scores?
  - depression?

Disability Index: ODI

- ODI is a validated questionnaire
  - assesses disability due to low back pain
  - common measure for spinal disorders
- 10 questions
  - domains such as walking, personal care, and sex life
- Questions rated 0–5
  - higher numbers ➤ greater disability
  - maximum score = 50
- Final score multiplied by 2
- Disability index from
  - **0 to 100**

PHQ-9

- Patient Health Questionnaire
- Used to measure depression

- 9 questions
- Total score calculated
  - Range: 0-27
- Clinical cutoff 15
  - Depression if PHQ-9 $\geqslant$ 15

PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems? *(Use "✓" to indicate your answer)*	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

FOR OFFICE CODING 0 + _____ + _____ + _____

=Total Score: _____

OHSU OREGON CLINICAL & TRANSLATIONAL *Research Institute*

Hypothetical study data

- Sample size $n = 10$
- Everyone received intervention
- Collected ODI at baseline (pre) and 1 month after surgery (post)
- Average within-person change (pre – post) in disability scores (ODI) is
  - $mean = 12$ ($SD = 25$)

- Next step:
  - Is this change statistically significant?
  - What's the p-value?

Hypothesis Test

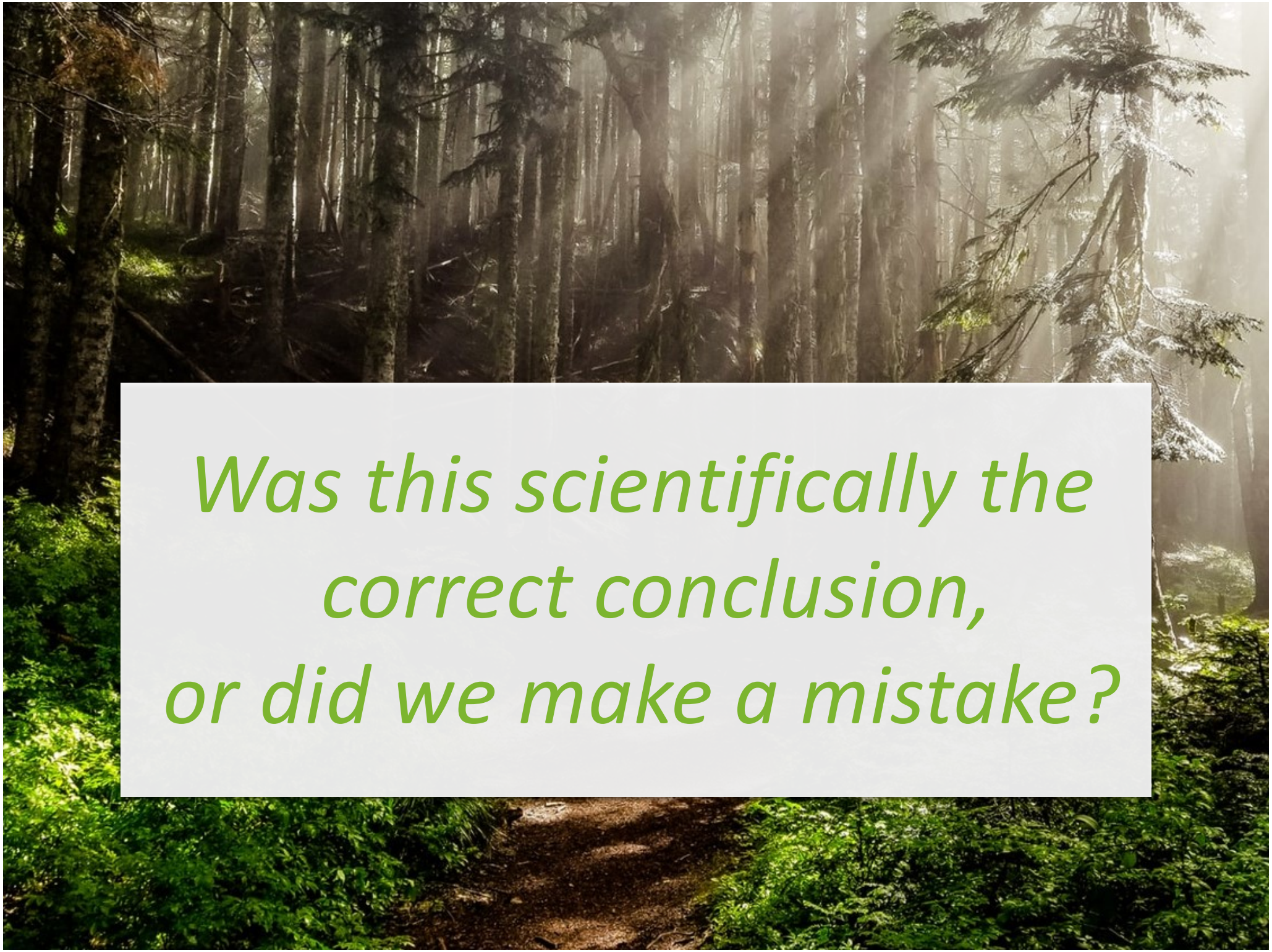
- Paired *t*-test (a 1-sample t-test)
- Null and alternative hypotheses

$$H_0 : \mu_D = 0$$
$$\text{vs. } H_A : \mu_D \neq 0$$

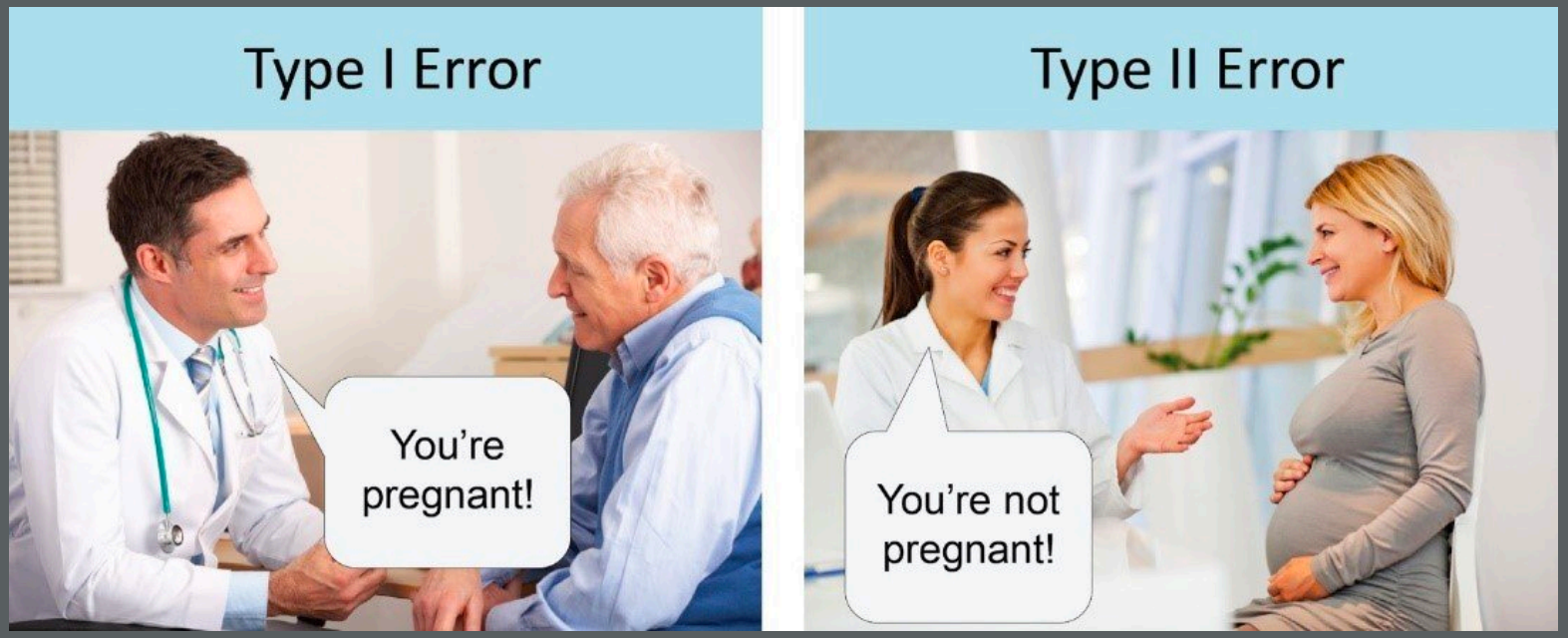
- Test statistic

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{12 - 0}{\frac{25}{\sqrt{10}}} = 1.52$$

- p-value = 0.16 ➤ fail to reject H_0
- Insufficient evidence to claim the intervention decreased ODI. 🤬


Was this scientifically the correct conclusion, or did we make a mistake?

Type I & II Errors



https://www.statisticssolutions.com/to-err-is-human-what-are-type-i-and-ii-errors/

Justice System Analogy

Justice System - Trial

	Defendant Innocent	Defendant Guilty
Reject Presumption of Innocence (Guilty Verdict)	**Type I Error**	**Correct**
Fail to Reject Presumption of Innocence (Not Guilty Verdict)	**Correct**	**Type II Error**

Statistics - Hypothesis Test

	Null Hypothesis True	Null Hypothesis False
Reject Null Hypothesis	**Type I Error**	**Correct**
Fail to Reject Null Hypothesis	**Correct**	**Type II Error**

http://www.intuitor.com/statistics/T1T2Errors.html

OHSU OREGON CLINICAL & TRANSLATIONAL *Research Institute*

Justice System Analogy

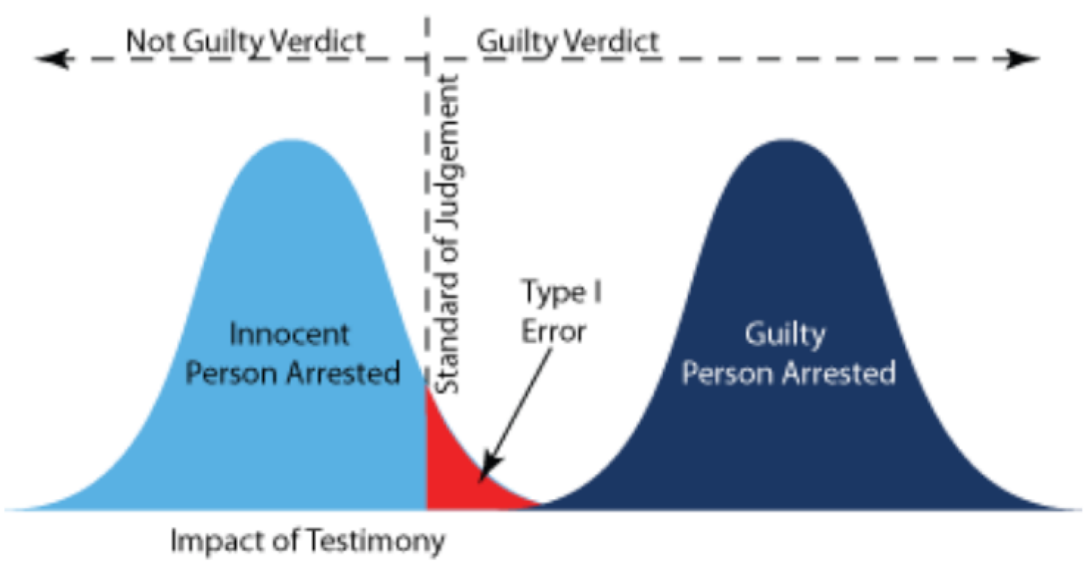


figure 3. Distribution of possible witnesses in a trial showing the probable outcomes with a single witness if the accused is innocent or obviously guilty..

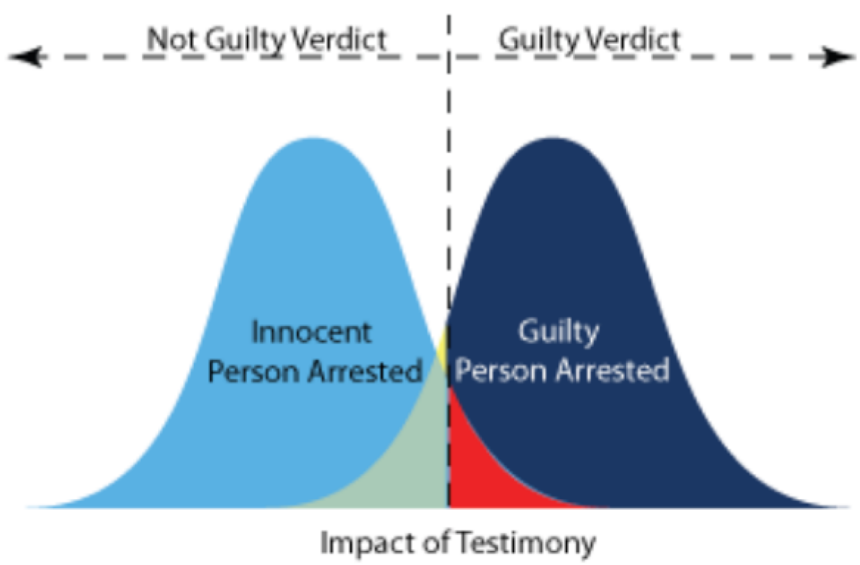


figure 4. Distribution of possible witnesses in a trial showing the probable outcomes with a single witness if the accused is innocent or not clearly guilty..

http://www.intuitor.com/statistics/T1T2Errors.html

OHSU OREGON CLINICAL & TRANSLATIONAL *Research Institute*

Type I & II errors

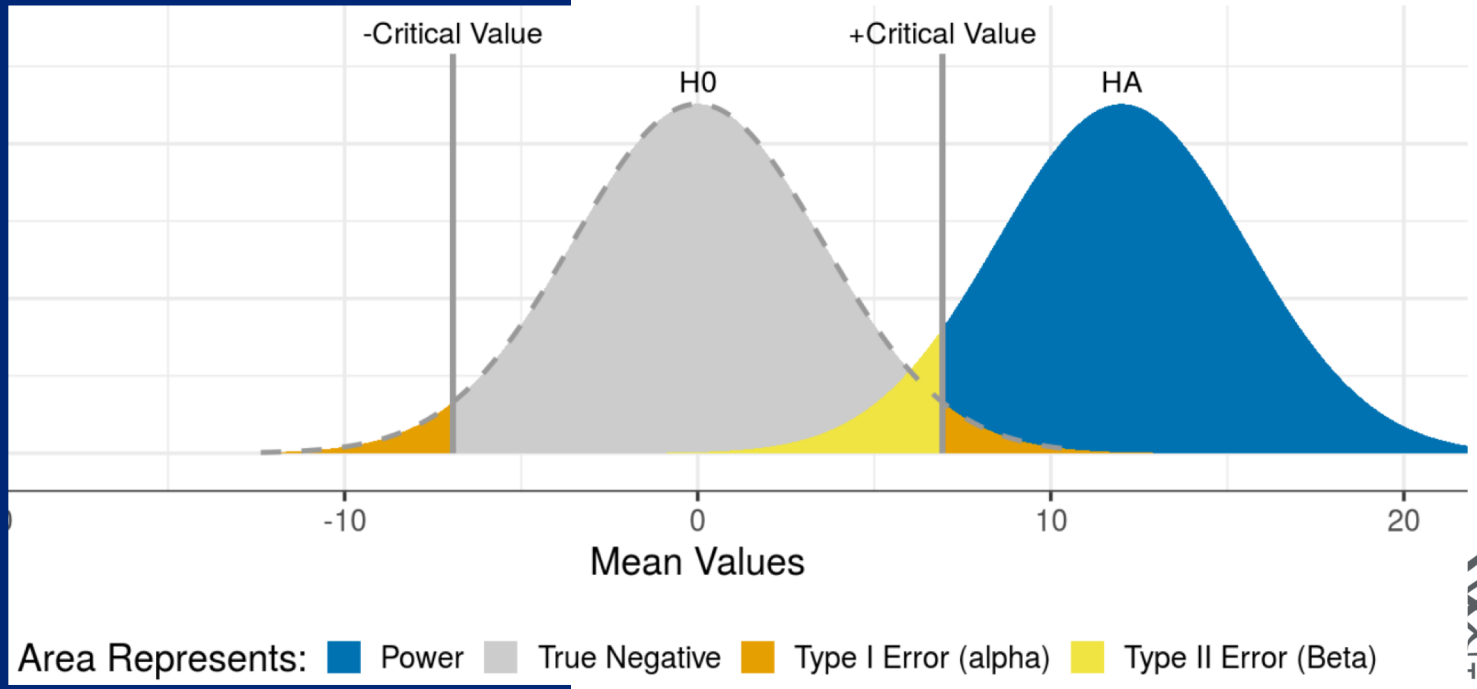
	Fail to reject null hypothesis	Reject null hypothesis
Null hypothesis is true	Correct! (true negative)	Type I error (false positive) probability = α
Null hypothesis is false	Type II error (false negative) probability = β	Correct! (true positive)

Ideally, we want

- small Type I & II errors and
- big power

Power =

P(correctly rejecting the null hypothesis)



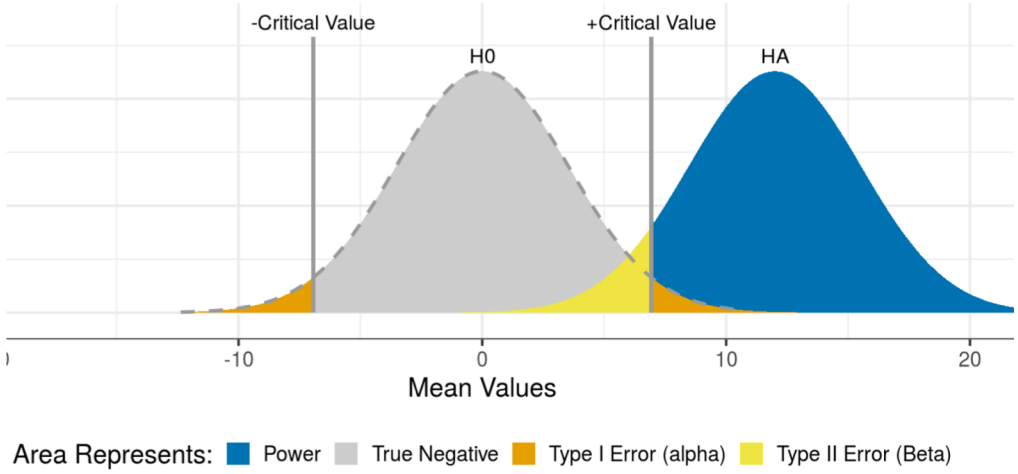
Power & Sample Size Applet

OREGON CLINICAL & TRANSLATIONAL Research Institute

Relationships between Type I & II errors and power

Type I & II errors

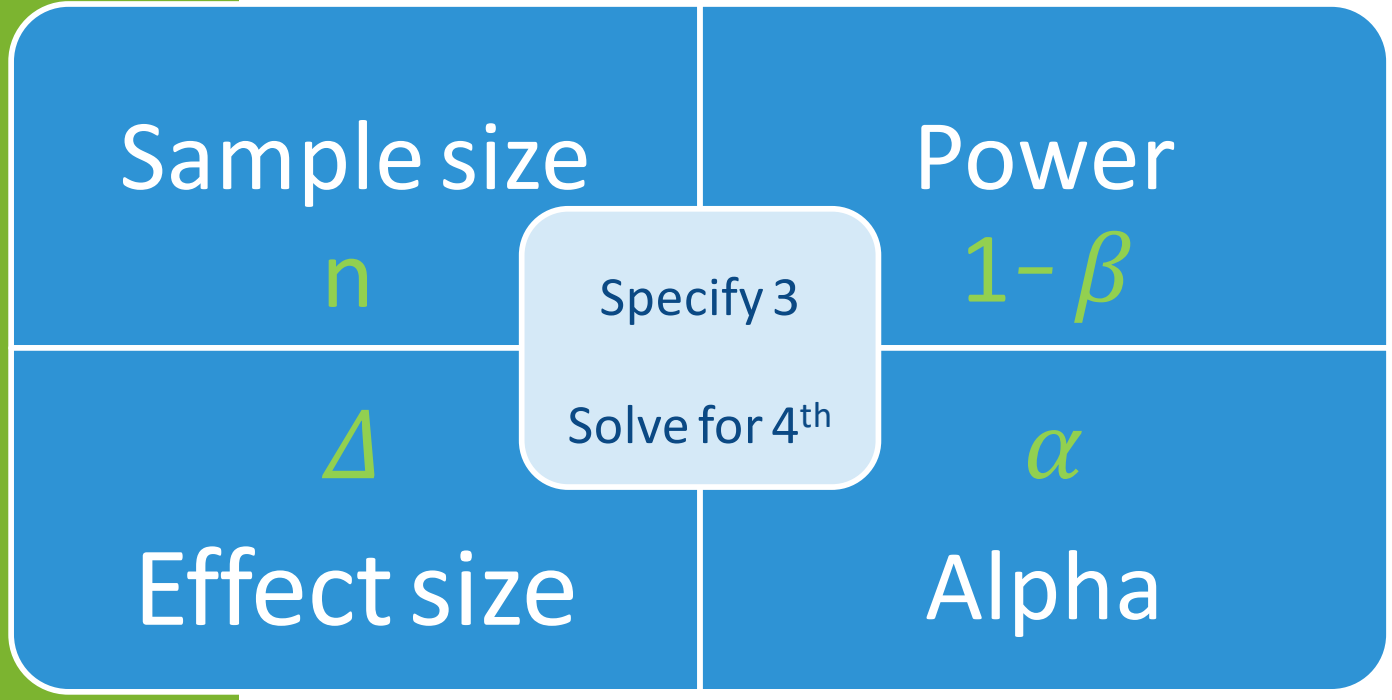
- ⇩ P(Type I error) leads to ⇧ P(Type II error)
  - We typically set α = P(Type I error) as 0.05



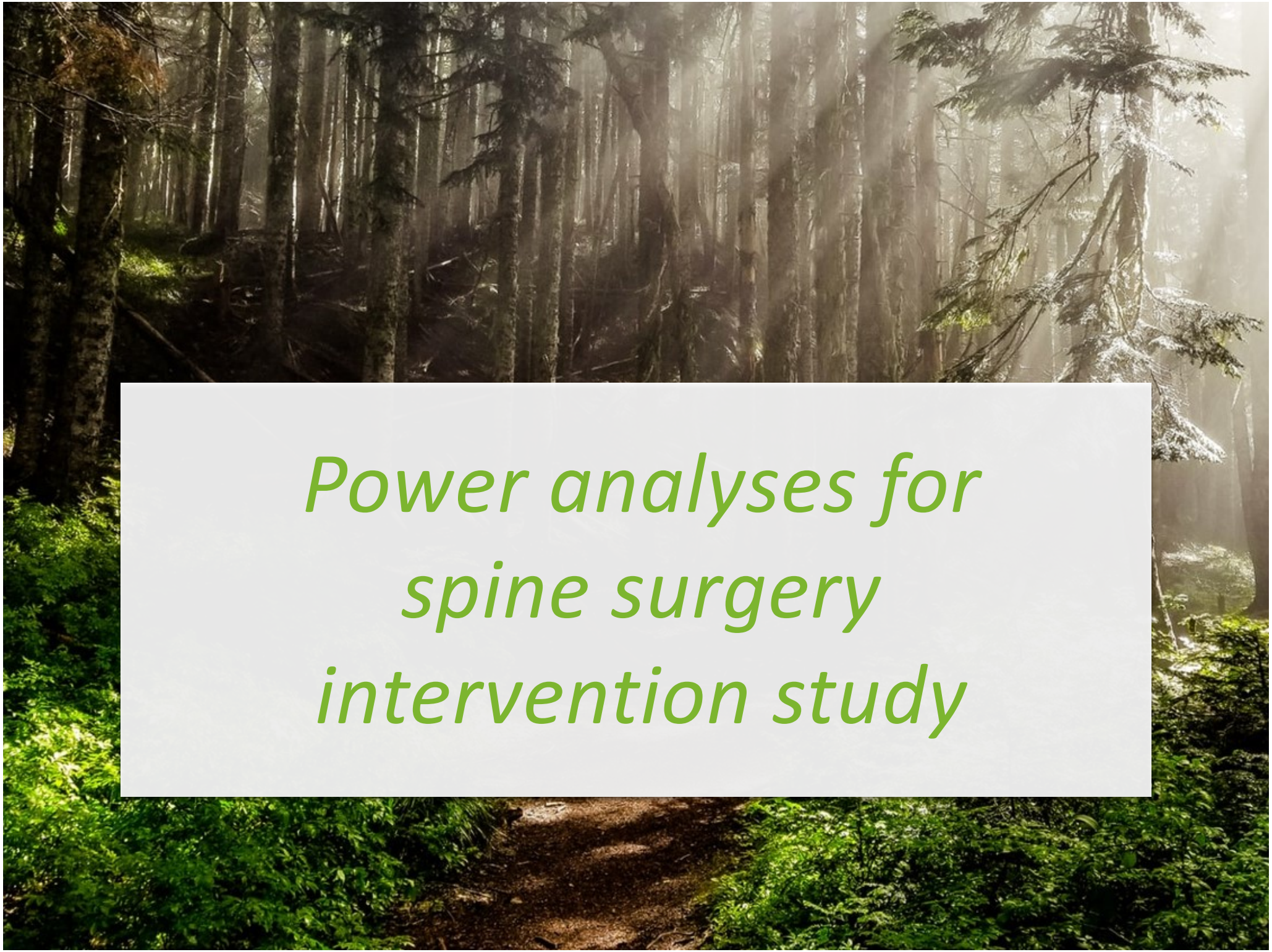
Power vs. Type II error

- Power = 1 - P(Type II error) = 1 - β
- As P(Type II error) ⇩, the power ⇧
- P(Type II error) ⇩ as the effect size ⇧

4 components to a power analysis



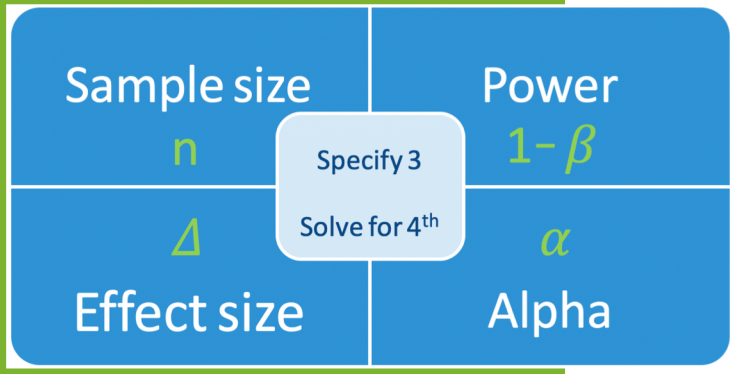
OHSU
OREGON CLINICAL & TRANSLATIONAL
Research Institute


Power analyses for
spine surgery
intervention study

Outcomes & Hypotheses

- Randomize patients to control and intervention groups
- Primary Outcome: **Disability**
  - Oswestry Disability Index (ODI)
  - Hypotheses:
    1. ODI will decrease after surgery (1 mo)
    2. Lower ODI in intervention group 4 mo after surgery
- Secondary Outcome: **Depression**
  - PHQ-9
  - Hypothesis:
    1. Fewer people with depression in intervention group 1 mo after surgery

What sample size do we need?

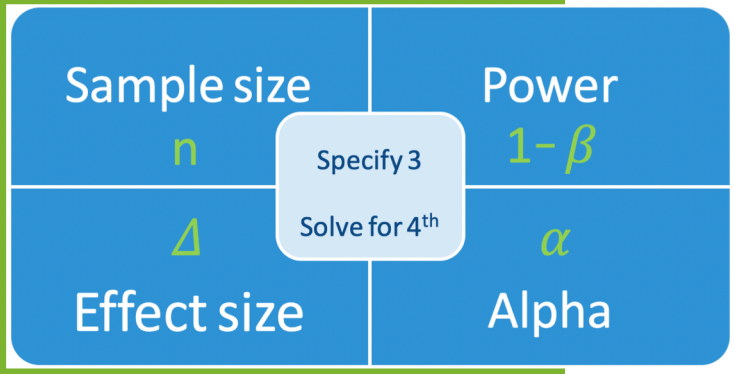


OREGON CLINICAL & TRANSLATIONAL Research Institute
OHSU

Study Design

- Hypothesis:
  - ODI will decrease after surgery (1 mo)
- Compare pre- vs. post-surgery ODI means
- Assume just one group *(to keep it simple for now…)*,
  - i.e. the intervention group
- Statistical analysis:
  - Need a paired t-test

- **What sample size do we need?**
  - *What do we need to know to estimate this?*



OHSU OREGON CLINICAL & TRANSLATIONAL Research Institute

How to estimate the **effect size**?

- Pilot or preliminary data
- Published literature
- Clinically meaningful difference

Published literature

ELSEVIER

CrossMark

THE SPINE JOURNAL

The Spine Journal 18 (2018) 1398–1405

Clinical Study

The 9-Item Patient Health Questionnaire (PHQ-9): an aid to assessment of patient-reported functional outcomes after spinal surgery

Andrew N. Tuck, BS[a], Melissa B. Scribani, MPH[b], Scott D. Grainger, RN, BS[c], Celeste A. Johns, MD[d], Reginald Q. Knight, MD, MHA[c,e,*]

Abstract

BACKGROUND CONTEXT: Preoperative depression is increasingly understood as an important predictor of patient outcomes after spinal surgery. In this study, we examine the relationship between depression and patient-reported functional outcomes (PRFOs), including disability and pain, at various time points postoperatively.

PURPOSE: The objective of this study was to analyze the use of depression, as measured by the 9-Item Patient Health Questionnaire (PHQ-9), as a means of assessing postoperative patient-reported disability and pain.

Link to paper

OHSU OREGON CLINICAL & TRANSLATIONAL *Research Institute*

Published literature

Table 2
Patient-reported functional outcomes—total cohort

	PHQ-9 score	VAS score for arm or leg	VAS score for neck or back	ODI score
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Preoperative	8.0 (6.4)	59.5 (30.5)	63.5 (27.4)	46.5 (18.4)
1 mo	5.0 (5.5)	30.2 (30.8)	38.9 (28.6)	33.9 (21.9)
4 mo	4.7 (5.8)	30.4 (31.7)	39.4 (30.7)	29.9 (22.6)
10 mo	4.8 (5.6)	36.7 (32.8)	44.3 (31.0)	32.4 (22.3)
24 mo	5.3 (6.4)	39.8 (33.2)	47.7 (30.7)	34.6 (23.4)
p-Value	<.0001	<.0001	<.0001	<.0001

PHQ-9, 9-Item Patient Health Questionnaire; VAS, visual analog scale; ODI, Oswestry Disability Index; SD, standard deviation.

OHSU OREGON CLINICAL & TRANSLATIONAL *Research Institute*

Effect size for change in mean

- For the effect size, we need **both** an estimate for
  - the **change in the mean**, and
  - The **standard deviation (SD) of the *differences***
- Problem: the published paper doesn't include the SD of the differences. *Typical!*
  - Need an estimate of the SD

$$SD_{diff} = \sqrt{SD_{pre}^2 + SD_{post}^2 - (2 \cdot Corr \cdot SD_{pre} \cdot SD_{post})}$$

  - SD_{pre} , SD_{post} = SD's of pre and post values
  - Corr = is the correlation between the pre and post intervention values
    - *How do we estimate the correlation???*

Estimating the SD of the differences

- Problem:
  - don't know SD of the differences
  - also don't know the correlation between the pre and post intervention values to help us estimate the SD of the differences…
- Solution:
  - Run the sample size calculation for varying SD values
  - Mean difference = 12.6
  - Power = 0.80, α = 0.05

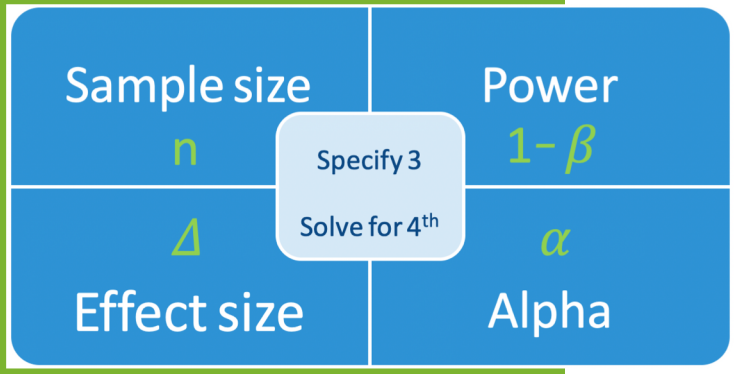
Corr	SD_{diff}	Sample size
0.20	25.6	35
0.50	20.4	23
0.80	13.2	11

https://wbakerrobinson.shinyapps.io/Flexible_One_Sample_T_test_shiny/

Outcomes & Hypotheses

- Randomize patients to control and intervention groups
- Primary Outcome: **Disability**
  - Oswestry Disability Index (ODI)
  - Hypotheses:
    1. ODI will decrease after surgery (1 mo)
    2. Lower ODI in intervention group 4 mo after surgery
- Secondary Outcome: **Depression**
  - PHQ-9
  - Hypothesis:
    1. Fewer people with depression in intervention group 1 mo after surgery

What sample size do we need?

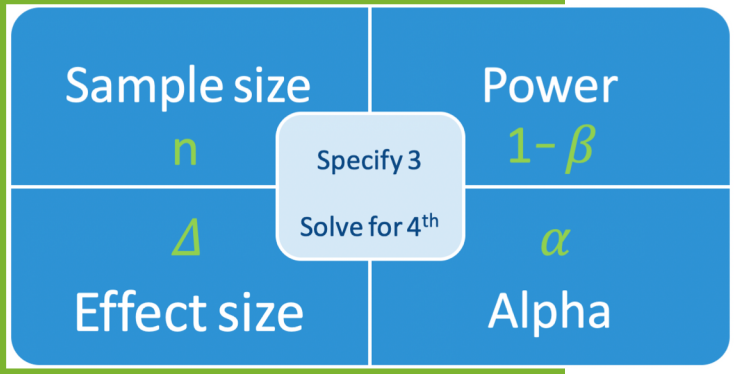


OREGON CLINICAL & TRANSLATIONAL Research Institute
OHSU

Study Design

- Hypothesis:
  - Lower ODI in intervention group 4 mo after surgery
- Now comparing two groups:
  - control & intervention
- Statistical analysis:
  - Need a 2-sample t-test (independent samples)

- **What sample size do we need?**
  - *What do we need to know to estimate this?*



OHSU OREGON CLINICAL & TRANSLATIONAL Research Institute

Published literature

Table 2
Patient-reported functional outcomes—total cohort

	PHQ-9 score	VAS score for arm or leg	VAS score for neck or back	ODI score
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Preoperative	8.0 (6.4)	59.5 (30.5)	63.5 (27.4)	46.5 (18.4)
1 mo	5.0 (5.5)	30.2 (30.8)	38.9 (28.6)	33.9 (21.9)
4 mo	4.7 (5.8)	30.4 (31.7)	39.4 (30.7)	29.9 (22.6)
10 mo	4.8 (5.6)	36.7 (32.8)	44.3 (31.0)	32.4 (22.3)
24 mo	5.3 (6.4)	39.8 (33.2)	47.7 (30.7)	34.6 (23.4)
p-Value	<.0001	<.0001	<.0001	<.0001

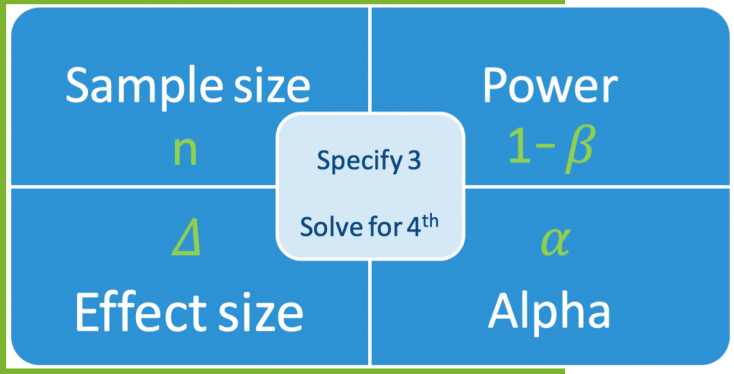
PHQ-9, 9-Item Patient Health Questionnaire; VAS, visual analog scale; ODI, Oswestry Disability Index; SD, standard deviation.

OHSU OREGON CLINICAL & TRANSLATIONAL *Research Institute*

2-sample t-test sample size calculation

- Published literature (4 mo after surgery)
  - Mean ODI = 29.9 (SD = 22.6)
  - We'll use this mean for our control group
  - We'll use the SD for both groups (*assuming the same*)

- **Effect size**
  - In addition to above statistics,
    - need the **difference between the groups**
  - Preliminary data?
  - What is clinically relevant?
  - *Determine sample sizes for varying values*



OREGON CLINICAL & TRANSLATIONAL *Research Institute*
OHSU

2-sample t-test sample size calculation

	80% Power		85% Power	
Difference in means	n per group	Total n	n per group	Total n
5	322	644	368	736
8	127	254	145	290
10	82	164	93	186

- Sample sizes above were calculated for a two-sample t-test, assuming
  - common SD = 22.6
  - power = 0.80, $\alpha = 0.05$
- Note:
  - we don't actually need the mean of the control group since the power calculation is based off of the difference in means

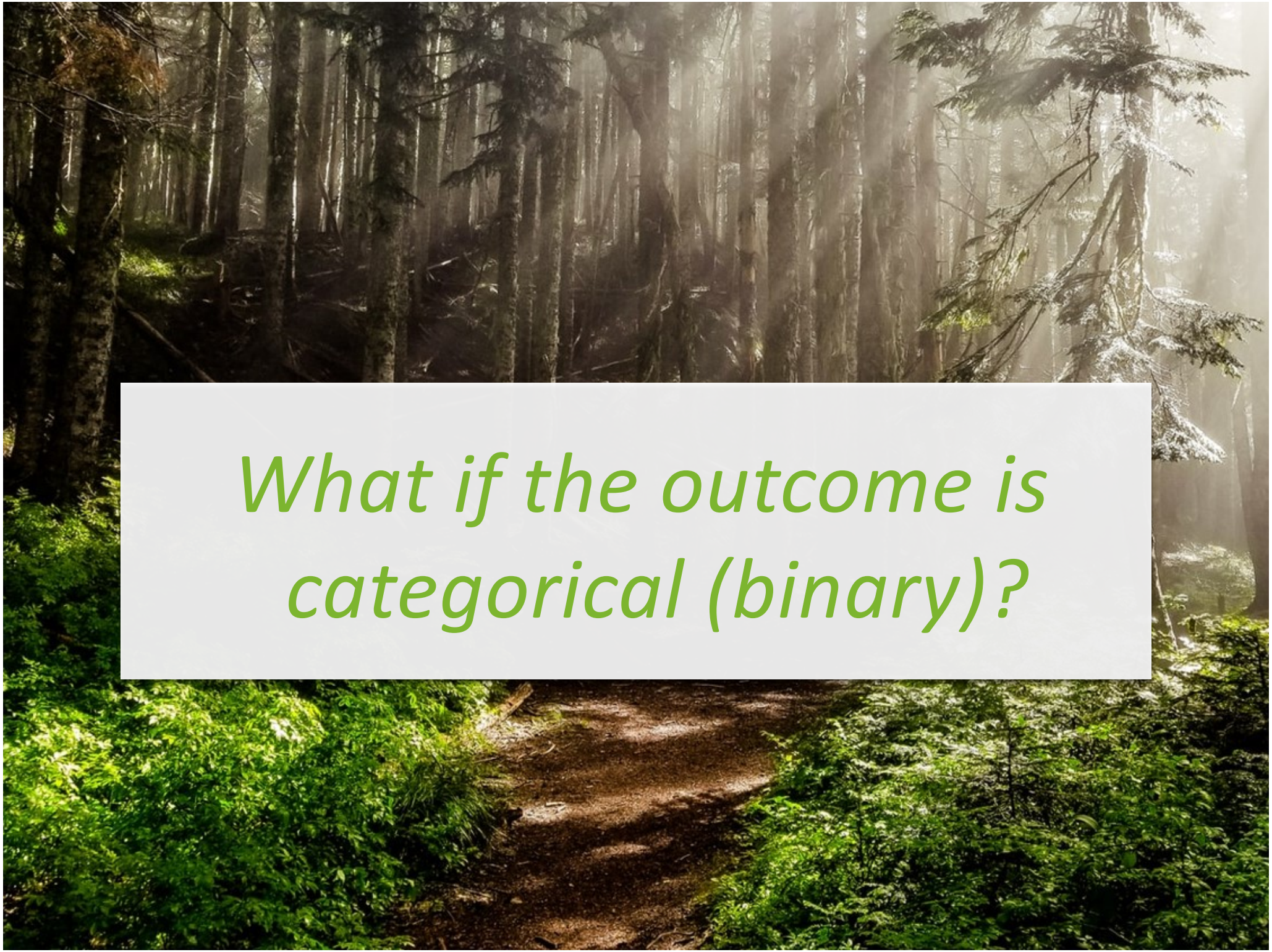
See R code to calculate power at https://berd-pss101-example-rcode.netlify.app/

2-sample t-test sample size calculation in R

```
(two_sample_t_calc1 <- power.t.test(delta = -5,
                                    sd = 22.6,
                                    sig.level = 0.05,
                                    power = 0.80,
                                    type = "two.sample",
                                    alternative = "two.sided"))
```

```
##
##      Two-sample t test power calculation
##
##               n = 321.6747
##           delta = 5
##              sd = 22.6
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

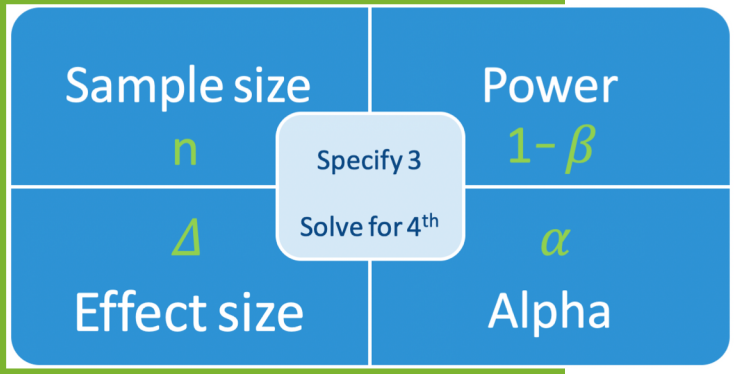
See R code to calculate power at https://berd-pss101-example-rcode.netlify.app/


What if the outcome is categorical (binary)?

Outcomes & Hypotheses

- Randomize patients to control and intervention groups
- Primary Outcome: **Disability**
  - Oswestry Disability Index (ODI)
  - Hypotheses:
    1. ODI will decrease after surgery (1 mo)
    2. Lower ODI in intervention group 4 mo after surgery
- Secondary Outcome: **Depression**
  - PHQ-9
  - Hypothesis:
    1. Fewer people with depression in intervention group 1 mo after surgery

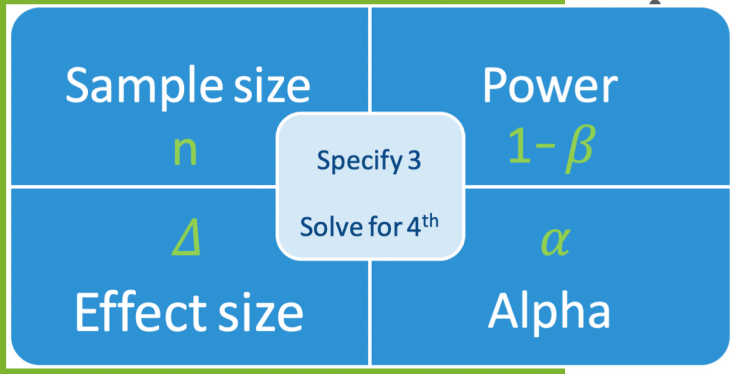
What effect size can we detect?



Analytic plan for depression

- Secondary Outcome: **Depression (PHQ-9)**
  - Hypothesis: Fewer people with depression in intervention group 1 mo after surgery
  - Create binary variable for PHQ-9 (range: 0-27)
    - Depression if PHQ-9 $\geqslant$ 15
- Analysis plan
  - Compare the proportion with depression 1 month after surgery in the control and intervention groups
  - 2-proprtions z-test
- Power analysis
  - Using the sample size from primary outcome,
  - **what difference in proportions can we detect** with 80% power?

What effect size can we detect?

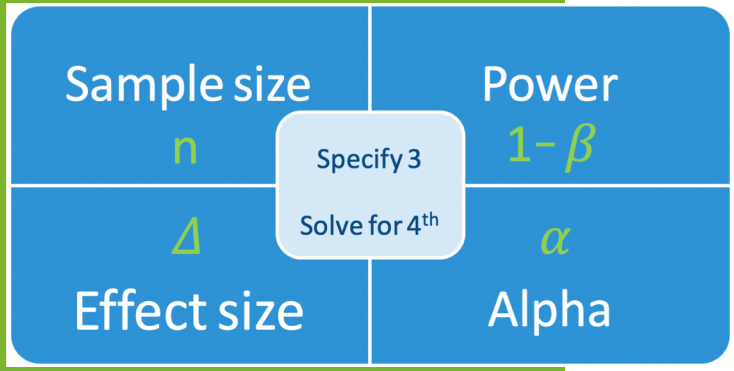


OREGON CLINICAL & TRANSLATIONAL Research Institute
OHSU

Power analysis for 2-prop z-test

- Components
  - Sample size (per group)
  - Power
  - Alpha
  - **Effect size**
    - proportion with depression in control group
    - difference in proportions
      - between control and intervention groups
  - (Continuity correction)
- We're going to use n = 164 (82 per group)
  - from the power table for the
  - primary outcome Hypothesis 2 with
  - difference in means = 10
  - for 80% power.

What effect size can we detect?



OREGON CLINICAL & TRANSLATIONAL Research Institute
OHSU

2-prop z-test sample size calculation

Control group proportion with depression	Decrease in proportions detected
0.10	0.096
0.30	0.177
0.50	0.212

- Are these differences "reasonable"?
  - Achievable?
  - Clinically meaningful?

- Detectable differences in proportions above were calculated for a 2-proportions z-test (Chi-squared), assuming
  - n = 82 per group
  - power = 0.80, $\alpha = 0.05$

See R code to calculate power at https://berd-pss101-example-rcode.netlify.app/

Online calculator: https://stattools.crab.org/R/Two_Arm_Binomial.html

OHSU OREGON CLINICAL & TRANSLATIONAL *Research Institute*


Things to keep in mind…

Things to keep in mind

- Make sure the statistical test used for the power analysis matches that in your analysis plan
  - *We only looked at very basic designs today*
  - *Your research is likely more complex*

- Calculate power or sample size
  - for varying parameter values

Questions?

Meike Niederhausen, PhD

niederha@ohsu.edu

Biostatistics & Design Program (BDP)

bdp@ohsu.edu

Acknowledgements

- Alicia Johnson, MPH
  - Biostatistics and Design Program (BDP)
- Will Baker-Robinson
  - MS Biostatistics student
  - https://github.com/wbakerrobinson
- Yiyi Chen, PhD
  - Associate Professor, Biostatistics
  - OHSU-PSU School of Public Health
  - Knight Cancer Center,
    - Biostatistics Shared Resource (BSR)

OHSU OREGON CLINICAL & TRANSLATIONAL *Research Institute*

What's next?

- *Power and Sample Size for Clinical Trials: An Introduction*
  - Thursday, Feb. 18, 2021, 2:00 - 3:00 pm
  - Presented by Yiyi Chen, PhD

- Abstract:
  - Sample size computation plays an important role in designing a clinical trial. The proposed sample size of a clinical trial depends on lots of factors such as the study design, the hypotheses, the study endpoints, the desired power and the significance level. This seminar will discuss how these factors are related to the computed sample size, illustrated with practical examples.
  - This is the second webinar in a series for power and sample size computation sponsored by BERD.

Thank You