ACKNOWLEDGMENTS

Acknowledgments

The development of the course was funded by National Institute of General Medical Science grant **R25GM111901** (Muller and Glueck co-PI) to the University of Florida, with a subcontract to the University of Colorado Denver.

Other Funding Support

The course covers results published by Muller and his colleagues. A full bibliography appears at www.SampleSizeShop.org.

Other Funding Support

Support of the research covered in the course came from a variety of grant sources, including NIH/NIDCR grant U54-DE019261, NIH/NCRR grant K30-RR022258, NIH/NHLBI grant R01-HL091005 and NIH/NIAAA grant R01-AA013458-01

Behavioral and Social Science Focus

The course was developed in response to RFA-OD-13-009, issued by the National Institutes of Health Office of Behavioral and Social Sciences Research, in cooperation with a wide range of the NIH Institutes and Centers that fund behavioral and social science research.

Behavioral and Social Science Focus

The course reflects the behavioral and social science focus of the RFA. The faculty of the course includes experts in behavioral and social science, along with colleagues who specialize in biostatistics, computer science, and health disparities.

Behavioral and Social Science Focus

Most example studies used in the course will be drawn from the published behavioral and social science literature, as well as from ongoing, NIH funded behavioral and social science studies.

Course software

The free, open-source, web-, tablet- and smart phone-based software used in the course (www.glimmpse.SampleSizeShop.org) was developed by Sarah Kreidler, DPT, MS, PhD, in collaboration with a team of colleagues including Glueck and Muller.

Course software

The software has a point-and-click, wizard style interface, and requires no programming, no mathematical notation, and no knowledge of software languages. Anyone with a web-browser can use it.

Course software

The software GLIMMPSE 2.0.0 extends an existing program called POWERLIB to include the work of Glueck and Muller (2003), and adds a graphical user interface.

Course software

Development of GLIMMPSE 1.0.0 was funded by an American Recovery and Re-investment Act supplement (3K07CA088811-06S) for NCI grant K07CA088811. Both the parent grant and the supplement were awarded to the University of Colorado Denver, (Deborah Glueck, PI).

Course software

Additional funding was provided by NIDCR 1 R01 DE020832-01A1 to the University of Florida (Keith Muller, PI; Deborah Glueck, University of Colorado site PI). The funding from NCI allowed the development of the GLIMMPSE software.

Course software

The additional funding from NIDCR permitted extensive architecture changes which support power for many general linear mixed models, as well as supporting beta testing and final software release activities.

COURSE TEAM

Deb Glueck, PhD
Co-PI, Course Director & Instructor

Associate Professor, Departments of Biostatistics & Informatics and Radiology | UC Denver

Dr. Glueck is the author of 36 manuscripts, split between applied scientific articles, and articles on power and sample size. She was or is currently the principal investigator or co-principal investigator of four NIH grants to develop and disseminate methods taught in this course. She is the lead biostatistician on research collaborations funded by a variety of NIH Institutes. She previously was PI of a NIH/NCI funded K award and co-PI on an NIH/NCI funded R03. Dr. Glueck is a winner of multiple mentoring awards at the University of Colorado Denver.

Keith E. Muller, PhD
Co-PI, Lecturer

Associate Chair, Department of Health Outcomes & Policy | University of Florida

Dr. Muller is a fellow of the American Statistical Association and first author of two widely-used texts on linear models. He has co-authored more than 130 peer-reviewed articles, with roughly half methodological and half collaborative. He was or is currently the principal investigator or co-principal investigator of four NIH grants to develop and disseminate methods taught in the course. His collaborations are funded by a variety of NIH and other government agencies.

Anna Barón, PhD
Lecturer

Professor, Department of Biostatistics & Informatics | UC Denver

Dr. Barón has over 30 years of experience collaborating with researchers in a number of areas including cancer, cardiology, health services research, cross-cultural psychiatry, diabetes, and HIV/AIDS prevention. Her statistical expertise is in discriminant analysis, survival analysis, and methods in statistical epidemiology. Dr. Barón's support of translational education is motivated by two ongoing collaborations, where the methodology and software taught in this course would have greatly enhanced planning at the design stage.

Jessica R. Shaw, MS student
Course Developer, Teaching Assistant

Student, MS Biostatistics – Statistical Genomics and Genetics Track | UC Denver

Jessi Shaw earned a BS in Marketing Management and International Business from Purdue University in 2013. Before pursuing a career in health research, she worked for the world's largest market research firm, providing big data analytics support to multiple Fortune 500 companies. Research experiences at the Purdue University Center for Cancer Research and the UC Davis MIND Institute motivated her to study biostatistics. In her time at UC Denver, Shaw has served as a Teaching Assistant for four courses and managed the restructuring of three.

COURSE LOGISTICS

Fulginiti Pavilion facilities

Additional bathrooms are located in the basement.

A lactation room is available in the basement via UC Denver badge access. If you do not have a UC Denver ID, ask an instructor for access.

Course emergency Contact:

Jessi Shaw, Teaching Assistant
Cell: 815-975-7515
jessica.shaw@ucdenver.edu

IMPORTANT: NO MARKERS ALLOWED

Dr. Glueck has a severe allergy to the ink contained in markers.

Please do not bring any markers into class or lab sessions.

COURSE SCHEDULE & POLICIES

Schedule: Day 1

1/20/16	Event/Task
9am - 12 pm	Lecture
12 - 1 pm	Lunch lecture on “Ethics of Power and Sample Size”
1 - 2 pm	Optional accessible walk and talk with faculty or break time
2 - 5 pm	Lecture

Schedule: Day 2

1/21/16	Event/Task
9am - 12 pm	Lecture
12 - 1 pm	Working lunch with faculty on class projects or optional break
1 - 2 pm	Optional accessible walk and talk with faculty or break time
2 - 5 pm	Lecture

Schedule: Day 3

1/22/16	Event/Task
9am - 12 pm	Lecture
12 - 1 pm	Working lunch with faculty on class projects or optional break
1 - 2 pm	Optional accessible walk and talk with faculty or break time
2 - 5 pm	Lecture

Schedule: Assignments 1-4

Date	Time	Event/Task
1/25/16	9-11 am	In-person or online lab: Assignments 1-4
	5 pm	**DUE:** Assignments 1-4

Schedule: Final assignment

Date	Time	Event/Task
1/27/16	12-1 pm	In-person or online lab: Final assignment
1/28/16	4 - 6 pm	In-person or online lab: Final assignment

Schedule: Final assignment

Date	Time	Event/Task
1/29/16	12 pm	**DUE:** 1 page statistical analysis plan and power or sample size analysis submitted via e-mail to Deborah.Glueck@ucdenver.edu. No in-person class.

Schedule: Final assignment feedback

Date	Time	Event/Task
2/4/16	5 pm	Graded statistical analysis plan and power or sample size analysis returned to students via e-mail. No in-person class.

Schedule: Final assignment feedback

Date	Time	Event/Task
2/5/16	12-1 pm	Interactive video conference via GoToMeeting with students and course faculty to answer questions on graded assignment. Optional in-person class.

Schedule

Given that this is an intensive 3-day course, we will take breaks every hour. Please feel free to eat in class and take additional breaks at your discretion. Please let your instructor know if you require additional accommodations.

Special considerations

The professors will gladly accommodate participants with physical disabilities or diagnosed learning disabilities, upon request.

Special considerations

For students requesting accommodations, contact the Office of Disability Resources and Services. Their staff will assist in determining reasonable accommodations as well as coordinating the approved accommodations. Phone number: (303) 724-5640. Location: Building 500, Room W1103. The physical address is 13001 E. 17th Place.

Disability Resources and Services Office

For nursing mothers, lactation rooms are available near the classroom. For parents with responsibility for small children who cannot attend the class in person, it may be possible to attend the course remotely via video-conferencing, upon consent of the instructors.

Attendance and missing classes

If receiving credit for this course, students must attend 5 out of 6 half days and all post-course meetings.

Students who do not meet this requirement will not be eligible for course credit.

Attendance and missing classes

Please e-mail us with specific concerns: Deborah.Glueck@ucdenver.edu.

If you miss class time, materials are in the course book.

ASSIGNMENTS AND GRADING

Assignments (15% per assignment)

Course learners will complete four small assignments, with rapid feedback, that will prepare students for the final assignment.

Assignments (15% per assignment)

Students will complete power and sample size analyses using GLIMMPSE for four example studies used in lectures. In completing the assignments, students will have access to assistance from the instructors and peers.

Assignments (15% per assignment)

For each example study, students will prepare a document that describes:

1. Study aims
2. Study design
3. Sampling design, recruitment period and the potential for missing data and dropout

Assignments (15% per assignment)

For each example study, students will prepare a document that describes:

4. A statistical analysis plan
5. Numerical results, tables and graphs for a power and sample size analysis

Final project (40%)

As the final project, each student will complete a power and sample size analysis for a fifth example study. One week after the end of lectures, students will turn in their respective final projects via e-mail.

Final project (40%)

The statistical analysis plan and power or sample size analysis should be a one page document with the following elements:

1. Statistical analysis plan
2. Power or sample size analysis
3. Power graph or table

Homework assignments

You will now receive the four assignments and the final project. You will also receive detailed solutions to the assignments. In completing the assignments, we encourage you to work with others, check your work against the solutions, and ask questions in the lab sessions. We highly recommend that you complete each assignment on your own before consulting the solutions.

Homework assignments

Labs and solutions are provided to help you become comfortable with **GLIMMPSE** and the process of writing a power and sample size analysis. Please use the assignments as an opportunity to master the material before the final project. While you will allowed to ask questions, you will be expected to complete the final project independently.

COURSE INTRODUCTION

The course will help you learn how to build studies that are successful

The goal of the course is to assist you in learning how to do power and sample size analyses for multilevel and longitudinal studies.

The goal of most scientific research is to test hypotheses

Hypotheses are theories about how the world works.

Hypotheses often describe the relationship between two or more events and allow scientists to describe average response for two or more groups.

For the course, you must first understand power and hypotheses

Power is the probability of rejecting a null hypothesis.

The scientific goal is to reject the null hypothesis.

In study design, power is used to calculate how many participants should be included in a study

Desired power → Required sample size

Statistics are simply tools for evaluating hypotheses

Statistics allow scientists to evaluate the chance that the null hypothesis is in fact true, given the observed data.

Scientific research depends on the calculation of probabilities

Cumulative distribution functions are equations used to calculate the probability of an outcome for given values of an independent variable.

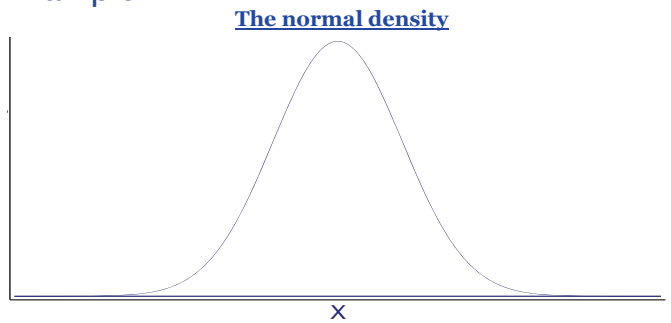
Example: Given a patient's daily diet contains 4000mg of sodium, the probability of a heart attack by age 60 is 20%.

Scientific research depends on the calculation of probabilities

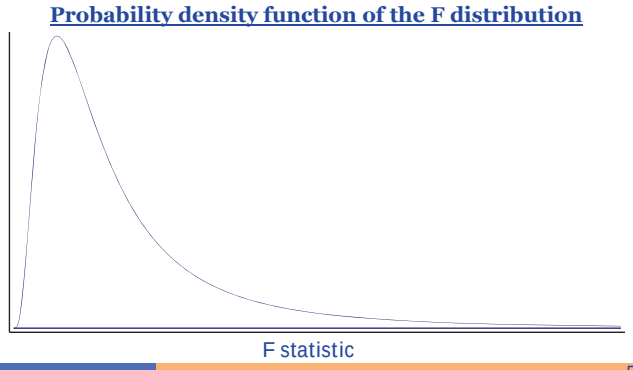
Many useful random variables that are continuous have density functions which are used to compute their cumulative distribution functions and in turn probabilities of interest.

When graphed, probability density functions take on highly characteristic shapes

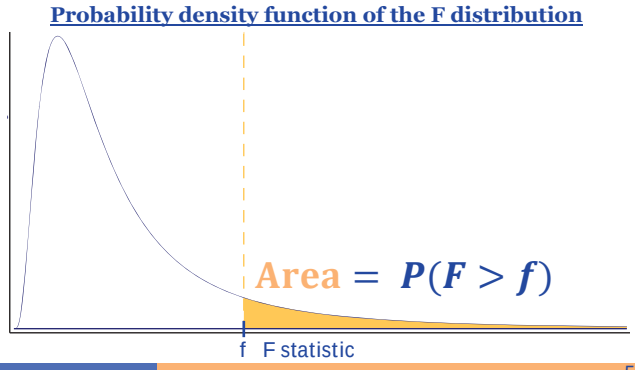
Example:



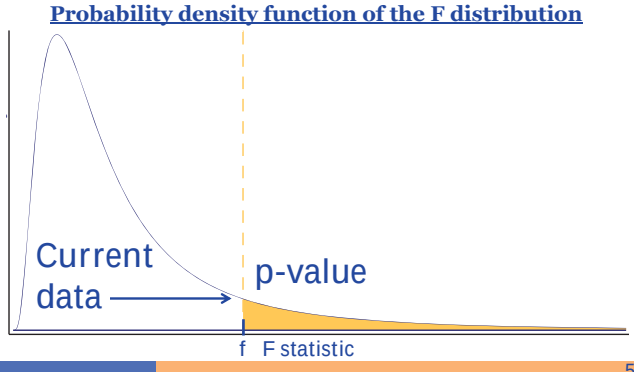
The central F distribution, frequently seen in multilevel and longitudinal studies, forms a skewed bell shape



Probability is calculated by finding the area under the curve of a probability density function



For tests based on F, a p-value is the probability of seeing data more extreme than the current data if the null is true



Small p-values lead researchers to achieve their goal of rejecting the null hypothesis

A Type I error occurs when a statistical test mistakenly rejects the null hypothesis when in fact the null hypothesis is true.

We only trust findings if several studies come to the same conclusion

Replicated findings are findings from subsequent studies that match the original findings.

Scientific knowledge grows from the steady accretion of replicated studies.

Science: public and replicable.

Outside of scientific studies, we usually think two events are related

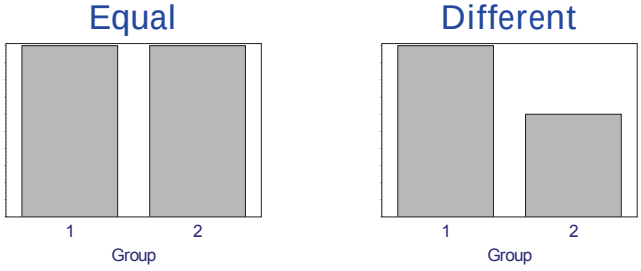
A null hypothesis claims that the two events are NOT related.

H_o is a common notation for the null hypothesis of no association

We *avoid notation* in the course and will redefine the null hypothesis every time we discuss it.

In this course we focus on hypotheses about average responses for groups

Average responses for two or more groups are either:



Power is the probability of rejecting a null hypothesis

Achieving an appropriately high power, such as 0.9, lends researchers confidence in the accuracy of a study's conclusions.

Power depends on only a few things

1. Study design, including sample size (how many participants are in the study)
2. Strength of relationship between events
3. Degree of variability in the system
4. Statistical test
5. Type I error rate

The course covers skills difficult to find elsewhere

At the end of the course you will be able to conduct sample size calculations for a wide variety of longitudinal and multilevel designs.

Accurate power and sample size are important for ethical, monetary, and time reasons

Incorrect sample sizes pose ethical dilemmas.

We will discuss the dilemmas in detail in a later lecture.

The course covers a set of skills applicable to many study designs

Many studies, from observational epidemiology to randomized controlled clinical trials, have longitudinal or multilevel features, or both.

The course includes instruction in the use of GLIMMPSE

GLIMMPSE is a free, point-and-click, open-source software package for multilevel and longitudinal power and sample size calculations.

GLIMMPSE is easily accessible

GLIMMPSE is available via most web browsers at www.SampleSizeShop.org.

The GLIMMPSE app is available via iTunes and Google Play.

You will learn to conduct accurate power and sample analyses in three steps

1. Define and understand the statistical inputs for power and sample size analysis
2. Use **GLIMMPSE** software to conduct your own calculations
3. Write the statistical, power and sample size section of grants

Day 1 Agenda

9 am – 12 pm

1. Longitudinal Studies
2. Studies with Single Level of Clustering
3. Studies with Multiple Levels of Clustering
4. Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Lunch lecture on “Ethics of Power and Sample Size”

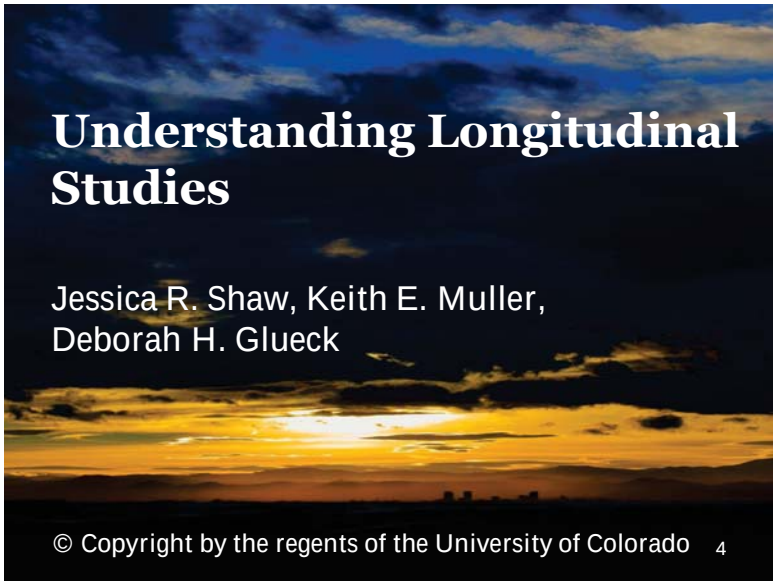
1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes



Learning Objectives

Define independent sampling unit (ISU).

Define unit of observation.

Define correlation.

Learning Objectives

Define between- and within-independent sampling unit factors.

Define longitudinal studies.

Learning Objectives

Understand the definition of a longitudinal study.

Recognize an example of a longitudinal study.

Understand how longitudinal designs result in complex correlation between measurements.

Each study can be described by its analytical units

Independent sampling unit (ISU) are statistically independent from any other unit," (p. 101).

The person is often the independent sampling unit.

Muller and Stewart, 2006

Each study can be described by its analytical units

The unit of observation is the measurement of interest within the independent sampling unit.

Muller and Stewart, 2006

Longitudinal studies are common in social and behavioral health research

A longitudinal study evaluates a research question by analyzing two or more measurements on the same research participant over time.

Correlation measures the straight-line relationship between two variables

Correlation "indicates the strength and direction of the relationship between two random variables," (p. 127).

Rosner, 2010, p. 127

Correlation ranges from -1 to 1

Correlation

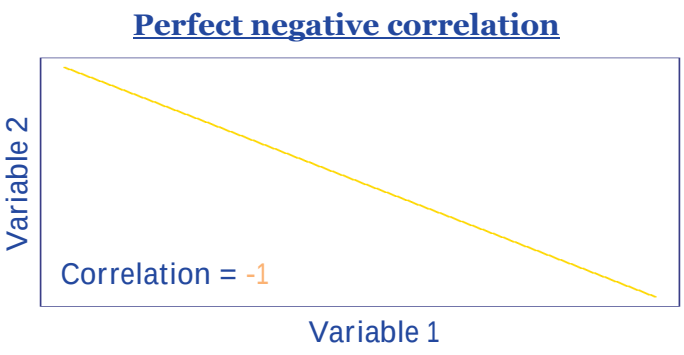
The sign of correlation indicates whether two variables are changing in the same direction

Negative correlation (between -1 and 0) indicates that two variables change in opposite directions.

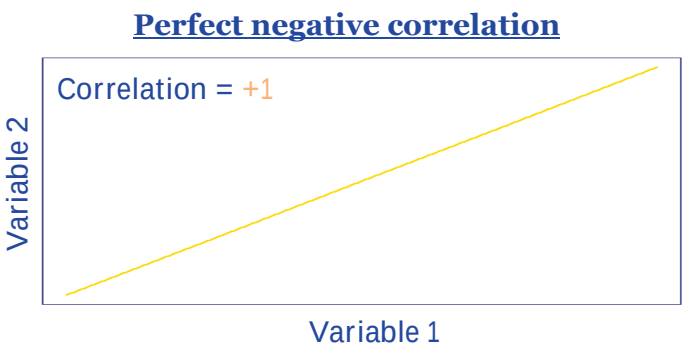
Positive correlation (between 0 and 1) indicates that variables change in the same direction.

Numbers further from zero represent stronger correlation

A correlation of -1 indicates that two variables change at constant rates in opposite directions



A correlation of +1 indicates that two variables change at constant rates in the same direction



A correlation of 0 indicates that two variables are unrelated

Variables with correlation equal to zero have rates and directions of change that are uncorrelated, or unrelated.

Variables that are independent always have correlation equal to zero

Mathematically speaking, variables with correlation equal to zero are *not necessarily* independent.

For the purposes of this class, we will focus on cases where correlation of zero indicates that two variables are independent.

Longitudinal studies induce correlation between measurements

Measurements from the same person taken at two or more times will be correlated.

Longitudinal measures are a type of repeated measures.

We discuss study designs in terms of 'factors'

A 'factor' indicates a dimension of interest such as time, or treatment versus no treatment.

Longitudinal studies may involve within-independent sampling unit factors

Within-independent sampling unit factors are features with values which differ within each independent sampling unit.

Example:

Distance walked, with different values for each day, Monday through Friday

Bray and Maxwell, 1985
Doncaster and Davey, 2007

Longitudinal studies may include between-independent sampling unit factors

Between-independent sampling unit factors take only one value for a single independent sampling unit. Different ISUs may have different values.

Example: Drug assignment in a randomized clinical trial

Bray and Maxwell, 1985
Doncaster and Davey, 2007

We examine a longitudinal study of pain perceived after a root canal

Vignette

Researchers conducted a study to determine if dental patients who are instructed to use a sensory focus have a different pattern of long-term memory of pain than participants who did not.

Logan, Baron and Kohout, 1995

We examine a longitudinal study of pain perceived after a root canal

Vignette, continued

Participants were selected and randomly assigned to either the intervention or non-intervention groups.

Logan, Baron and Kohout, 1995

We examine a longitudinal study of pain perceived after a root canal

Vignette, continued

Those in the intervention group listened to automated audio instructions to pay close attention only to the physical sensations in their mouth.

Logan, Baron and Kohout, 1995

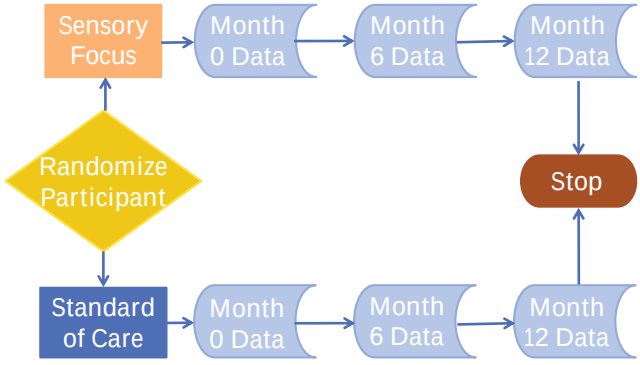
We examine a longitudinal study of pain perceived after a root canal

Vignette, continued

Patients in the no intervention group listened to automated audio instruction on a neutral topic to control for media and attention effects.

Logan, Baron and Kohout, 1995

Observed pain after root canal was measured at 0, 6, and 12 months



Logan, Baron and Kohout, 1995

The study assessed the impact of treatment on post-root canal pain

Null hypothesis:

Participants receiving the sensory focus treatment experience the same pattern of pain over time as participants receiving the standard of care treatment.

Logan, Baron and Kohout, 1995

The study assessed the impact of treatment on post-root canal pain

Independent sampling unit: Patient

Unit of observation: Patient perceived pain at a specific time

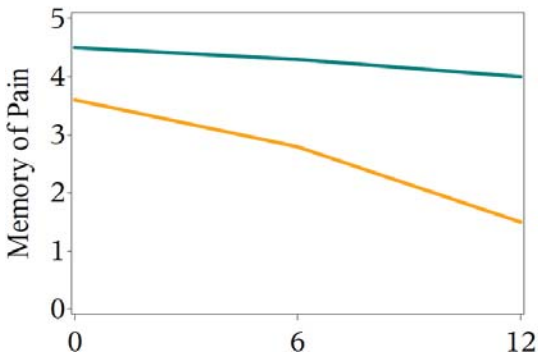
Measurements taken from a given patient at months zero, six, and 12 were correlated

Within-independent sampling unit factor: Time, with measurements of perceived pain recorded for each person at month 0, 6, and 12

Outcomes were analyzed for each individual over a 12 month period

Between-independent sampling unit factor: Treatment assignment to standard of care or sensory focus

Pain recall over time differed by treatment group



REVIEW OF LEARNING OBJECTIVES

Define a longitudinal study

A study that aims to answer a research question by analyzing two or more measurements on the same independent sampling unit (often a research participant) over time.

Define independent sampling unit

Observations from one independent sampling units are statistically independent of observations from any other distinct unit (p. 101).

Muller and Stewart, 2006

Which measure indicates both the direction and strength of an association?

- Amplitude
- Correlation

Explain how within- and between-independent sampling factors differ

Within-person factors occur within an independent sampling unit whereas between-person factors occur for each independent sampling unit as a whole.

How do longitudinal designs result in correlation between measurements?

Two or more measurements from one person are not independent from one another.

Day 1 Agenda

9 am – 12 pm

1. Longitudinal Studies
2. Studies with Single Level of Clustering
3. Studies with Multiple Levels of Clustering
4. Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Lunch lecture on “Ethics of Power and Sample Size”

1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes

Learning Objectives

Define a level of data.

Define a single level study.

Understand how single level cluster designs induce correlation.

Learning Objectives

Learn to recognize a single level study.

Recognize that some authors use the terms group- or cluster-randomized trial or observational cluster or hierarchal study, instead of using the level terminology.

Learning Objectives

Define a level of data.

Define a single level study.

Understand how single level cluster designs induce correlation.

A level of data is a position within a hierarchy of units

Study designs can have any number of levels.

We will examine both single- and multi-level models.

To better understand levels of data, consider the hierarchical structure of schools

School

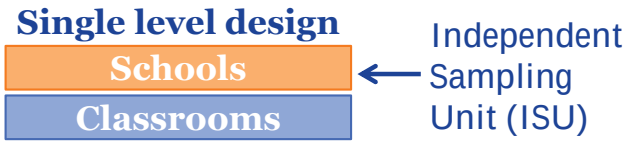
Classroom within school

Students within classrooms

Single level studies involve only one layer of correlation

Levels of correlation:

1. Classrooms within schools are correlated.



Multilevel designs can involve any number of levels

Levels of correlation:

1. Classrooms within schools.
2. Students within classrooms.

Two level design

Single level studies contain just one layer of correlation

Units of observation within the independent sampling unit are correlated.

Learning Objectives

Learn to recognize a single level study.

Recognize that some authors use the terms group- or cluster-randomized trial or observational study, instead of using the level terminology.

To better understand single level designs, we discuss an example

Vignette

A single level study examined the efficacy of a workplace training program to reduce alcohol consumption. Researchers randomized workplaces to two treatment groups.

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

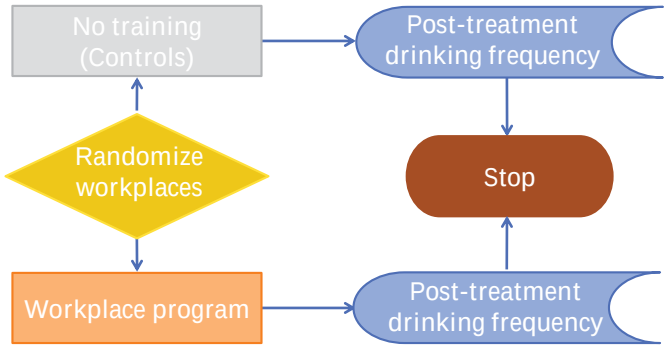
To better understand single level designs, we discuss an example

Vignette, continued

The first treatment included a workplace training program and the second treatment included no training. Post-treatment drinking frequency for each worker was measured as the outcome of interest.

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

Employees were surveyed after randomization to study groups

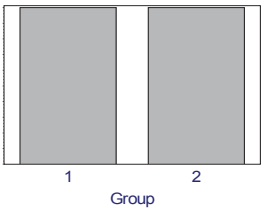


Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

Researchers hoped to observe an association between treatment and reduction in drinking

Null hypothesis:

There is no difference in post-treatment drinking frequency between workers who receive no training and workers who receive the workplace program.



This was a single level design with one outcome measurement in time

Level of correlation: Workers in the workplace

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

This was a single level design with one outcome measurement in time

Independent sampling unit: Workplace

Unit of observation: Drinking frequency for each employee after treatment

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

This study evaluated the impact of treatment on employee behaviors

Between-independent sampling unit factor:

Intervention (standard of care or workplace drinking program)

Within-independent sampling unit factor:

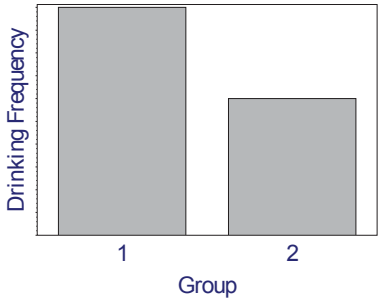
Cluster membership

Interest in factor depends on setting

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

Analysis compared post-treatment drinking frequency between the two treatment groups

Scientific goal: **REJECT** the null hypothesis



Measurements of drinking frequency of workers within a workplace are correlated

Some workers drink together.

Some workers attend sobriety programs together.

Workers discuss drinking habits with each other.

Randomization allowed researchers to make two important assumptions

1. The average correlation between workers in intervention groups **is equal to** the average correlation between workers in the control groups.
2. Pre-existing employee factors did not bias study outcomes.

Learning Objectives

- Learn to recognize a single level study
- Recognize that some authors use the terms group- or cluster-randomized trial or observational study, instead of using the level terminology

Some authors use the terms group- or cluster randomized trial or observational study

Authors use these terms instead of using the "level" terminology.

Recall the alternative terminology when reviewing manuscripts.

Murray, D. M., 1998

Groups are collections of smaller units which share similarities

Groups are non-random, naturally occurring features of hierarchies, "identified by some physical, geographic, social, or other connection among their members," (p. 1).

Murray, D. M., 1998

Groups are collections of smaller units which share similarities

Examples of groups include:

- Worksites
- Schools
- Hospitals
- Clinics
- Neighborhoods
- Cities

Murray, D. M., 1998

Research designs often involve groups

A group randomized trial involves random assignment of "identifiable groups that are not constituted at random," (p.1).

Murray, D. M., 1998

Clusters are synonymous with groups

Survey sampling authors prefer the term "cluster" to group.

REVIEW OF LEARNING OBJECTIVES

What is a single level design?

A research design with only one layer of correlated units.

In the example, at what level was the data analyzed?

Randomized study group

In the example design, what were the independent sampling units?

Workplaces

In the example, which units were correlated?

Workers within a workplace

ADDITIONAL REVIEW OF LEARNING OBJECTIVES

To review, we examine another study

Vignette

Scientists conducted a study to assess the feasibility of using health information technology to administer a health risk assessment called the Young Adult Health Care Survey to adolescents.

Adapted from Kadivar et al., 2014

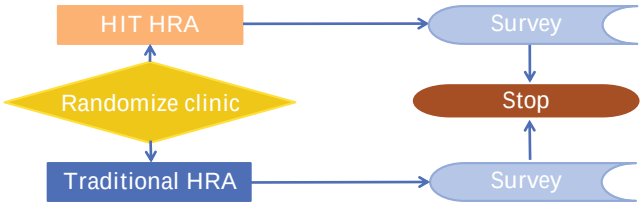
Nesting

Vignette, continued

By necessity, 9 of 11 clinics were assigned to the intervention and 2 to the comparison condition. Patients completed the risk assessment survey and then rated the quality of their health care experience.

Adapted from Kadivar et al., 2014

Before answering some questions, study the following diagram



Acronym Key	
HRA	Health Risk Assessment
HIT	Health Information Technology
YAHCS	Young Adult Health Care Survey

Adapted from Kadivar et al., 2014

What type of study is this?

- Longitudinal
- Single level
- Multilevel

Adapted from Kadivar et al., 2014

What was the null hypothesis?

Null hypothesis:

There is no difference in adolescents' reported quality of care between clinics assigned to use health information technology and clinics assigned to use a traditional health risk assessment

Adapted from Kadivar et al., 2014

What was the independent sampling unit?

Independent sampling unit: Clinic

Adapted from Kadivar et al., 2014

What was the unit of observation?

Unit of observation: Patient survey scores

Adapted from Kadivar et al., 2014

How many levels of correlation were in this study?

One:

– Patients within the same clinic were correlated

Adapted from Kadivar et al., 2014

Day 1 Agenda

9 am – 12 pm

1. Longitudinal Studies
2. Studies with Single Level of Clustering
3. Studies with Multiple Levels of Clustering
4. Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Lunch lecture on "Ethics of Power and Sample Size"

1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes

Learning objectives

Define multilevel study.

Recognize a multilevel study.

Understand the advantages of a multilevel study.

Learning objectives

Recognize that some authors use the terms group- or cluster-randomized trial or observational cluster or hierarchal study, instead of using the level terminology.

Understand how multilevel studies induce correlation.

A multilevel study involves one or more hierarchical groups of study participants

Authors use different but equivalent terminology.

Examples:

- "Two-level study"
- "Two-level group design"
- "Two-level cluster design"

Members of a level, also referred to as a group or cluster, share experiences which induce correlation

Multilevel studies involve two or more layers of correlation.

Murray, D. M., 1998

For example, reading scores from students in the same school are correlated

Shared teacher → induces a correlation

Shared school → induces a correlation

Two level design

Schools	ISU
Classrooms	Level 1
Students	Level 2

Murray, D. M., 1998

We examine an example of a multilevel, randomized controlled trial

A randomized controlled trial is a study in which participants are randomized into either a study group which receives an intervention or a control group which does not (p. 139).

May have more than two groups.

Last, John M, 2000

The study assessed the effectiveness of a literacy intervention

Vignette

Researchers conducted a cluster randomized control trial to evaluate the effectiveness of a web-based literacy intervention called ABRACADABRA (ABRA).

Adapted from Piquette, et al., 2014

The study assessed the effectiveness of a literacy intervention

Vignette, continued

The study included 24 classrooms within 12 elementary schools within a single school district. Researchers assumed that schools within the district were under local control and were therefore independent.

Adapted from Piquette, et al., 2014

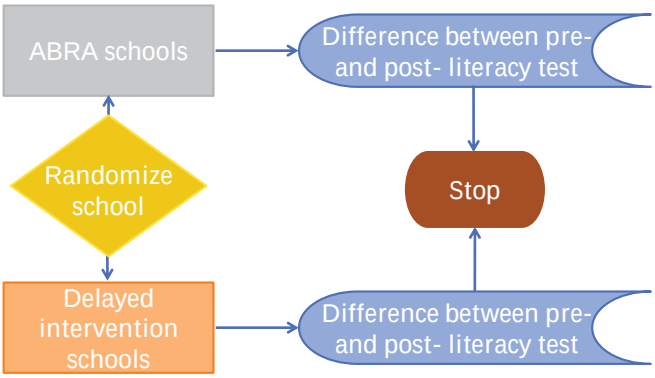
The study assessed the effectiveness of a literacy intervention

Vignette, continued

Schools were randomized into the intervention group or the control group. Change in literacy was evaluated using pre- and post-tests to determine whether ABRA technology significantly improved literacy in elementary school children.

Adapted from Piquette, et al., 2014

Randomization of study conditions took place at the school level



Adapted from Piquette, et al., 2014

The goal of the study was to assess the effectiveness of a literacy intervention

Null hypothesis:

There is no significant difference in literacy between elementary students in the intervention group and those in the control group.

Adapted from Piquette, et al., 2014

Researchers adjusted for correlation to isolate the impact of the intervention

Levels of correlation:

- Classrooms within schools
- Students within each classroom

Failure to adjust for correlation would inflate type I error rate.

Adapted from Piquette, et al., 2014

Change in literacy scores was chosen as the measure of intervention effectiveness

Independent sampling unit:

School

Unit of observation:

Difference between pre- and post-test performance

Adapted from Piquette, et al., 2014

The goal of the study was to assess the effectiveness of a literacy intervention

Between-independent sampling unit factor:

Randomization group

Within-independent sampling unit factor:

Cluster member

Interest in factor depends on setting

Adapted from Piquette, et al., 2014

ABRACADABRA was associated with greater improvements in early literacy

By accounting for existing organizational groups, the multilevel design enabled researchers to isolate the ABRACADABRA program as the primary cause of literacy improvement.

Adapted from Piquette, et al., 2014

Can you describe a multilevel study?

A research design with multiple layers of correlated units.

REVIEW OF LEARNING OBJECTIVES

Day 1 Agenda

9 am – 12 pm

1. Longitudinal Studies
2. Studies with Single Levels of Clustering
3. Studies with Multiple Levels of Clustering
4. Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Lunch lecture on “Ethics of Power and Sample Size”

1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes

Learning Objectives

Understand the definition of a multilevel and longitudinal study.

Recognize an example of a multilevel and longitudinal study.

Understand how multilevel and longitudinal designs result in correlation between measurements.

Recall the definition of a longitudinal study

A longitudinal study evaluates a research question by analyzing two or more measurements on the same research participant over time.

Recall the definition of a multilevel study

A multilevel study involves two or more layers of correlation.

Observations within a level, also referred to as a group or cluster, experience similar influences, which induces correlation.

Studies with longitudinal and multilevel features have multiple types of correlation

Longitudinal features induce correlation within levels, also known as groups or clusters.

Multilevel features induce correlation within levels, also known as groups or clusters.

We examine a study with both multilevel and longitudinal features

Vignette

Researchers conducted a group-randomized, multilevel longitudinal study to test the effectiveness of a preventive alcohol use intervention. The study focused on urban, low-income, and multi-ethnic populations.

Komro et al., 2008

We examine a study with both multilevel and longitudinal features

Vignette, continued

Schools in 22 communities were recruited. Communities were randomized into an intervention or control group.

Komro et al., 2008

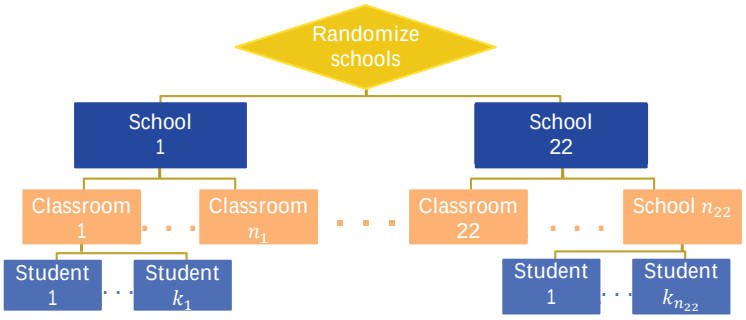
We examine a study with both multilevel and longitudinal features

Vignette, continued

Students were surveyed at the end of 6th grade, 7th grade, and 8th grade to measure alcohol use.

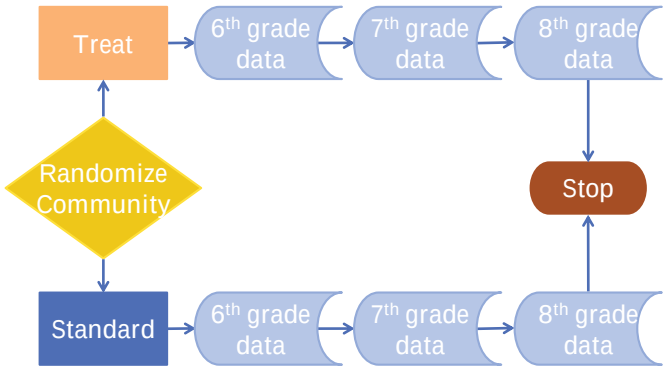
Komro et al., 2008

We examine a study with both multilevel and longitudinal features



Komro et al., 2008

We will examine a study with both multilevel and longitudinal features



Komro et al., 2008

This study aimed to assess the effectiveness of an intervention

Null hypothesis:

There is no significant difference in average alcohol use between communities that receive the standard of care and communities that receive the intervention.

Komro et al., 2008

Students were observed within schools in each community

Independent sampling unit: Community

Unit of observation: Measurement at one point in time for one student

Komro et al., 2008

The study investigated the impact of the proposed intervention over time

Within-independent sampling unit factor: Time, with repeated measurements taken at the end of 6th grade, 7th grade, and 8th grade

Between-independent sampling unit factor:

Assignment to standard of care or treatment

Komro et al., 2008

This study involved two types of correlation

1. Multilevel:

   Correlation within groups or clusters

2. Longitudinal:

   Correlation between repeated measurements

Komro et al., 2008

Recall, levels experience common influences which induce correlation

Schools within the same community were correlated.

Murray, 1998

Further, longitudinal studies induce correlation between measurements

Measurements from the same student at each time point were correlated.

Rosner, 2010, p. 127

Unfortunately, Project Northland Chicago was ineffective

Findings revealed the need for population-specific intervention studies.

Komro et al., 2008

REVIEW OF LEARNING OBJECTIVES

True or false?

Any given study can have only one type of correlation.

TRUE **FALSE**

True or false?

Measurements on within-independent sampling unit factors will not be correlated.

TRUE **FALSE**

True or false?

Measurements on between-independent sampling unit factors will not be correlated.

TRUE **FALSE**

Day 1 Agenda

9 am – 12 pm

1. Longitudinal Studies
2. Studies with Single Levels of Clustering
3. Studies with Multiple Levels of Clustering
4. Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Lunch lecture on “Ethics of Power and Sample Size”

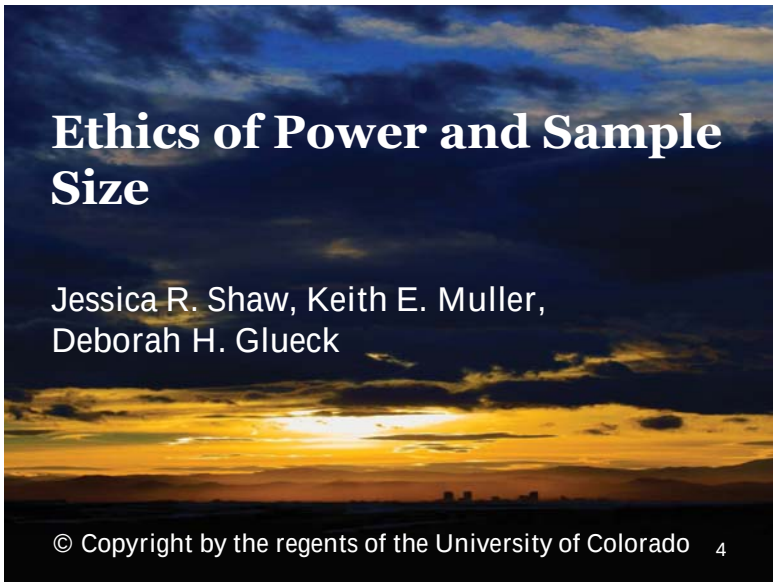
1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes



Learning Objectives

Understand some basic ethical concepts underlying trial design.

In the context of study design, define equipoise, therapeutic misconception, informed consent.

Learning Objectives

Understand ethical concerns in overpowered studies.

Understand ethical concerns in underpowered studies.

Understand the importance of early sample size and power planning.

BROAD ETHICAL CONCEPTS IN STUDY DESIGN

Four key concepts underlie the ethics of power and sample size

1. Informed consent
2. Equipoise
3. Therapeutic misconception
4. Ethical balance

Informed consent is an ethical requirement of research intended to protect research participants

Informed consent requires that individuals freely agree to participate after being thoroughly informed of the purpose, risks, and benefits of the research.

Investigators must inform potential participants of foreseeable limitations to the value of the research

Participants must be informed if the study requires a sample size that will be difficult to recruit.

Faden et al., 2014

Ethical scientific research requires the assumption of equipoise between study conditions

Equipoise is the assumption that study conditions are equally useful.

The null hypothesis is that study conditions are in equipoise

Recall, the null hypothesis assumes that a study factor is NOT associated with a difference in outcomes.

Example: The effect of Treatment A is equal to the effect of Treatment B.

Therapeutic misconception occurs when a participant or an investigator confuses research with treatment

Treatment is the administration of a known effective solution to a health problem.

Research is conducted in order to develop effective solutions to health problems.

Henderson, GE, et al., 2007

Research does not promise the same benefits to participants as treatment

Treatment directly benefits the *individual*.

Research helps *society* as a whole and may or may not benefit individual participants.

Henderson, GE, et al., 2007

Several factors may contribute to therapeutic misconception

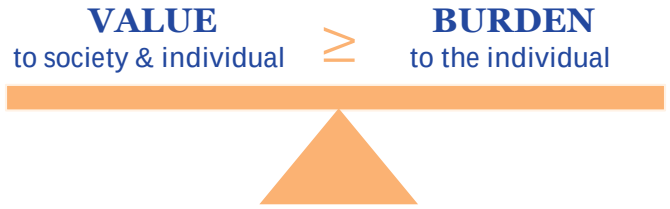
An *investigator* may be overly optimistic about the benefits of participation in her research.

A *participant* may lack the scientific background necessary to fully understand the risks, benefits, and limitations of the research.

Henderson, GE, et al., 2007

Researchers have a responsibility to maintain ethical balance

Ethical balance is maintained if the value of research to society is greater than or equal to the burden to the individual.



Bacchetti et al., 2005

Researchers have a responsibility to maintain ethical balance

The burden to the individual is the projected risk, discomfort, or inconvenience to the participant.

The value to society is the projected value of research to participants or to society as a whole.

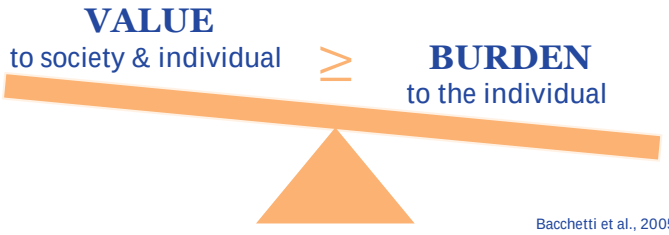
Bacchetti et al., 2005

The primary potential benefit to a study participant is the receipt of an *effective* treatment or intervention

Additional benefits may include incidental care and diagnostics, provided as part of the study, at no cost to the participant.

Studies providing minimal benefit to participants can be unethical

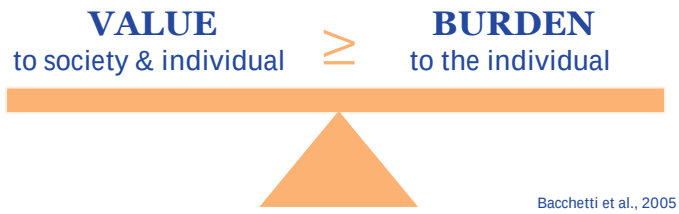
Research generates a net burden when the projected value of research to society is less than the projected burden to the individual.



Bacchetti et al., 2005

Ethical balance depends on the cumulative benefit of research to the individual and to society

A study which generates an *individual* net burden may still maintain ethical balance.



Bacchetti et al., 2005

Proposed research studies involving human participants must be approved by an IRB

An institutional review board (IRB) is responsible for reviewing ethical concerns in proposed studies.

The IRB may decide that a study is exempt from full review.

Ethical balance is the basis of all Institutional Review Board (IRB) approvals and rejections

The IRB evaluates the risk to benefit ratio of each study.

Inaccurate power and sample size calculations can upset the ethical balance

We will discuss ethical dilemmas posed by underpowered studies and overpowered studies.

ETHICAL DILEMMAS IN UNDERPOWERED STUDIES

Underpowered studies may result from a failure to distinguish between efficacy and effectiveness

Efficacy is a measure of mean treatment effect in a perfect world.

Effectiveness is a measure of treatment effect given imperfect compliance with treatment.

Power calculated assuming perfect compliance may not be sufficient to detect effects

Compliance is rarely perfect.

One should calculate power under real world assumptions about compliance.

Underpowered studies can fail to maintain ethical balance

Underpowered studies reduce the potential to find clinically significant differences.

Underpowered studies use scarce resources that could otherwise be used by other researchers.

Underpowered studies place burden on participants without adequate benefits.

Lenth, 2001

Researchers are responsible for forecasting realistic sample sizes in grant applications

Researchers must account for potential loss to follow-up.

Researchers must recruit the sample size projected in grant.

Lenth, 2001
Edwards et al., 1997

Why might you choose to do a trial which is underpowered?

It may simply be impossible to accrue a large enough sample for the study.

Example:

Recruitment may be difficult in a population which is culturally averse to participation in research.

Lenth, 2001

Why might you choose to do a trial which is underpowered?

Underpowered studies may still contribute to overall scientific knowledge, even if they do not achieve high power or statistical significance.

Lenth, 2001

Why might you choose to do a trial which is underpowered?

Underpowered studies may still contribute to overall scientific knowledge, even if they do not achieve high power or statistical significance.

Why might you choose to do a trial which is underpowered?

Underpowered studies may inspire discussion and further research.

Lenth, 2001

Underpowered studies are ubiquitous in the behavioral and social sciences

Maxwell, Lau, and Howard (2015, *American Psychologist*)

Is Psychology Suffering From a Replication Crisis?
What Does "Failure to Replicate" Really Mean?

http://dx.doi.org/10.1037/a0039400

Lenth, 2001

Underpowered studies are ubiquitous in the behavioral and social sciences

"...failures to replicate may not be failures at all, but rather are the result of low statistical power in single replication studies, and the result of failure to appreciate the need for multiple replications in order to have enough power to identify true effects."

Lenth, 2001

ETHICAL DILEMMAS IN OVERPOWERED STUDIES

Overpowered studies can be unethical by exposing a greater number of individuals to harm than necessary

The burden and direct benefit to each *individual* remains the same.

The *number* of individuals burdened increases.

Studying multiple aims simultaneously may confound sample size selection

Sufficient power for one endpoint may result in too much power for another endpoint.

Researchers must balance the risks and rewards of the research.

Patient input is useful in ethically balancing risks and rewards.

Lenth, 2001

An overpowered study may detect statistically significant differences without clinical significance

Researchers must ask, “Are the results relevant?”

Lenth, 2001

Are overpowered studies common in randomized clinical trials of rheumatoid arthritis?

Celik and Yazici (2014) concluded "Most RCTs in RA enroll more patients than needed. This is costly and has the immediate consequence of exposing needless number of patients to potential harm."

List 40 specific examples

DOI: 10.1111/eci.12337

Lenth, 2001

DISCUSSION

Day 1 Agenda

9 am – 12 pm

1. Longitudinal Studies
2. Studies with Single Levels of Clustering
3. Studies with Multiple Levels of Clustering
4. Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Ethics of Power and Sample Size

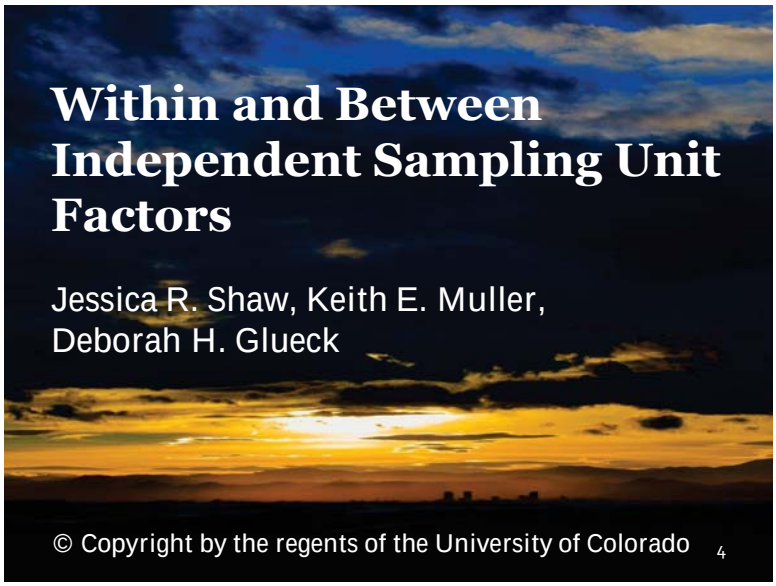
1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes



Within and Between Independent Sampling Unit Factors

Jessica R. Shaw, Keith E. Muller, Deborah H. Glueck

Learning Objectives

Define between-independent sampling unit factors.

Define within-independent sampling unit factors.

First, let's review the concept of an independent sampling unit

(ISU) Independent sampling units

Have observations which are statistically independent from observations from any other unit (p. 101).

Muller and Stewart, 2006

Longitudinal studies can involve both within- and between-independent sampling unit factors

Within-independent sampling unit factors are values which vary within independent sampling units.

Between-independent sampling unit factors occur at different levels, for each independent sampling unit as a whole.

Bray and Maxwell, 1985
Doncaster and Davey, 2007

Within-independent sampling unit factors can vary over time or space

Examples:

- Variable over time: Measurements of distance walked per day over one month
- Variable over space: Measurements of each digit of the hand

Bray and Maxwell, 1985
Doncaster and Davey, 2007

Between-independent sampling unit factors vary between groups. The groups may also be clusters in a single or multilevel design.

Examples:

- Treatment condition
- Genotype
- Gender
- Disease status

Bray, James H. and Scott E. Maxwell
Doncaster, C. Patrick and Andrew J. H. Davey

Researchers further classify factors to identify the source of variation

Observational factors: pre-existing factors which are naturally-occurring

Interventional factors: factors which are introduced by experimental design

Researchers classify study designs based on their inclusion of between and within-ISU factors

- Pure within-ISU factor design
- Pure between-ISU factor design
- Between by within-ISU factor design

Note: ISU = Independent sampling unit

Pure within-ISU designs can include observational or interventional factors

Example: Observational, within-ISU factors

Sample	Women (N=100)
# Study conditions	None
# Measurements	3 per woman
Between-ISU factor	None
Within-ISU factor	Time

Note: ISU = Independent sampling unit

Pure within-ISU designs can include observational or interventional factors

Example: Interventional, within-ISU factors

Sample	Women (N=100)
# Study conditions	3 (Treatments A, B, C)
# Measurements	3 (A, B, C)
Between-ISU factor	None
Within-ISU factor	Treatment

Note: ISU = Independent sampling unit

Similarly, pure between-ISU designs can include observational or interventional factors

Example: Observational, between-ISU factor

Sample	Men (N=100) Women (N=100)
# Study conditions	None
# Measurements	1 per participant
Between-ISU factor	Gender
Within-ISU factor	None

Note: ISU = Independent sampling unit

Similarly, pure between-ISU designs can include observational or interventional factors

Example: Interventional, between-ISU factor

Sample	Coloradans (N=200)
# Study conditions	2 (Treatments A, B)
# Measurements	1 per participant
Between-ISU factor	Treatment
Within-ISU factor	None

Note: ISU = Independent sampling unit

Between by within-independent sampling unit designs are common

Inclusion of both between and within-independent sampling unit factors allows investigators to answer more complex research questions about the pattern of response to treatment over time.

We consider how between and within-ISU factors shaped the following longitudinal studies

1. An observational longitudinal study
2. An interventional longitudinal study

1. Consider the following observational longitudinal study*

Vignette

Researchers planned a nine month, longitudinal study to assess whether a sensory focus was effective at reducing pain in root canal patients. In addition, investigators wanted to know whether it was equally effective in men and women.

*Hypothetical extension of Logan et al., 1995

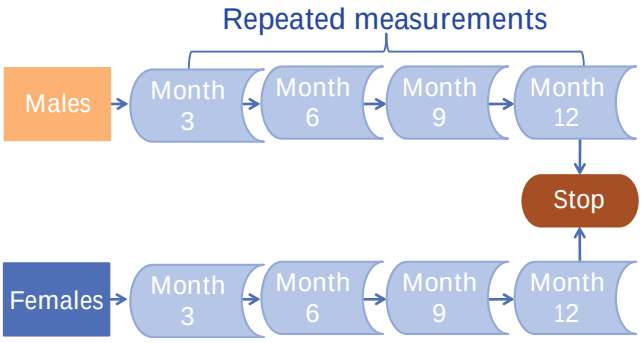
Vignette

A sample of 50 men and 50 women agreed to participate in the study. Each participant required a root canal. Patients were asked to report pain experienced at months 3, 6, 9, and 12 following the procedure.

Vignette

Researchers analyzed the data to identify differences between men and women in average pain experienced and in the pattern of pain experience over time.

After the initial visit, each participant reported pain four times



Analysis evaluated differences in pain experience associated with each of the study's ISU factors

Within-ISU analysis evaluated differences over time.

Note: ISU = Independent sampling unit

Analysis evaluated differences in pain experience associated with each of the study's ISU factors

Between-ISU analysis compared average pain reported in females to pain reported in males.

Note: ISU = Independent sampling unit

*Hypothetical extension of Logan et al., 1995

Analysis evaluated differences in pain experience associated with each of the study's ISU factors

Between by within-ISU analysis compared the pattern of patients' pain over time (time-by-treatment interaction).

Note: ISU = Independent sampling unit

*Hypothetical extension of Logan et al., 1995

2. Recall the following interventional longitudinal study*

Vignette

Researchers conducted a study to determine if dental patients who are instructed to use a sensory focus have a different pattern of long-term memory of pain than participants who did not.

Logan et al., 1995

Vignette, continued

Participants were selected and randomly assigned to either intervention or no intervention.

Logan et al., 1995

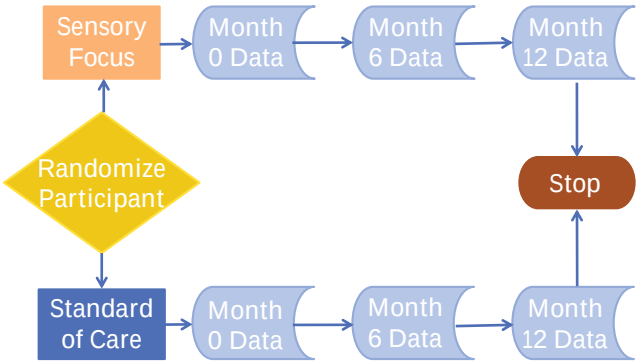
Vignette, continued

Patients in the *intervention group* listened to automated audio instructions to pay close attention only to the physical sensations in their mouth.

Patients in the *no intervention group* listened to automated audio instruction on a neutral topic to control for media and attention effects.

Logan et al., 1995

Observed pain after root canal was measured at 0, 6, and 12 months



Logan et al., 1995

Analysis evaluated differences in pain experience associated with each of the study's ISU factors

Within-person factors were used to compare the pattern of pain over time.

Between person factors were used to compare mean pain reported in the sensory focus treatment to mean pain reported in the standard of care treatment group.

Logan et al., 1995

Both studies used repeated measures, but their between-ISU factors differed

Feature	Interventional	Observational
Sample	General population (N=100)	Men (N=50) Women (N=50)
# study conditions	2 (Sensory focus, standard of care)	None
# measurements	2 per participant	2 per participant
Between-ISU factor	Treatment	Gender
Within-ISU factor	Time	Time

Note: ISU = Independent sampling unit

*Hypothetical extension of Logan et al., 1995

REVIEW OF LEARNING OBJECTIVES

Indicate whether the following scenarios are between- or within-factors

We measure the height of each individual participant at ages 8, 12 and 16.

BETWEEN **WITHIN**

Indicate whether the following scenarios are between- or within-factors

We compute the average height for males and females.

BETWEEN **WITHIN**

Day 1 Agenda

9 am – 12 pm

1. Longitudinal Studies
2. Studies with Single Levels of Clustering
3. Studies with Multiple Levels of Clustering
4. Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Lunch lecture on “Ethics of Power and Sample Size”

1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes

Learning Objectives

Define statistical models.

Understand how statistical models facilitate outcome prediction.

Understand how statistical models facilitate hypothesis testing.

Learning Objectives

Map study design features to statistical models.

Map scientific goals to research hypotheses.

Learning Objectives

Define between-ISU hypotheses.

Define within-ISU hypotheses.

Define between- by within-ISU hypotheses.

UNDERSTANDING STATISTICAL MODELS

Statistical models help summarize associations between outcomes and predictors

Accurate power and sample size calculation requires a basic understanding of statistical models.

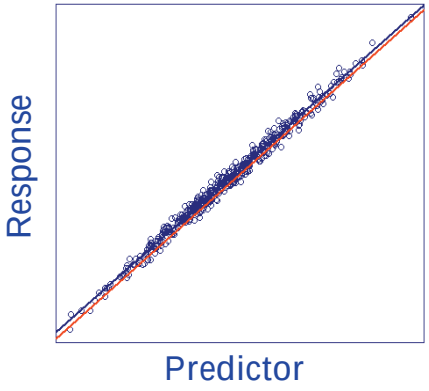
Models are used to predict outcomes and generate testable hypotheses

Statistical models calculate a predicted value of a response variable for each value or level of a predictor.

Predicted values inform hypotheses tests, which ask whether observed values are significantly different than we would expect.

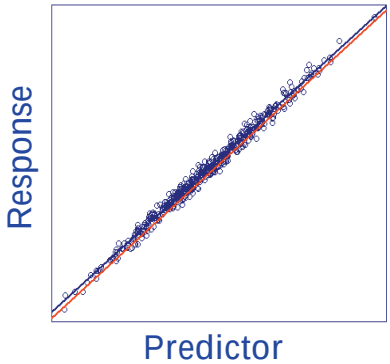
Plotting the predicted values produces a line of prediction

The line of prediction is the line that fits best when drawn through previously observed values in the response variable.



Hypothesis tests often evaluate whether the slope of the predicted line is equal to zero

The standardized slope of the predicted line represents the size of the effect of the predictor variable on the response variable.



Statisticians aim to select a 'good enough' statistical model to represent reality

Statistical models approximate reality using information available.

"Essentially, all models are wrong, but some are useful."—Box and Draper, 1987

Box et al., 1987
Chang et al., 2010

MAPPING SCIENTIFIC GOALS TO STATISTICAL MODELS

Different terminology is used for study designs than for statistical models

Understanding diverse terminology can ease interdisciplinary collaboration.

Understanding statistical terminology will prepare you to use GLIMMPSE

Recall, **GLIMMPSE** is a free, point-and-click, open-source software package for multilevel and longitudinal power and sample size calculations.

Terminology changes but the role of each study element remains the same

Differential Terminology by Context

Study design	Statistical model
Level	Cluster
Unit of observation	Single observed response
Between- ISU factors	Predictors or covariates
Within-ISU factors	Repeated measures

GLIMMPSE uses the statistical model terminology

Statistical Model Terminology

Term	Example
Clusters	Students in classrooms
Single observed response	Test scores measured 6 months after treatment
Predictor	Instructional method
Repeated measures	Test scores measured once per year for three years

Predictors are ISU factors that affect the outcome(s) of interest

Predictors may:

- Be interventional or observational
- Be fixed, or chosen by design
- Be random and unknown before the experiment is observed, but known without appreciable error once observed

Note: ISU = Independent sampling unit

Fixed predictors are chosen by design

Most often, the scientific goal is to understand the association between fixed **predictors** and **response variables.**

Fixed predictors are explanatory variables of interest defined in sample selection or randomization

Example	Predictor	Design Decision
Sample selection	Sex	Inclusion of 100 male participants and 100 female participants
Randomization	Treatment	Random assignment of 100 participants to the treatment condition and 100 to the control condition

Random predictors are quantifiable factors that vary by chance but are of experimental interest

Examples include:

- Blood pressure
- Alcohol intake

Note: ISU = Independent sampling unit

Random covariates are factors which impact the outcome but are not the primary explanatory variable

The value of a random covariate is not known before an experiment is observed.

Including a random covariate can help isolate the true effect of the predictor on the outcome(s) of interest by reducing the error variance.

A repeated measures model is used to describe a study in which a response variable is measured multiple times

Repeated measures may take many forms:

1. Longitudinal repeated measures
2. Multivariate repeated measures
3. Spatial repeated measures

Longitudinal repeated measures are within-independent sampling unit factors that vary with time

Example:

Distance walked by each participant, on each day, over one month.

Bray and Maxwell, 1985
Doncaster and Davey, 2007

Spatial repeated measures are within-independent sampling unit factors that vary over space

Example:

Repeated measurements of each digit of the hand for each participant in the study.

Bray and Maxwell, 1985
Doncaster and Davey, 2007

Multivariate repeated measures are observations measured multiple times for *multiple outcomes*

Important note:

MultiVARIATE ≠ MultiVARIABLE

MultiVARIATE ≠ MultiVARIABLE

Multivariate → Multiple outcomes

Multivariable → Multiple predictors or covariates

COMMON STATISTICAL MODELS

Three statistical models are often used for multilevel and longitudinal studies

Three models:

1. The univariate model
2. The multivariate model
3. The mixed model

We will not discuss how to fit the models for data analysis.

The univariate, multivariate and mixed models have a common structure

A model is a statement about a population:

response = prediction + error.

Data analysis uses a sample of data to find estimates of the parts of the model.

We observe the response values.

The univariate model is used to analyze experiments with a single outcome measured at only one point in time

Univariate models can only be applied to uncorrelated data, as found in studies with no clustering and a single outcome measure.

The multivariate model is used to analyze experiments with multiple outcomes or hypotheses

Multivariate outcomes may be multiple time points, multiple outcome variables measured at one time point, or multiple outcome variables measured at multiple time points.

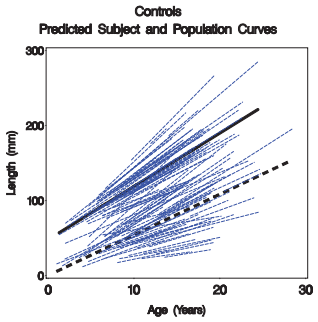
The mixed model is a more flexible model for analyzing experiments with multiple outcomes

The mixed model is more flexible because it makes fewer assumptions than the multivariate model.

We will discuss model assumptions in the next lecture.

In one form of the mixed model, a separate regression line is calculated for each independent sampling unit

In designs with the individual as the ISU, the mixed model allows for each person to have his or her own slope and intercept.



Note: ISU = Independent sampling unit

Both multivariate and mixed models can be used to analyze correlated data

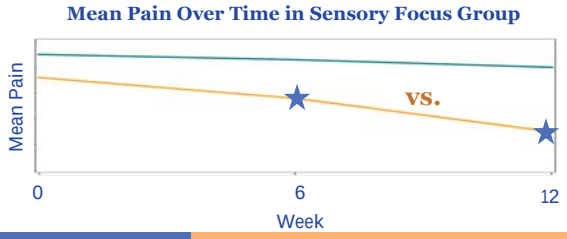
Multivariate or mixed models are employed in studies with:

- A single level and multiple outcomes
- Multiple levels and a single outcome
- Multiple levels and multiple outcomes.

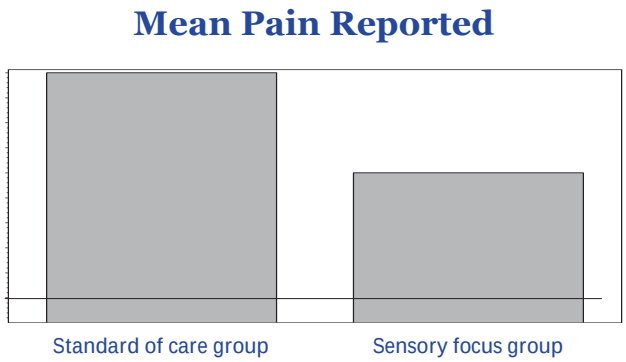
DEFINING STUDY HYPOTHESES

Within-ISU hypotheses are useful for tracking changes in an ISU over time

Within-independent sampling unit hypotheses compare measurements over time or space.

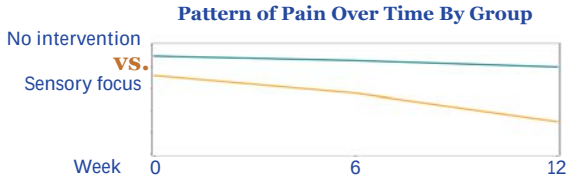


Between-ISU hypotheses help identify differences between study groups



Within by between-ISU hypotheses evaluate outcomes over multiple dimensions

Within by between-independent sampling unit hypotheses compare the patterns of an outcome over repeated measures across groups.



REVIEW OF LEARNING OBJECTIVES

Recall the longitudinal study of pain perceived after a root canal

Vignette

Researchers conducted a study to determine if dental patients who are instructed to use a sensory focus have a different pattern of long-term memory of pain than participants who did not.

Adapted from Logan et al., 1995

Vignette, continued

Participants were selected and randomly assigned to either intervention or no intervention.

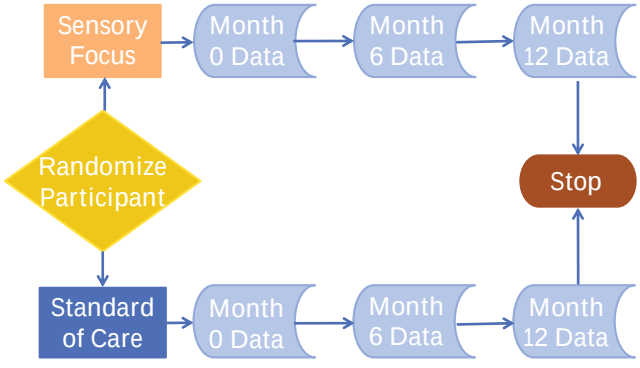
Adapted from Logan et al., 1995

Vignette, continued

Those in the intervention group listened to automated audio instructions to pay close attention only to the physical sensations in their mouth. Patients in the no intervention group listened to automated audio instruction on a neutral topic to control for media and attention effects.

Adapted from Logan et al., 1995

Observed pain after root canal was measured at 0, 6, and 12 months



Adapted from Logan et al., 1995

Researchers analyzed three distinct hypotheses

1 Between-independent sampling unit null hypothesis:

Mean pain experienced by patients in the sensory focus treatment group does not significantly differ from that experienced by patients in the standard of care group.

Adapted from Logan et al., 1995

Researchers analyzed three distinct hypotheses

2 Within-independent sampling unit null hypothesis:

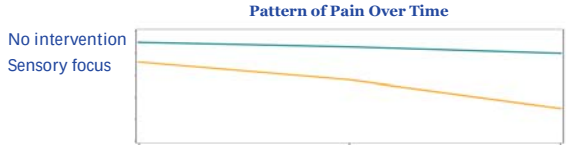
Mean pain experienced does not vary significantly over time.

Adapted from Logan et al., 1995

Researchers analyzed three distinct hypotheses

3 Between by within-independent sampling unit null hypothesis:

The **pattern** of pain perception over time is different for the sensory-focus group than for the control group.



Adapted from Logan et al., 1995

What was the independent sampling unit?

Answer: Patient

Adapted from Logan et al., 1995

What was the unit of observation?

Answer: Pain reported at a specific time

Adapted from Logan et al., 1995

What was the between-ISU factor?

Answer: Intervention group (standard of care or sensory focus)

Adapted from Logan et al., 1995

What was the between-ISU hypothesis?

There is no significant difference in the mean pain experienced by patients in the sensory focus group and the standard of care group.

What was the within-ISU factor?

Answer: Time

Adapted from Logan, Baron, and Kohout, 1995

What was the within-ISU hypothesis?

There is no significant difference in mean pain reported over time.

What was the between by within-ISU hypothesis?

The pattern of pain experienced over time does not differ between patients in the sensory focus group and those in the standard of care group.

Day 1 Agenda

9 am – 12 pm

1. Longitudinal Studies
2. Studies with Single Levels of Clustering
3. Studies with Multiple Levels of Clustering
4. Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Lunch lecture on “Ethics of Power and Sample Size”

1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes

Learning Objectives

Define the multivariate model.

Define the mixed model.

Define a residual.

Understand how model assumptions impact power analysis.

Learning Objectives

Understand the assumptions of the multivariate model.

Define reversible mixed models.

Understand the assumptions of reversible mixed models.

THREE COMMON STATISTICAL MODELS

Recall, three common statistical models are used for multilevel and longitudinal studies

Three common models:

1. The univariate model
2. The multivariate model
3. The mixed model

We will not discuss how to fit the models for data analysis.

The univariate model is used to analyze experiments with a single outcome measured at only one point in time

Univariate models can only be applied to uncorrelated data, as found in studies with no clustering and a single outcome measure.

The multivariate model is used to analyze experiments with multiple outcomes or hypotheses

Multivariate outcomes may be multiple time points, multiple outcome variables measured at one time point, or multiple outcome variables measured at multiple time points.

The mixed model is a more flexible model for analyzing experiments with multiple outcomes

The mixed model is more flexible because it makes fewer assumptions than the multivariate model.

We will discuss model assumptions in the next lecture.

UNDERSTANDING CORRELATION, RESIDUALS, AND VARIANCE

Three interrelated terms are used to describe model assumptions

1. Correlation
2. Residuals
3. Variance

Recall, correlation is one measure of the relationship between two variables

Correlation "indicates the strength and direction of the relationship between two random variables," (p. 127).

Rosner, 2010, p. 127

Correlation ranges from -1 to 1

Correlation

The sign of correlation indicates whether two variables are changing in the same direction

Negative correlation (between -1 and 0) indicates that two variables change in opposite directions.

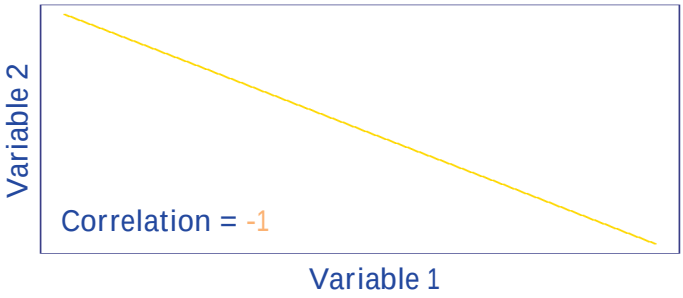
Positive correlation (between 0 and 1) indicates that variables change in the same direction.

Numbers further from zero represent stronger correlation

Correlation

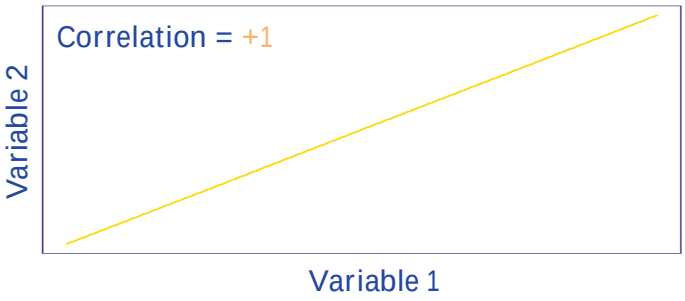
A correlation of -1 indicates that two variables change at constant rates in opposite directions

Perfect negative correlation



A correlation of +1 indicates that two variables change at constant rates in the same direction

Perfect positive correlation



A correlation of 0 indicates that two variables are unrelated

Variables with correlation equal to zero have rates and directions of change that are uncorrelated, or unrelated.

Variables that are independent always have correlation equal to zero

Mathematically speaking, variables with correlation equal to zero are not necessarily independent.

For the purposes of this class, we will focus on cases where correlation of zero indicates that two variables are independent.

The univariate, multivariate and mixed models have a common structure

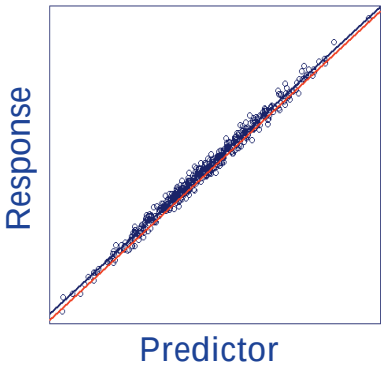
A model is a statement about a population:

response = prediction + error.

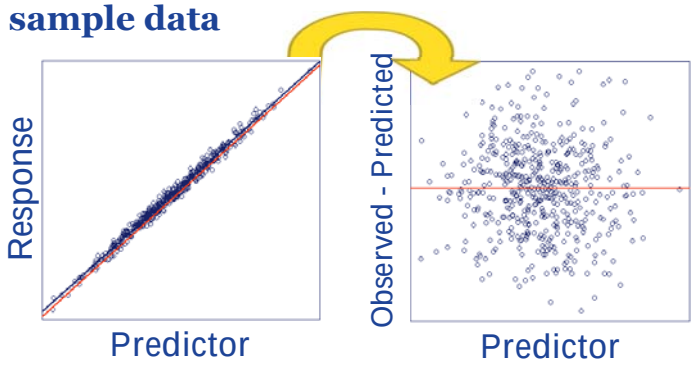
Data analysis uses a sample of data to find estimates of the parts of the model.

We observe the response values.

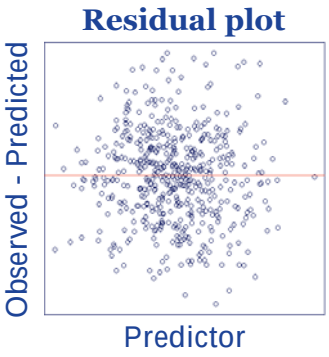
Recall, data analysis gives estimates called predicted values



Graphing differences between observed and predicted values shows how well the estimated model fits the sample data

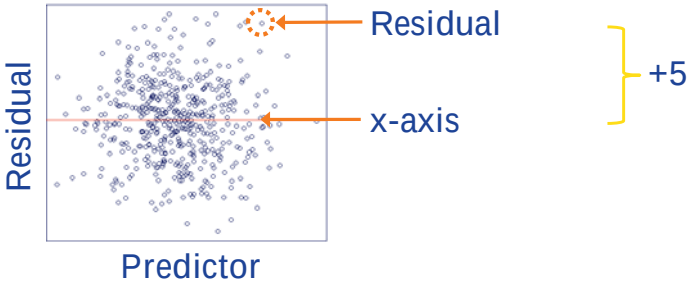


Graphing the differences gives a residual plot, frequently used to check model assumptions



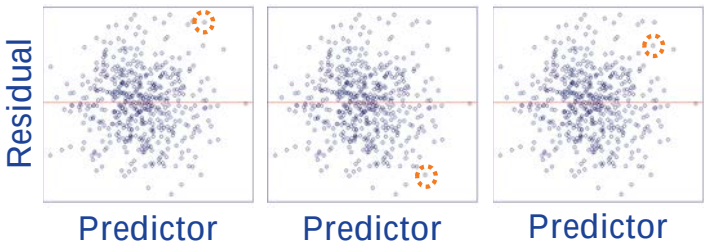
A residual is the difference between a single observed outcome value and the model predicted response value

Residual = observed value – predicted value



Residuals vary in magnitude and may be positive or negative

Residual value		
+5	-5	+3

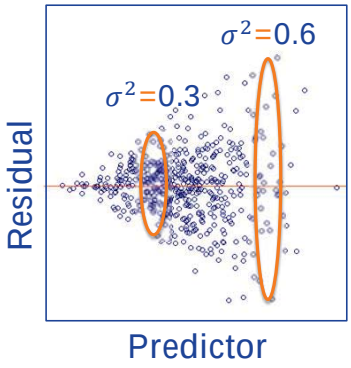


Variance is a summary measure used to describe the spread of many residuals

The variance of residuals summarizes the degree to which the values observed in an experiment differ from the values predicted by the estimated statistical model.

Variance is related to the range of values observed in an experiment

A greater range in observed values usually corresponds to greater variance in the residuals.



UNIVARIATE MODEL ASSUMPTIONS

Violating model assumptions may invalidate a data or power analysis

Studies discussed in this course employ one of three key statistical linear models:

1. The univariate model
2. The multivariate model
3. The mixed model

We will now discuss the assumptions of the univariate, multivariate and mixed linear models

The assumptions of the three models overlap, with some variation.

The univariate model has five assumptions

1. Independence
2. Homogeneity
3. Linearity
4. Existence
5. Normal (Gaussian) errors

1. INDEPENDENCE

The independence assumption requires that the values for each ISU be independent of values for every other ISU

If the independence assumption is met, then the correlation between independent sampling units (ISUs) is equal to zero.

The correlation between ISUs must be equal to zero, but observations within an ISU may still be correlated

Examples of independent units include unrelated people or separate schools located in different cities.

2. UNIVARIATE HOMOGENEITY

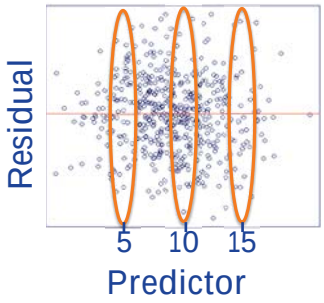
The homogeneity assumption requires that data values vary in predictable and consistent ways

The specific requirements of the homogeneity assumption depend on whether a study has one outcome, or many.

Muller and Fetterman, 2002

The univariate homogeneity assumption requires a constant variance of the errors for all values of the predictor

Residual variance can estimate error variance.



Predictor	Variance of residuals
5	0.6
10	0.6
15	0.6

Muller and Fetterman, 2002

Residual analysis is outside the scope of the course

Residuals are homogenous only for balanced designs.

In contrast to the variance of the hypothetical errors, the variance of residuals typically depends on the predictor value.

Data analysts account for that and use modified residuals (Muller and Fetterman, Chapter 3).

Muller and Fetterman, 2002

3. LINEARITY

The linearity assumption mandates a linear relationship between predictors and response variables

The linearity assumption means that the observed change in the response variable is approximately the same for each one unit increase in a predictor.

Muller and Fetterman, 2002

4. EXISTENCE

The existence assumption requires that observations are real values with a variance less than infinity

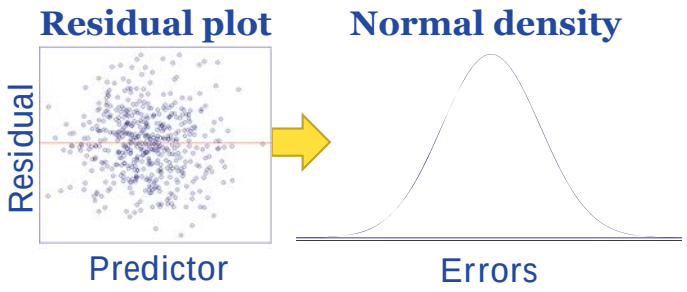
Any finite set of values has finite variance.

Concepts of infinity play a useful role in statistical and mathematical theories across the sciences, from astrophysics to zoology (black holes; catastrophe theory in psychology).

Muller and Fetterman, 2002

5. NORMALITY

The normality assumption requires that the errors follow a normal distribution (residuals estimate errors)

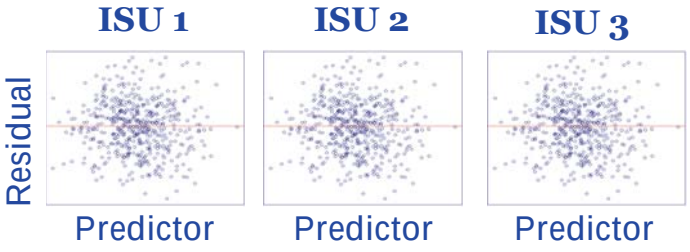


Multivariate model assumptions parallel the univariate model, with some elaborations

1. Independence
2. Multivariate homogeneity
3. Linearity
4. Existence
5. Multivariate normal (Gaussian) errors
6. Complete data

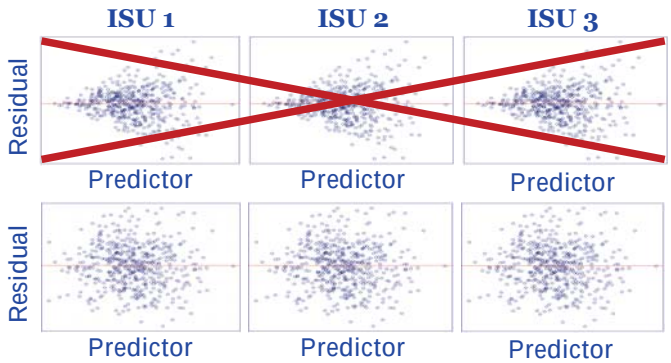
2. MULTIVARIATE HOMOGENEITY

The multivariate homogeneity assumption requires that errors (estimated by residuals) from each ISU have the same pattern of correlation and variance



Muller and Fetterman, 2002

With multivariate homogeneity assumption, the error variance does not depend on predictor values



6. COMPLETE DATA

The complete data assumption requires that none of the measurements be missing for any ISU

Some missing

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$
$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$
$y_{6,1}$	NA	$y_{6,3}$	$y_{6,4}$
$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$
$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$
NA	$y_{9,2}$	NA	$y_{9,4}$

None missing

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$
$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$
$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$
$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$

To fit a multivariate model with the following data, ISUs 6 and 9 must be deleted entirely

Original data

To fit a multivariate model with the following data, ISUs 6 and 9 must be

Data used

deleted entirely

To justify the deletion of missing observations for the multivariate model, data must be missing at random

Data may only be deleted if missingness is not determined by the outcome or predictors.

MIXED MODEL ASSUMPTIONS

The mixed model assumptions parallel the multivariate model with elaborations and one exception

1. Independence (but N=1 can be ok)
2. Multivariate homogeneity
   (but can allow groupwise heterogeneity)
3. Linearity (more general form)
4. Existence
5. Normal (Gaussian) errors
6. ~~Complete data~~

Unlike the multivariate model, the mixed model can accommodate missing data

Original data

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$
$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$
$y_{6,1}$	NA	$y_{6,3}$	$y_{6,4}$
$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$
$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$
NA	$y_{9,2}$	NA	$y_{9,4}$

Data used

$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$
$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$
$y_{6,1}$	NA	$y_{6,3}$	$y_{6,4}$
$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$
$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$
NA	$y_{9,2}$	NA	$y_{9,4}$

In general, mixed models have fewer restrictions than multivariate models

In addition to missing data, many mixed models can accommodate mistimed data and repeated covariates.

Kachman, 2015

In this course, we only consider 'reversible' mixed models

'Reversible' mixed models are models which are equivalent to multivariate models.

Mixed models are 'reversible' under the following three conditions

1. Each ISU contains the same number of units of observation.
2. Units of observation are measured at the same times, the same locations, or are measured for the same variables.
3. Predictor variables are only measured once.

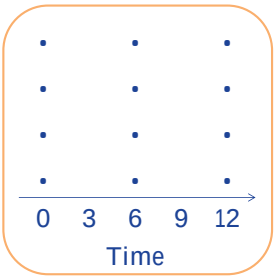
REVIEW OF LEARNING OBJECTIVES

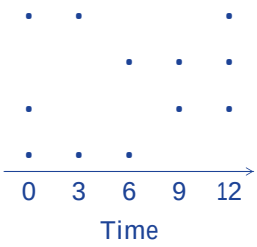
Which study design has an equal number of observations for each ISU?

	Observation		
ISU	1	2	3
1	.	.	.
2	.		.
3	.	.	
4	.		.
5	.	.	
6		.	.
7	.	.	

	Observation		
ISU	1	2	3
1	.	.	.
2	.	.	.
3	.	.	.
4	.	.	.
5	.	.	.
6	.	.	.
7	.	.	.

Which study design has equal timing of measurements?





Day 1 Agenda

9 am – 12 pm

1. Understanding Longitudinal Studies
2. Understanding Studies with Single Levels of Clustering
3. Understanding Studies with Multiple Levels of Clustering
4. Understanding Multilevel and Longitudinal Studies

Day 1 Agenda

12 – 1 pm

5. Lunch lecture on “Ethics of Power and Sample Size”

1 – 2 pm

Independent Research and Mentoring

Day 1 Agenda

2 – 5 pm

6. Within and Between Independent Sampling Unit Factors
7. Understanding the Hypothesis
8. Model Assumptions
9. Continuous, Binary, and Poisson Outcomes

Learning Objectives

Define continuous outcomes.

Define binary outcomes.

Define Poisson outcomes.

Learn to identify continuous, binary, and Poisson outcomes.

Continuous outcomes are variables with "a potentially infinite number of possible values along a continuum," (p.37)

Example: When do children first walk?

Participant	Age in months
1	11.0
2	11.5
3	8.0
4	12.3

Last, 2000, p.37

Binary outcomes are variables having only two possible values

Binary outcomes are often coded as 'dummy variables.'

Example: Did a student graduate from high school?

Outcome	Code
Graduate	1
Do not graduate	0

Last, 2000, p. 17

Poisson outcomes are variables in the form of counts

Example: How many car accidents occurred in Denver in the last two years?

Year	# Car Accidents
2015	320
2014	362

Last, 2000, p.136

REVIEW OF LEARNING OBJECTIVES

Identify the following outcomes as continuous, binary, or Poisson

- BMI
- Steps per day
- High blood pressure
- Number of heart attacks
- Drinks per week
- Weight
- Obesity

- Continuous
- Poisson
- Binary
- Poisson
- Poisson
- Continuous
- Binary

Day 2 Agenda

9 am – 12 pm

10. Choosing the Test
11. Correlation Structure
12. Power and Type I Error
13. Alignment of Power and Data Analysis

12 – 1 pm

Working lunch with faculty or optional break

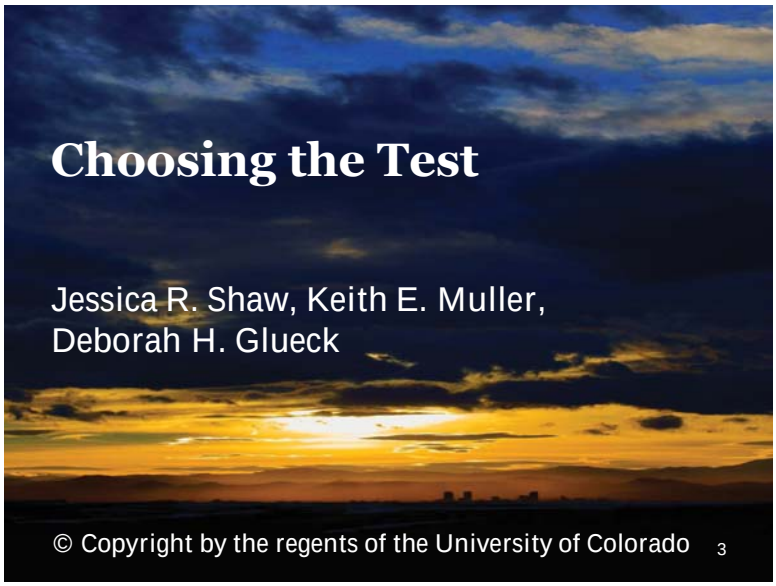
Day 2 Agenda

1 – 2 pm

Optional accessible walk and talk with faculty or break time

2 – 5 pm

14. Summarizing Your Design
15. Predicting Missing Data
16. Accounting for Missing Data and Dropout



Learning Objectives

Review models used for multilevel and longitudinal studies.

Understand which hypothesis test statistics are used for each model.

Learning Objectives

Understand criteria for comparing tests

Learn how to choose an appropriate test for data analysis, and aligned power and sample size analysis

REVIEW MODELS USED FOR MULTILEVEL AND LONGITUDINAL STUDIES

Recall, two statistical models are used most often for multilevel and longitudinal studies

Common models:

1. The multivariate model
2. The mixed model

We will not discuss how to fit the models for data analysis.

Reversible mixed models are mixed models that can also be expressed as multivariate models

The mixed model is more general.

All multivariate models are special cases of mixed models.

A mixed model meeting the conditions on the next slide is reversible.

Recall, GLIMMPSE calculates power and sample size for a mixed model under the following assumptions

1. Each ISU contains the same number of units of observation
2. Units of observation are measured at the same times, the same locations, or are measured for the same variables
3. Predictor variables are only measured once

SELECT THE HYPOTHESIS TEST STATISTIC USED FOR EACH MODEL

Tests for multivariate models can be divided into two groups

1. Univariate approach to repeated measures
2. Multivariate approach to repeated measures

Univariate approach to repeated measures tests assume commensurate data

Commensurate data have values observed on the same measurement scale.

Example:

Alcohol intake is measured 3, 6, and 9 months following treatment.

Multivariate approach to repeated measures tests allow the analysis of multivariate data

Examples:

Researchers measure intake of fats, proteins, and carbohydrates at weeks 0, 2, and 4.

Researchers measure height, weight and head circumference.

UNDERSTAND CRITERIA FOR COMPARING TESTS

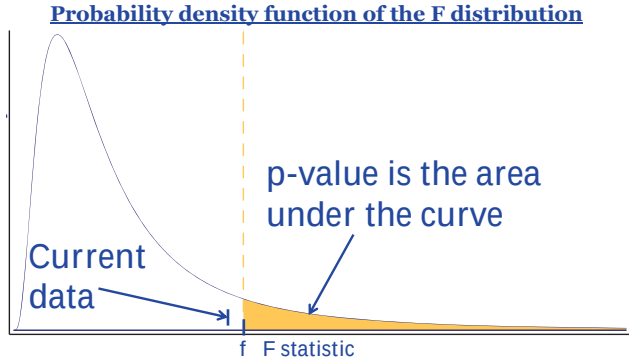
Recall the definition of the null hypothesis

A null hypothesis claims that the two events are not related.

Small p-values lead researchers to achieve their goal of rejecting the null hypothesis

A Type I error occurs when a statistical test mistakenly rejects the null hypothesis when in fact the null hypothesis is true.

Recall, a p-value is the probability of seeing data more extreme than the current data if the null is true



The claimed or nominal Type I error rate is chosen by investigators

The rate is a number between 0 and 1.

It represents the hoped for probability that a test will reject the null hypothesis if the null hypothesis is true.

The actual Type I error rate can be estimated by simulation

Consider running the experiment many times and recording the results.

$$\text{Empirical Type I error} = \frac{\#\text{ rejections}}{\#\text{ experimental replications}}$$

Some testing paradigms have accurate Type I error rates

A Type I error rate is accurate if the actual rate is equal to the claimed rate.

Example:

The claimed rate is 0.05 and the actual rate is 0.05.

Some testing paradigms do not achieve the claimed Type I error rate

The actual Type I error rate may be higher or lower than the claimed rate.

Example:

The claimed rate may be 0.05, but the actual rate may be 0.25.

Inflated Type I error rates decrease the chance of replication of results

An inflated Type I error rate occurs when a test rejects the null hypothesis (which is true) at a higher rate than the desired alpha level chosen by the investigator.

Why is it important whether a test has an accurate Type I error rate?

When Type I errors occur, scientists draw wrong conclusions more often than desired.

Type I errors are not just statistical problems

Mistaken conclusions can directly affect patient care.

Example:

A Type I error may result in the approval of a drug that is ineffective.

Inflated Type I error rates decrease the chance of replication of results

Failure to replicate results means that two studies draw different conclusions.

Conflicting recommendations make patients lose trust in science

Example:

A study tests the null hypothesis that eating cranberries does not increase risk for cancer. Researchers reject the null and conclude that eating cranberries causes cancer. A second study fails to reject the null hypothesis and concludes that eating cranberries does not cause cancer. The public is unsure of the safety of eating cranberries.

Researchers must accept that some experiments will incorrectly reject the null hypothesis

When the actual Type I error rate is bigger than the claimed one, the study designer is increasing the chance that the study cannot be replicated.

We only trust findings if several studies come to the same conclusion

Replicated findings are findings from subsequent studies that match the original findings.

Scientific knowledge grows from the steady accretion of replicated studies.

A uniformly most powerful test achieves the greatest power among all reasonable tests of the same size

Uniformly most powerful tests only exist for certain special cases.

Example:

A uniformly most powerful test exists for the general linear hypothesis in the general linear univariate model (e.g., testing means equal).

Under certain conditions, different tests coincide

Coincidence occurs when two tests provide exactly same p-values, lead to the same inferences, and produce the same power and sample size when designing a study.

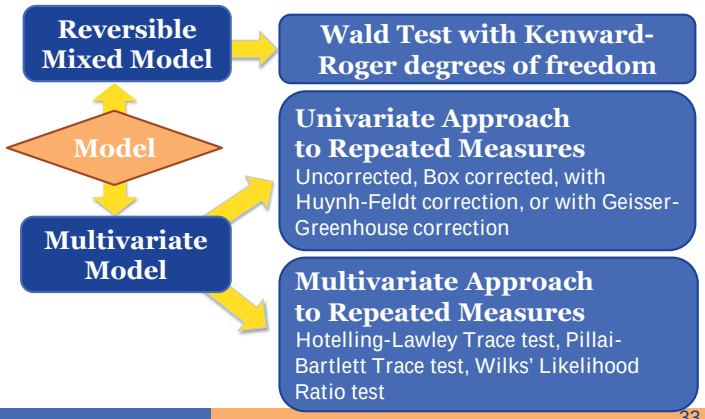
Coincidence can simplify the process of selecting a test.

We suggest two criteria for evaluating tests

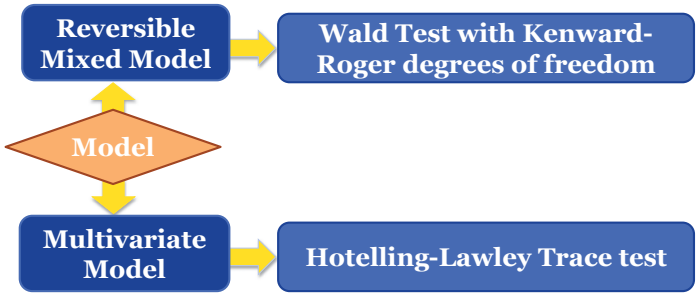
1. Check to see if the actual Type I error rate is close to claimed Type I error rate.
2. Choose the test with accurate Type I error rate that has the highest power.

CHOOSE AN APPROPRIATE TEST

GLIMMPSE incorporates a variety of tests, outlined in the map below



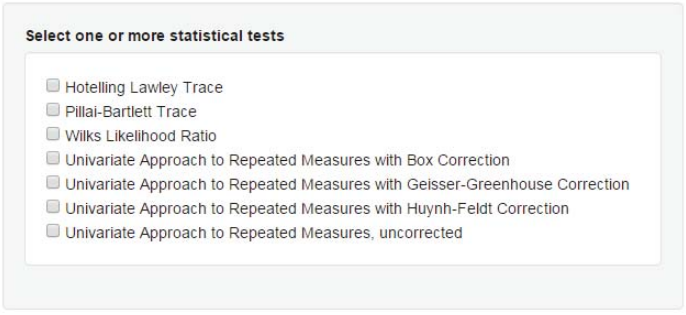
We will discuss why most testing decisions can be made using the following reduced map



Some tests may provide better power for certain designs/hypotheses

This lecture focuses on heuristics that apply to most, but not all, situations.

To calculate power for a multivariate model in GLIMMPSE, you may select from seven tests



If you are uncertain of which test to use, we recommend using the Hotelling-Lawley Trace test

The Hotelling-Lawley Trace test often coincides with a popular mixed model test.

# of repeated measures	# of groups of ISUs	Recommended test
≤ 2	≤ 2	Hotelling-Lawley Trace test
≤ 2	≥ 3	
≥ 3	≤ 2	
≥ 3	≥ 3	

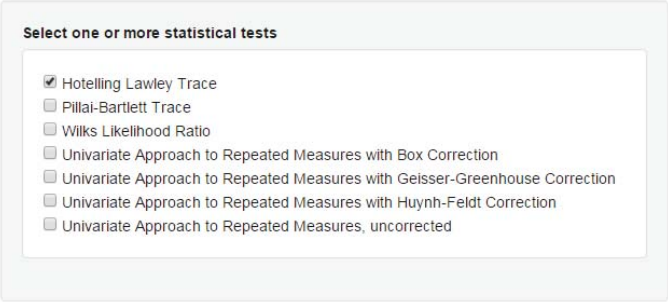
Recall, GLIMMPSE only calculates power for reversible mixed models

The Wald test with Kenward-Roger degrees of freedom for reversible mixed models approximately coincides with the Hotelling-Lawley test.

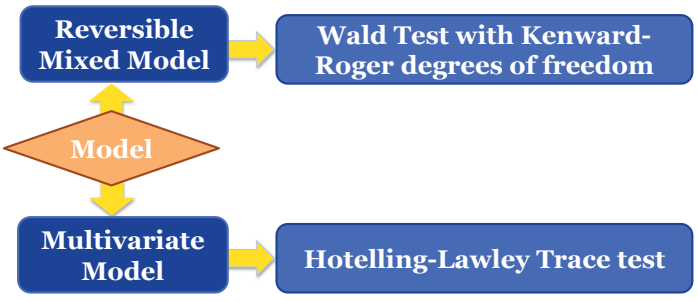
The Hotelling-Lawley trace test coincides with the Wald test under the following three conditions

1. No missing observations
2. Each ISU has same number of observations
3. No repeated predictors

To calculate power for a reversible mixed model in GLIMMPSE, select the Hotelling-Lawley Trace test



Accurate power and sample size can be calculated for most study designs with the Hotelling-Lawley Trace test



Particularly complicated designs may require a different test

Visit **www. SampleSizeShop.org** for resources on power analysis for complex designs.

Cheng, Edwards, Maldonado-Molina, Komro, and Muller (2010)

Muller, LaVange, Ramey, and Ramey (1992)

REVIEW OF LEARNING OBJECTIVES

Why can the Hotelling-Lawley Trace test be used for both multivariate and mixed models?

Coincidence means that the Wald test with Kenward-Roger degrees of freedom for the reversible mixed model coincides with the Hotelling-Lawley test in reversible mixed models.

Day 2 Agenda

9 am – 12 pm

10. Choosing the Test
11. Correlation Structure
12. Power and Type I Error
13. Alignment of Power and Data Analysis

12 – 1 pm

Working lunch with faculty or optional break

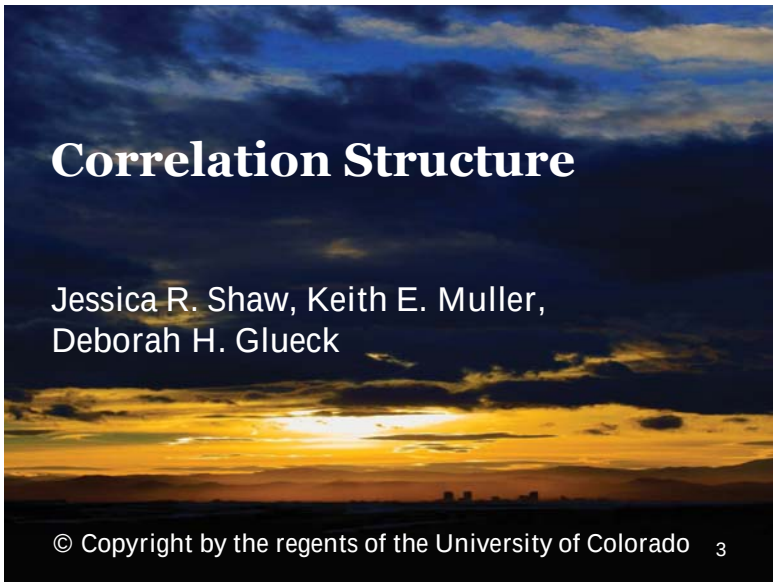
Day 2 Agenda

1 – 2 pm

Optional accessible walk and talk with faculty or break time

2 – 5 pm

14. Summarizing Your Design
15. Predicting Missing Data
16. Accounting for Missing Data and Dropout



Learning Objectives

Review correlation concepts.

Define variance.

Define correlation structure.

Understand how correlation structure influences power and data analysis.

Learning Objectives

Define compound symmetry.

Understand how clusters of observations induce compound symmetric correlation structures.

Learning Objectives

Understand that multivariate study designs can induce a variety of different correlation structures.

Understand how multivariate designs influence correlation structure.

For accurate power analysis, the expected correlation structure of observations must be summarized

Correlation structure summarizes the correlation between pairs of observations.

We will discuss correlation structures induced by multilevel and longitudinal studies

Clustering and repeated measures both influence correlation structure.

Specification of the correlation structure is critical to achieve correct power and sample size calculation

Correct correlation structure

↓

Accurate sample size

Incorrect correlation structure

↓

Inaccurate sample size

↓

Ethical dilemmas

REVIEW: CORRELATION CONCEPTS

Recall, correlation is a measure of the relationship between two variables

Correlation "indicates the strength and direction of the relationship between two random variables," (p. 127).

Rosner, 2010, p. 127

Correlation ranges from -1 to 1

Correlation

Numbers further from zero represent stronger correlation

Correlation

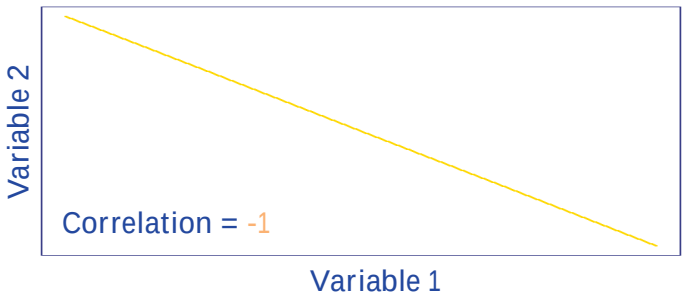
The sign of correlation indicates whether two variables are changing in the same direction

Negative correlation (between -1 and 0) indicates that two variables change in opposite directions.

Positive correlation (between 0 and 1) indicates that variables change in the same direction.

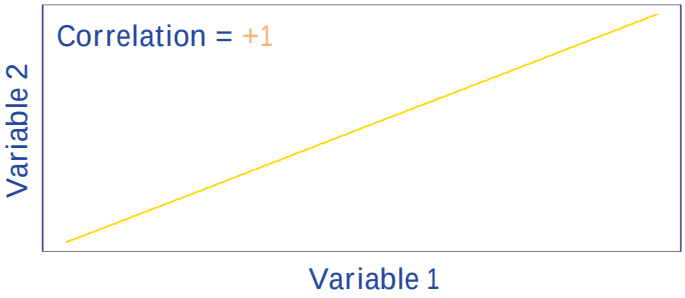
A correlation of -1 indicates that two variables change at constant rates in opposite directions

Perfect negative correlation



A correlation of +1 indicates that two variables change at constant rates in the same direction

Perfect negative correlation



A correlation of 0 indicates that two variables are unrelated

Variables with correlation equal to zero have rates and directions of change that are uncorrelated, or unrelated.

Variables that are independent always have correlation equal to zero

Mathematically speaking, variables with correlation equal to zero are *not necessarily* independent.

For the purposes of this class, we will focus on cases where correlation of zero indicates that two variables are independent.

A correlation matrix is a concise way of summarizing patterns of correlation

Correlation

Variable	1	2	3	4
1	1.0	0.8	0.5	0.1
2	0.8	1.0	0.1	0.5
3	0.5	0.1	1.0	0.8
4	0.1	0.5	0.8	1.0

The correlation between a variable and itself is always +1.0

Correlation

Variable	1	2	3	4
1	1.0	0.8	0.5	0.1
2	0.8	1.0	0.1	0.5
3	0.5	0.1	1.0	0.8
4	0.1	0.5	0.8	1.0

Glueck, D.H.

CORRELATION STRUCTURE OF CLUSTERS

Clustering within a level in a study induces a correlation structure called compound symmetry

Recall that levels, sometimes referred to as groups or clusters, share similarities which induce correlation.

Compound symmetric correlation structures exhibit two notable features

1. All independent sampling units have the same variance.
2. The correlation between any two independent sampling units is the same, no matter which two are chosen.

Compound symmetric correlation structures exhibit two symmetries

Correlation

ISU	1	2	3	4
1	1.0	0.5	0.5	0.5
2	0.5	1.0	0.5	0.5
3	0.5	0.5	1.0	0.5
4	0.5	0.5	0.5	1.0

Variance

ISU	σ^2
1	0.3
2	0.3
3	0.3
4	0.3

CORRELATION STRUCTURE OF REPEATED MEASURES

The pattern of correlation between repeated measures is usually different than the pattern within clusters

Observations in longitudinal studies typically do not follow a compound symmetric pattern of correlation.

Longitudinal studies induce correlation between measurements

Measurements from the same person taken at two or more times will be correlated.

Longitudinal measures are a special case of repeated measures.

Rosner, 2010

The variability of observations may change over the course of repeated measures designs

Example:

The variance in the outcome observed at measurement one may be different from the variance observed at measurement two.

Similarly, correlation may change over the course of repeated measures designs

Example:

The correlation between measurements one and two typically will be higher than the correlation between measurements one and four.

In studies with spatial repeated measures, a variable is measured repeatedly over space

Spatial repeated measures designs have complex correlation structures.

Previous data should be used to model the correlation in spatial repeated measures designs

Correlation may decrease as distance increases.

Multivariate repeated measures studies require similarly complex correlation structures

In multivariate designs, correlation is the result of multiple response variables being measured for each independent sampling unit.

Correlation structures for multivariate repeated measures studies should also be based on previous data

The inter-relationship of multiple response variables may be complex.

Modeling correlation structures is made easier by existing technology

Software simplifies the implementation of complex models.

A user specifies a model for each level and then the software combines levels to generate complete model.

REVIEW OF LEARNING OBJECTIVES

True or false? Nature can determine covariance structures

- TRUE
- FALSE

True or false? It is sometimes okay to ignore correlation within clusters

- TRUE
- FALSE

True or false? Organizational structures can affect correlation

- TRUE
- FALSE

Day 2 Agenda

9 am – 12 pm

10. Choosing the Test
11. Correlation Structure
12. Power and Type I Error
13. Alignment of Power and Data Analysis

12 – 1 pm

Working lunch with faculty or optional break

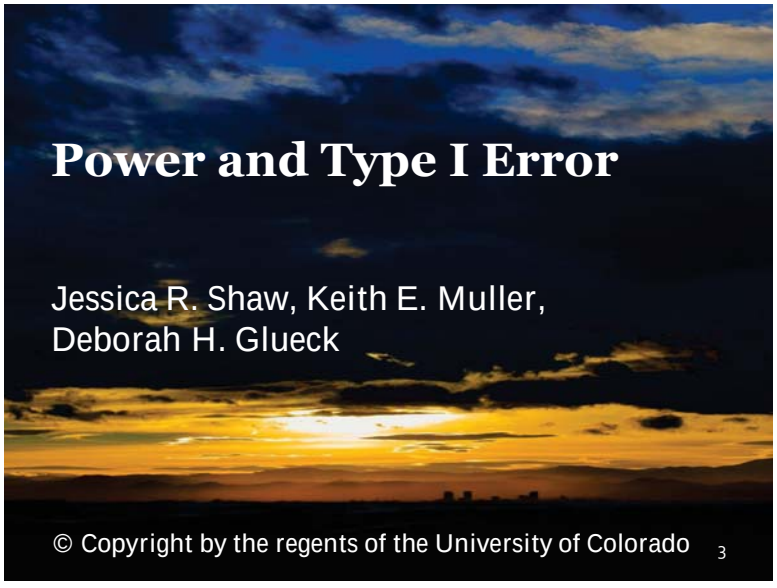
Day 2 Agenda

1 – 2 pm

Optional accessible walk and talk with faculty or break time

2 – 5 pm

14. Summarizing Your Design
15. Predicting Missing Data
16. Accounting for Missing Data and Dropout



Learning Objectives

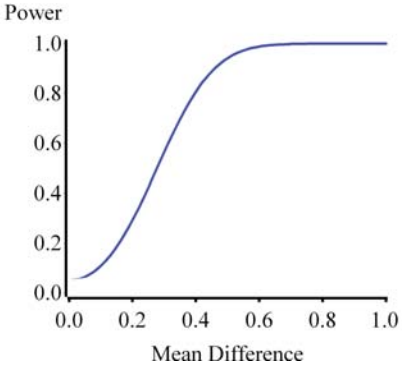
Define power.

Define Type I error.

Understand how to specify Type I error in GLIMMPSE.

DEFINITIONS

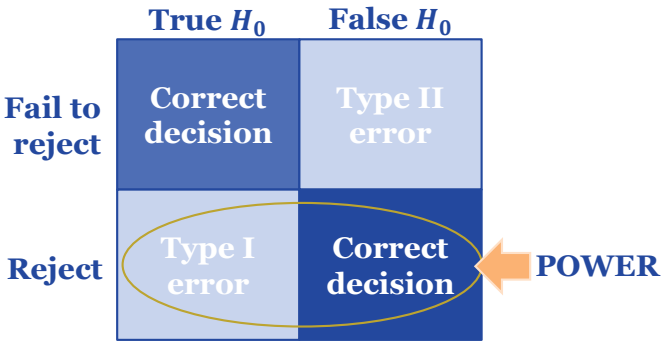
We use power curves to discuss the relationship between power and Type I error



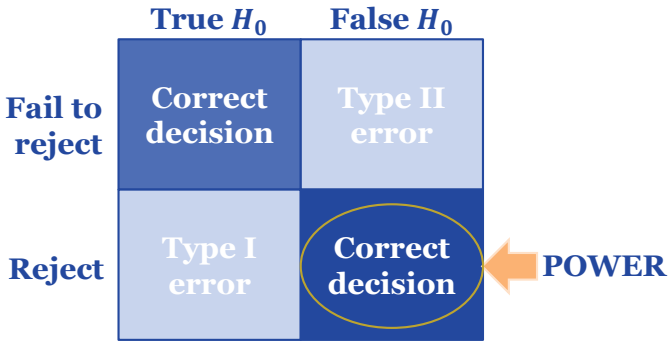
Type I error is the probability of an incorrect rejection of the null hypothesis, H_0

	True H_0	False H_0
Fail to reject	Correct decision	Type II error
Reject	Type I error	Correct decision

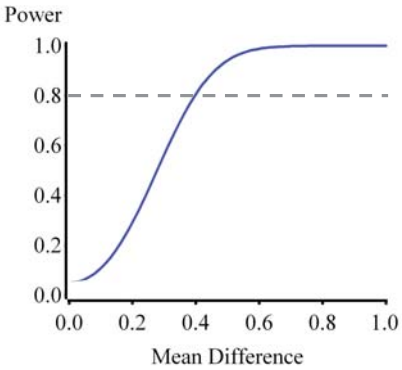
Power is the probability of rejecting the null hypothesis, H_0



Non-statisticians often define power as the probability of a correct rejection

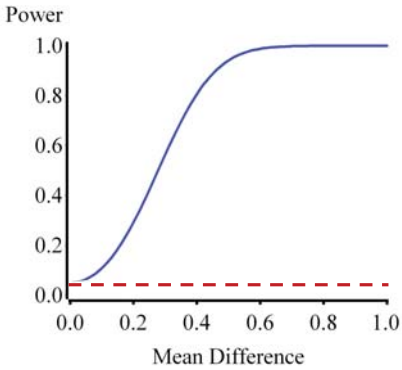


Power can be written as a number between 0 and 1

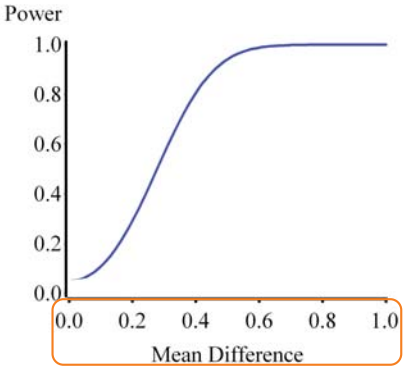


0.80 ⇔ 80% rejects

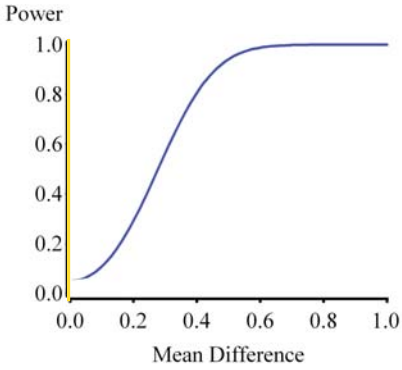
Power never falls below the Type I error rate, alpha



Null and alternative hypotheses are described in terms of mean difference

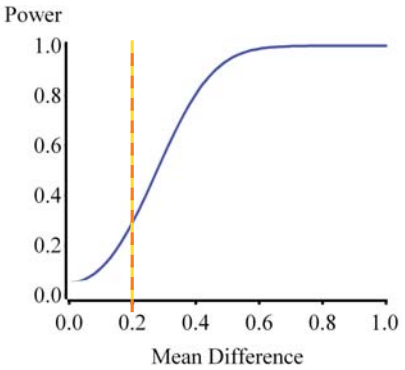


A mean difference of zero represents the null hypothesis being true



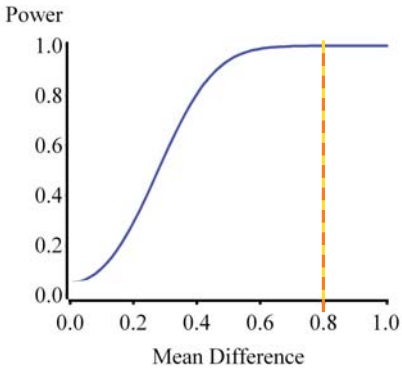
	True H_0	False H_0
Fail to reject	Correct decision	Type II error
Reject	Type I error	Correct decision

A mean difference not equal to zero represents an alternative hypothesis



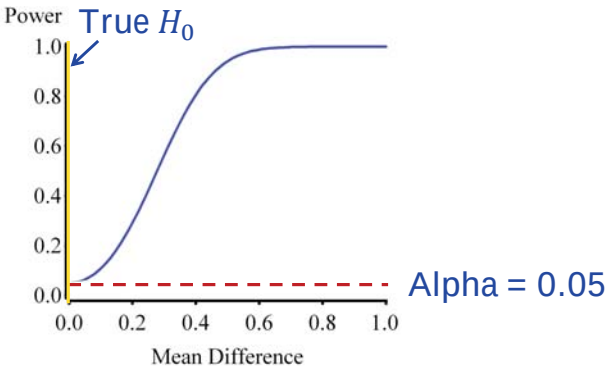
	True H_0	False H_0
Fail to reject	Correct decision	Type II error
Reject	Type I error	Correct decision

The x axis of a power curve represents many possible alternative hypotheses

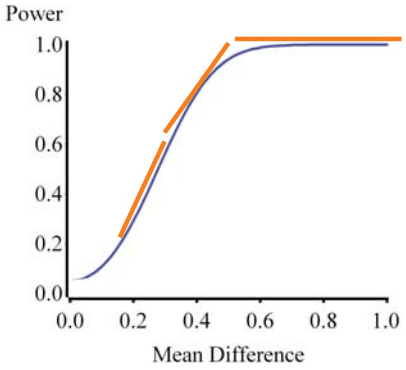


	True H_0	False H_0
Fail to reject	Correct decision	Type II error
Reject	Type I error	Correct decision

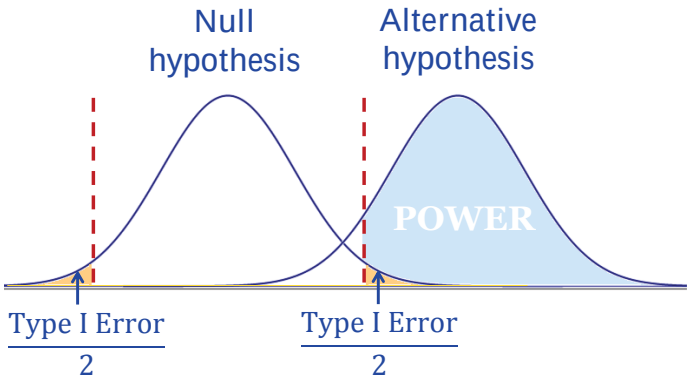
Power is equal to the Type I error rate, alpha, when the null hypothesis is true



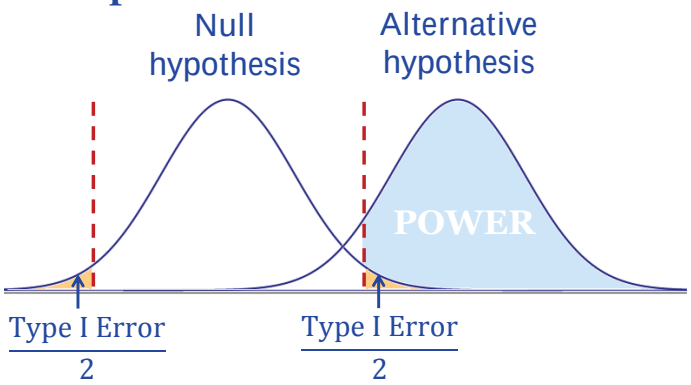
The power curve eventually flattens out as the mean difference gets larger



As the Type I error rate decreases, power also decreases

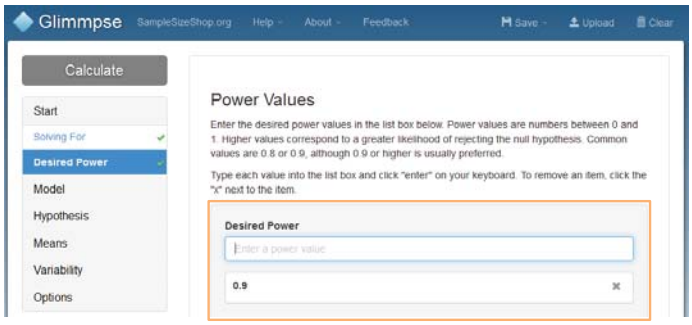


A lower Type I error requires a greater sample size to achieve the same power

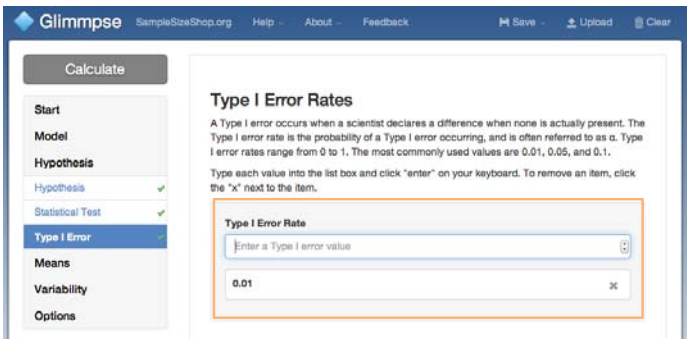


USING GLIMMPSE

If solving for required sample size, specify the power you would like to achieve in the Start menu



After summarizing your design in the Model section, specify the Type I error rate and press 'Enter'



REVIEW OF LEARNING OBJECTIVES

Define Type I error

Type I error is the probability of rejection of the null hypothesis, when the null hypothesis is in fact true.

Define power

Power is the probability of rejecting the null hypothesis.

Which one of the following is a plausible power value?

a. 0.10
b. 1.2
c. -3.6
d. 500%

Which one of the following is the most desirable power?

a. 0.10
b. 0.90
c. 0.40
d. 0.60

Day 2 Agenda

9 am – 12 pm

10. Choosing the Test
11. Correlation Structure
12. Power and Type I Error
13. Alignment of Power and Data Analysis

12 – 1 pm

Working lunch with faculty or optional break

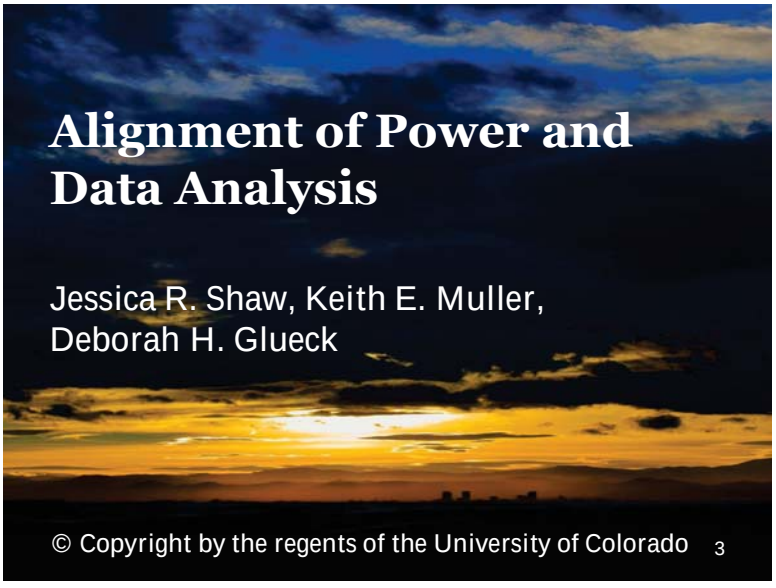
Day 2 Agenda

1 – 2 pm

Optional accessible walk and talk with faculty or break time

2 – 5 pm

14. Summarizing Your Design
15. Predicting Missing Data
16. Accounting for Missing Data and Dropout



Learning objectives

Define alignment in the context of power analysis.

Learn what features must be aligned to find the sample size correctly.

Understand what problems can happen if power and data analysis are not aligned.

Alignment means that the power analysis and data analysis describe the same experiment

The power analysis and the planned data analysis must share the same data collection, scientific goal, hypothesis, analysis, statistical model, statistical test, and Type I error.

We will discuss how to conduct an aligned power analysis for the longitudinal study of pain recall

Vignette

Researchers conducted a study to determine if participants who were instructed to use a sensory focus had a different pattern of long-term memory of pain than participants who did not use a sensory focus.

Logan, Baron, and Kohout, 1995

Vignette, continued

Participants were selected and randomly assigned to either intervention, the sensory focus, or no intervention.

Logan, Baron, and Kohout, 1995

Vignette, continued

Those in the intervention group listened to automated audio instructions to pay close attention only to the physical sensations in their mouth. Patients in the no intervention group listened to automated audio instruction on a neutral topic to control for media and attention effects.

Logan, Baron, and Kohout, 1995

Recall, this study analyzed the impact of treatment on the memory of pain

Null hypothesis:

Participants receiving the sensory focus treatment experience the same pattern of pain over time as participants receiving the standard of care treatment.

Logan, Baron, and Kohout, 1995

The scientists wished to compare the experience of the two groups

What was measured: Patients were asked to rate their memories of pain at baseline, and at 3 and 6 months.

Scientific goal: The goal of the study was to compare the intervention group to the standard of care group in terms of their pattern of memory of pain over time.

The scientists wished to compare the experience of the two groups

Null hypothesis: No difference in pattern of pain over time between groups, or no group by time interaction.

Planned analysis: The scientists planned a repeated measures analysis of variance.

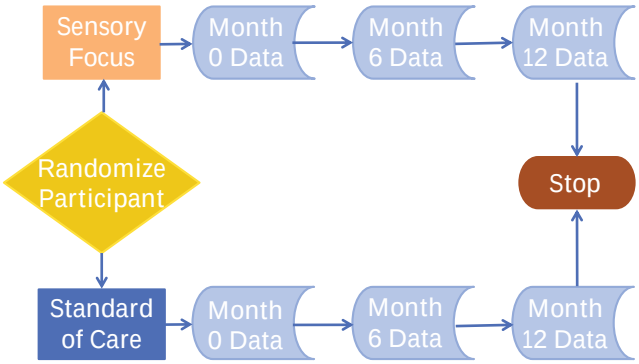
The scientists planned a repeated measures analysis of variance

Model: The scientists planned to fit a general linear multivariate model, with the repeated measurements of memory of pain as the outcome, and indicator variables for group as the predictors.

The scientists planned a repeated measures analysis of variance

Statistical test: The scientists planned to use a Hotelling-Lawley test to test interaction between group and time with an 0.05 Type I error rate.

Observed pain after root canal was measured at 0, 6, and 12 months



Logan, Baron, and Kohout, 1995

Power analysis must be aligned with the planned data analysis

Study Component	Power Analysis	Data Analysis
Type I error	0.01	0.01
Hypothesis	Time x treatment interaction using repeated measures ANOVA and Hotelling-Lawley trace	Time x treatment interaction using repeated measures ANOVA and Hotelling-Lawley trace

CONSEQUENCES OF MISALIGNMENT

Power and sample size analysis can be incorrect for a variety of reasons

Three common reasons are:

1. Wrong criterion
2. Wrong correlation structure
3. Wrong hypothesis

Some authors incorrectly suggest using confidence interval width as a criterion to select sample size

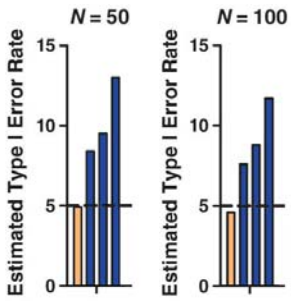
Power and confidence interval width are different criteria.

Jiroutek et al., 2003, give an example where sample size could be four times too large (106 vs. 24), or roughly 3 times too small, if one uses a width rather than a power criteria.

Fitting the wrong correlation for longitudinal studies can similarly inflate the Type I error rate

Example: Researcher fit the model assuming compound symmetry.

- True structure is compound symmetric.
- True structure is not compound symmetric.



Adapted from Gurka *et al.*, 2011

Power depends on the test

*Table 11. Approximate Power (×100) of W for Alternate Designs and Contrasts, N = 100, **B** and **Σ** based on IHDP Control Data*

Design (analysis, **C** matrix)	**B** multiplier	Time contrast, **U** matrix: Last time	Linear only	Linear and quadratic
Balanced Four Intervals, Subjects at IHDP Means (Four Group ANOVA, Overall)	.5	83	43	36
	1.0	>99	97	95
	2.0	>99	>99	>99
IHDP Ratios of Four Intervals, Subjects at IHDP Means (Four Group ANOVA, Overall)	.5	82	40	33
	1.0	>99	96	92
	2.0	>99	>99	>99
Balanced Four Intervals, IHDP Spread Within (Cubic Polynomial, All Slopes)	.5	83	35	26
	1.0	>99	93	84
	2.0	>99	>99	>99
IHDP Ratios of Four Intervals, IHDP Spread Within (Cubic Polynomial, All Slopes)	.5	51	18	14
	1.0	99	62	49
	2.0	>99	>99	>99

Muller et al., 1992

REVIEW OF LEARNING OBJECTIVES

List the components of studies that must match to achieve alignment

- Same experimental design
- Same data collection
- Same scientific goal
- Same hypothesis
- Same analysis
- Same model
- Same statistical test
- Same Type I error

Day 2 Agenda

9 am – 12 pm

10. Choosing the Test
11. Correlation Structure
12. Power and Type I Error
13. Alignment of Power and Data Analysis

12 – 1 pm

Working lunch with faculty or optional break

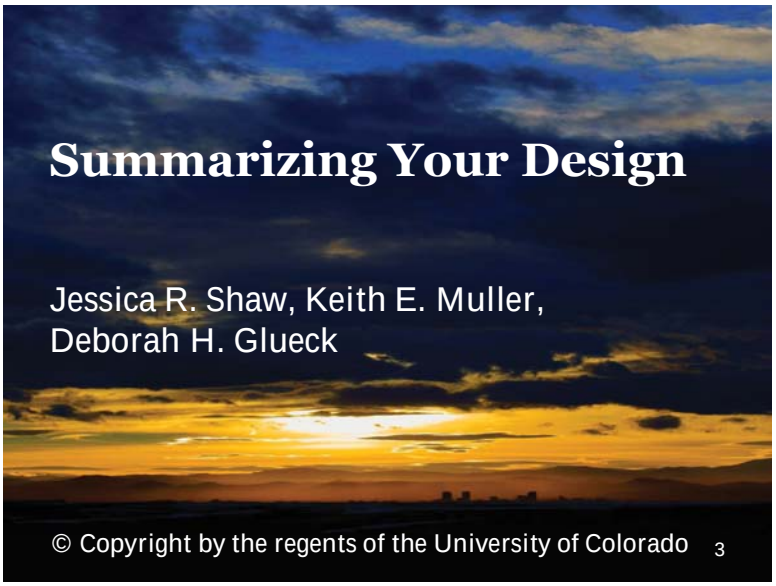
Day 2 Agenda

1 – 2 pm

Optional accessible walk and talk with faculty or break time

2 – 5 pm

14. Summarizing Your Design
15. Predicting Missing Data
16. Accounting for Missing Data and Dropout



Learning Objectives

Introduce GLIMMPSE software.

Understand how to use GLIMMPSE to summarize a study design.

This course provides instruction in the use of GLIMMPSE for power and sample size calculation

Recall, GLIMMPSE is a free, point-and-click open-source software package that enables power and sample size calculation for multilevel and longitudinal studies.

GLIMMPSE is available in a variety of formats

Available via standard web browsers and mobile devices at

GLIMMPSE.SampleSizeShop.org

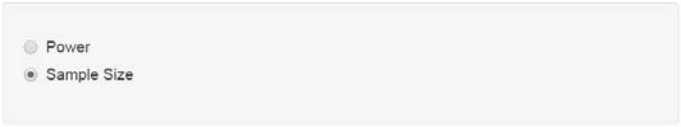
USING GLIMMPSE

First, visit the GLIMMPSE home page and select 'Guided Study Design'

GLIMMPSE.SampleSizeShop.org

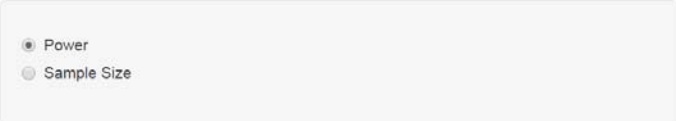
From the 'Start' menu, indicate what you are interested in calculating

You may calculate the sample size required to attain a desired level of power



This option is useful in planning studies with few restrictions on recruitment.

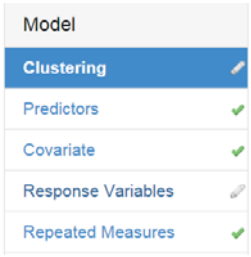
Alternatively, you may calculate the power that can be achieved given known limitations on recruitment



Remember, power and sample size calculations must be done *before* a study is conducted.

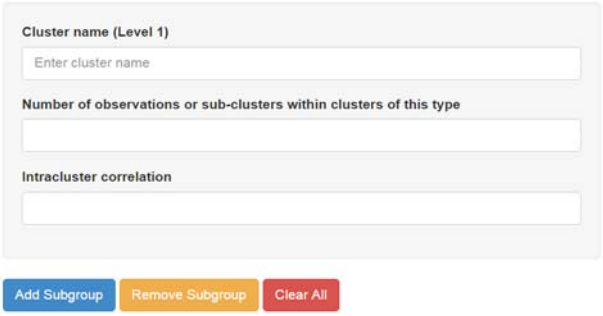
GLIMMPSE will then require information about five characteristics of the study design

1. Clustering
2. Predictors
3. A covariate
4. Response variables
5. Repeated measures



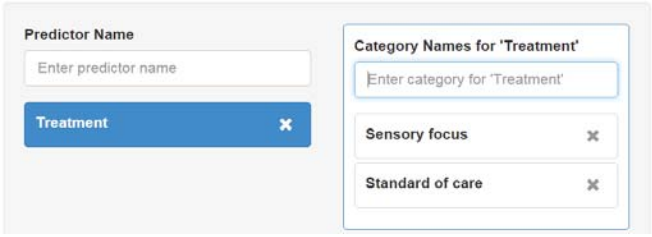
1. Clustering

Indicate whether the units of observation in your study are grouped into clusters.



2. Predictors

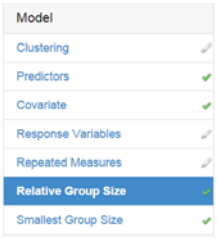
Type in the category values of each predictor variable.
For example:



Press Enter.

2. Predictors

Indicate whether the groups assigned to each level of the predictor variable are of equal size.



Example:

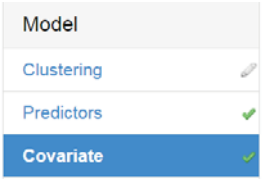
Treatment	Relative Group Size
Sensory focus	1
Standard of care	3

2. Predictors

If solving for *power*, indicate the number of participants or groups of participants that will be recruited in the smallest subgroup.

3. A covariate

Under the 'Model' menu, indicate if you would like to control for a single normally distributed covariate.



3. A covariate

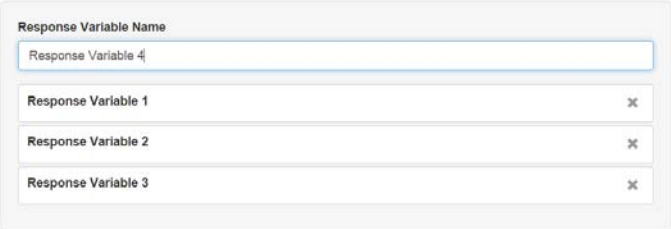
GLIMMPSE cannot currently calculate power for multiple normally or non-normally distributed covariates.

Glueck and Muller (2003) described an approximate method of adjustment.

Ignoring useful covariates gives conservative power.

4. Response variables

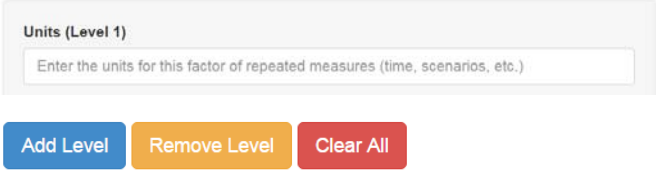
GLIMMPSE can accommodate multiple response variables in a single analysis



You may prefer multiple univariate analyses.

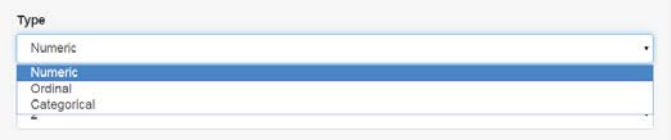
5. Repeated measures

Specify up to three levels of repeated measures.



5. Repeated measures

Classify the type of repeated measures.



Type	Example
Numeric	BMI over time
Ordinal	Hardness of cement
Categorical	Right arm, left arm

5. Repeated measures

Indicate number and spacing of repeated measurements.

Example:

Three measurements at months 3, 6, and 9.

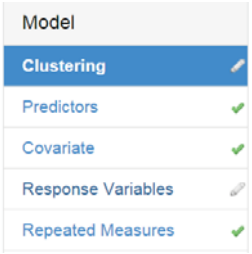
PRACTICE

For practice, we will summarize the designs of our example studies

- Single level
- Longitudinal
- Multilevel
- Multilevel longitudinal

For each study, we will identify the following characteristics required by GLIMMPSE

1. Clustering
2. Predictors
3. A covariate
4. Response variables
5. Repeated measures



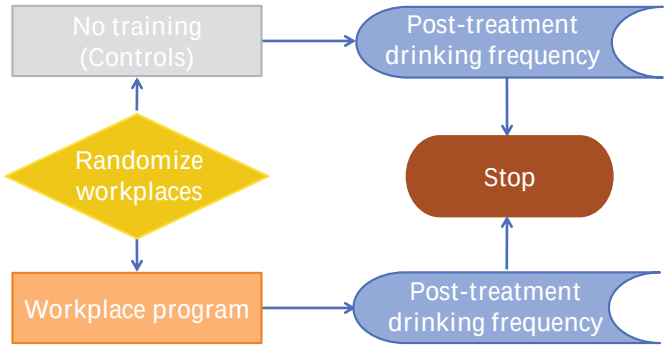
PRACTICE: SINGLE LEVEL STUDY

Vignette

A single level study examined the efficacy of a workplace training program to reduce alcohol consumption. Researchers randomized workplaces to two treatment groups. The first treatment included a workplace training program and the second treatment included no training. Post-treatment drinking frequency for each worker was measured as the outcome of interest.

Adapted from Reynolds et al., 2015

Single level study

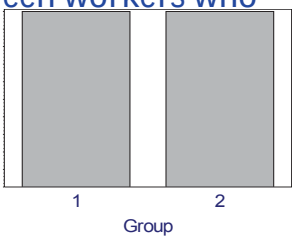


Adapted from Reynolds et al., 2015

Single level study

Null hypothesis:

There is *no difference* in post-treatment drinking frequency between workers who receive no training and workers who receive the workplace program.



Adapted from Reynolds et al., 2015

Single level study

Level(s): One level consisting of workers in a workplace

Independent sampling unit (ISU): Workplace

Unit of observation: Drinking frequency for each employee after treatment

Adapted from Reynolds et al., 2015

Single level study

Between-independent sampling unit factor:

Intervention (standard of care or workplace drinking program)

Within-independent sampling unit factor:

Worker within a workplace

Adapted from Reynolds et al., 2015

Statistical modelling terminology: Single level study

Clustering: Workers within the workplace

Predictors: Exposure to workplace substance abuse prevention program

Covariates: None

Adapted from Reynolds et al., 2015

Statistical modelling terminology: Single level study

Response variable: Frequency of drinking alcohol

Repeated Measures: None

Adapted from Reynolds et al., 2015

LONGITUDINAL STUDY

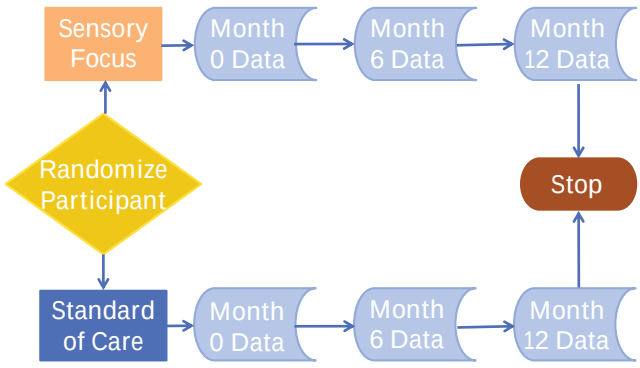
Vignette

Researchers conducted a study to determine if dental patients who are instructed to use a sensory focus have a different pattern of long-term memory of pain than participants who did not. Participants were selected and randomly assigned to either intervention or no intervention.

Vignette

Those in the intervention group listened to automated audio instructions to pay close attention only to the physical sensations in their mouth. Patients in the no intervention group listened to automated audio instruction on a neutral topic to control for media and attention effects.

Longitudinal study



Longitudinal study

Null hypothesis:

Participants receiving the sensory focus treatment experience the same pattern of pain over time as participants receiving the standard of care treatment.

Longitudinal study

Independent sampling unit (ISU): Patient

Unit of observation: Patient perceived pain at a specific time

Within-ISU factor: Time

Between-ISU factor: Treatment assignment to standard of care or sensory focus

Logan, Baron and Kohout, 1995

Statistical modelling terminology: Longitudinal study

Response variables: Perceived pain

Repeated measures: Three equally spaced measurements at months 0, 6 and 12

Logan, Baron and Kohout, 1995

Statistical modelling terminology: Longitudinal study

Clustering: None

Predictors: Treatment group (sensory focus or standard of care)

Covariate: None

Logan, Baron and Kohout, 1995

PRACTICE: MULTILEVEL STUDY

Vignette

Researchers conducted a cluster randomized control trial to evaluate the effectiveness of a web-based literacy intervention called ABRACADABRA (ABRA). The study included 24 classrooms within 12 elementary schools within a single school district. Researchers assumed that schools within the district were under local control and were therefore independent.

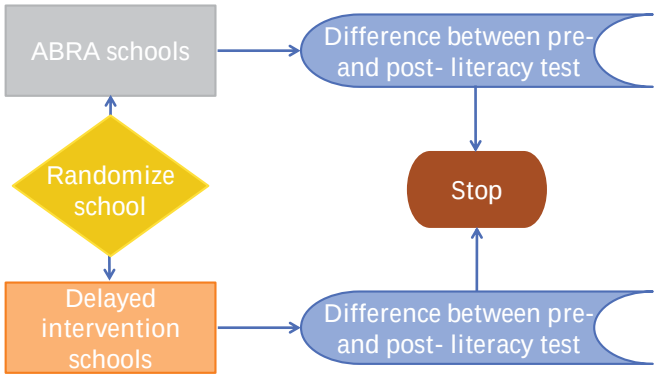
Adapted from Piquette et al., 2014

Vignette

Schools were randomized into the intervention group or the control group. Change in literacy was evaluated using pre- and post-tests to determine whether ABRA technology significantly improved literacy in elementary school children.

Adapted from Piquette et al., 2014

Multilevel study



Adapted from Piquette et al., 2014

Multilevel study

Null hypothesis:

There is no significant difference in literacy between elementary students in the intervention group and those in the control group.

Adapted from Piquette et al., 2014

Multilevel study

Levels of correlation: Classrooms within each school, students within each classroom

Independent sampling unit (ISU): School

Unit of observation: Difference between pre- and post-test performance

Adapted from Piquette et al., 2014

Multilevel study

Between-ISU factor: Randomization group

Within-ISU factors: classroom, child

Adapted from Piquette et al., 2014

Statistical modelling terminology: Multilevel study

Clustering: Classrooms within each school, students within each classroom

Predictors: Study condition (ABRACADABRA or delayed intervention)

Covariates: None

Adapted from Piquette et al., 2014

Statistical modelling terminology: Multilevel study

Response variable: Difference between pre- and post-test scores

Repeated measures: None

Adapted from Piquette et al., 2014

PRACTICE: MULTILEVEL LONGITUDINAL STUDY

Vignette

Researchers conducted a group-randomized, multilevel longitudinal study to test the effectiveness of a preventive alcohol use intervention. The study focused on urban, low-income, and multi-ethnic populations.

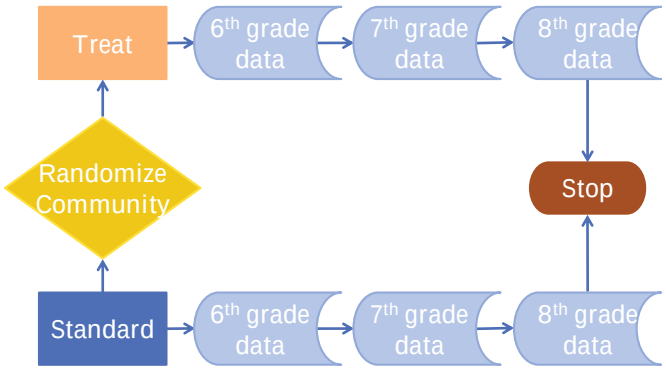
Komro et al., (2008)

Vignette

22 communities of schools were recruited. Communities were randomized into a intervention or control group. Students were surveyed at the end of 6th grade, 7th grade, and 8th grade to measure alcohol use.

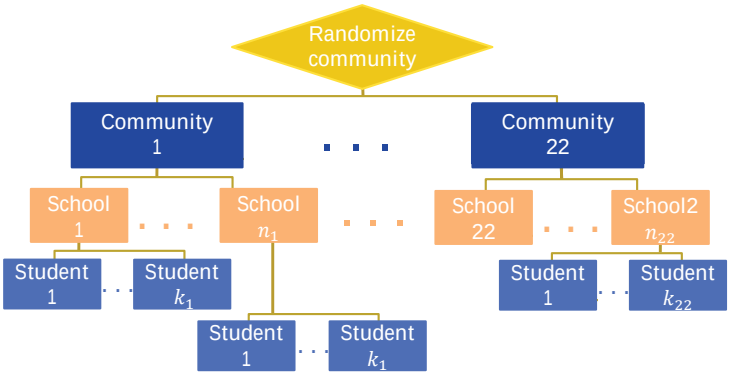
Komro et al., 2008

Clustering and longitudinal study



Logan, Baron and Kohout, 1995

Multilevel longitudinal study



Komro et al., (2008)

Multilevel longitudinal study

Null hypothesis:

There is no significant difference in the pattern of alcohol use across time between communities that receive the standard of care and communities that receive the intervention.

Komro et al., (2008)

Multilevel longitudinal study

Independent sampling unit (ISU): Community

Unit of observation: Student alcohol use measured at one point in time

Within-ISU factor: Time

Between-ISU factor: Intervention group

Komro et al., (2008)

Statistical modelling terminology: Multilevel longitudinal study

Clustering: Schools within each community, students within each school

Predictor: Study condition (intervention group or control group)

Covariate: None

Komro et al., (2008)

Statistical modelling terminology: Multilevel longitudinal study

Response variables: Alcohol use scores

Repeated measures: Grade level (6th grade, 7th grade, 8th grade)

Komro et al., (2008)

REVIEW OF LEARNING OBJECTIVES

What five characteristics should be used to summarize a study design?

1. **Clustering**
2. **Predictors**
3. **A covariate**
4. **Response variables**
5. **Repeated measures**

Day 2 Agenda

9 am – 12 pm

10. Choosing the Test
11. Correlation Structure
12. Power and Type I Error
13. Alignment of Power and Data Analysis

12 – 1 pm

Working lunch with faculty or optional break

Day 2 Agenda

1 – 2 pm

Optional accessible walk and talk with faculty or break time

2 – 5 pm

14. Summarizing Your Design
15. Predicting Missing Data
16. Accounting for Missing Data and Dropout

Learning objectives

Define missing data.

Understand the types of missing data.

Learn sources for predicting missing data in a design.

INTRODUCTION TO MISSING DATA

Missing data occurs when one or more outcome measurements fail to be recorded

Multilevel and longitudinal studies frequently produce missing data.

Missing data can occur for many reasons

Examples:

- Inconsistent participation
- Study drop-out
- Machine failure
- Data entry errors

Missing data complicates estimation and inference in three key ways

1. May bias estimates
2. May affect hypothesis test accuracy
3. Typically reduces power

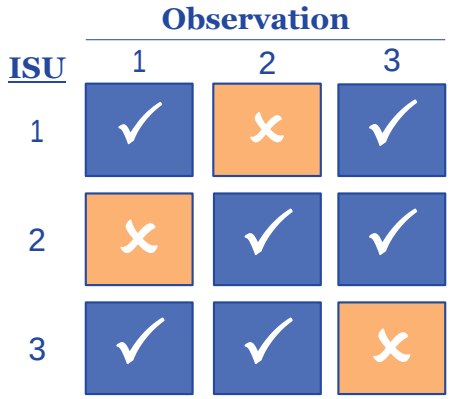
Most power analyses assume that data is missing at random

Data missing at random means that the probability of an observation being missing is not influenced by unobserved data.

Example: Missingness in a study of a cancer treatment is not influenced by patient values not observed.

TYPES OF MISSING DATA

We usually assume data is missing at random



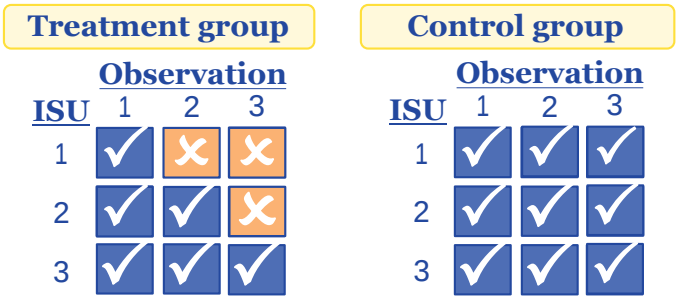
Sometimes missing data may be a result of dropout

Example:

Missing data may be a result of dropout if all other observations after the first missing data point are also unobserved.

Data may be missing at different rates for different levels of a predictor

Example: Missing data may be a result of treatment-related dropout



Investigators must evaluate missing data for patterns of loss to follow-up

Person	Observation 1	2	3
1	✓	✗	✗
2	✓	✓	✗
3	✓	✓	✓

Present: **2/3**

Missing: **1/3**

In particular, be aware of differentially missing data, wherein one group displays higher dropout

Treatment group

ISU	Observation 1	2	3
1	✓	✗	✗
2	✓	✓	✗
3	✓	✓	✓

Control group

ISU	Observation 1	2	3
1	✓	✓	✓
2	✓	✓	✓
3	✓	✓	✓

Treated participants may drop out at a higher rate if the treatment is time consuming or has adverse side effects

IMPLICATIONS FOR STUDY PLANNING

The reasons for missing data may affect power analysis

Differential dropout can affect design choices.

If more women miss visits, one may need to recruit more women. Does bias result?

If some treatment group members develop treatment related toxicity and leave the study, the effect has to be taken into account.

Missing data percentage and pattern can be predicted through several methods

Methods include:

- Literature review
- Internal pilot studies
- Planned pilot studies
- Published or unpublished studies from your laboratory
- Clinical knowledge of study population

As an example, consider the excerpt below from a published study

RESULTS: Out of a total of 68 patients, 60 (88.2%) completed the study treatment without serious adverse events. Treatment in two (2.9%) patients was discontinued due to elevated AST or ALT levels to more than three times the upper limit of normal, and noncompliance or loss of follow-up in six (8.8%) patients. Of the 60 patients who completed the study treatment, mean fasting plasma glucose, A1C, fasting plasma insulin, mean ALT and homeostasis model assessment for insulin resistance were all significantly reduced. Normal AST and ALT levels were achieved and maintained for at least three consecutive measurements and through to the end of the study period in 20 (33.3%) patients. Weight increased by a mean of 2.6 +/- 2.4 kg ($p < 0.001$).

Based on the study, an investigator could predict the percentage missing due to adverse side effects, non-compliance, and loss to follow-up.

The key statement reports drop-out, noncompliance, and loss to follow-up

RESULTS: Treatment in two (2.9%) patients discontinued due to elevated AST or ALT levels to more than three times the upper limit of normal, and noncompliance or loss of follow-up in six (8.8%) patients.

We will further discuss other methods of prediction tomorrow

Other methods include:

- Literature review
- Internal pilot studies
- Planned pilot studies

Many data analysis approaches do not allow any missing data

Recall that the general linear multivariate model does not allow any missing outcomes.

Researchers should design studies to minimize risk of missing data

Best practices include using reliable technology, establishing data validation processes, and minimizing the time and discomfort required of participants.

One approach for design is to first calculate sample size assuming no missing data and then adjust

After initial sample size calculation, design adjustments can be made to account for the predicted *amount* and *pattern* of missing data.

What is missing at random data?

Missing at random means that the chance that an outcome measurement is missing is not related to the unobserved data.

REVIEW OF LEARNING OBJECTIVES

How can you identify when data is not missing at random?

You can analyze patterns in missing data numerically or graphically.

Day 2 Agenda

9 am – 12 pm

10. Choosing the Test
11. Correlation Structure
12. Power and Type I Error
13. Alignment of Power and Data Analysis

12 – 1 pm

Working lunch with faculty or optional break

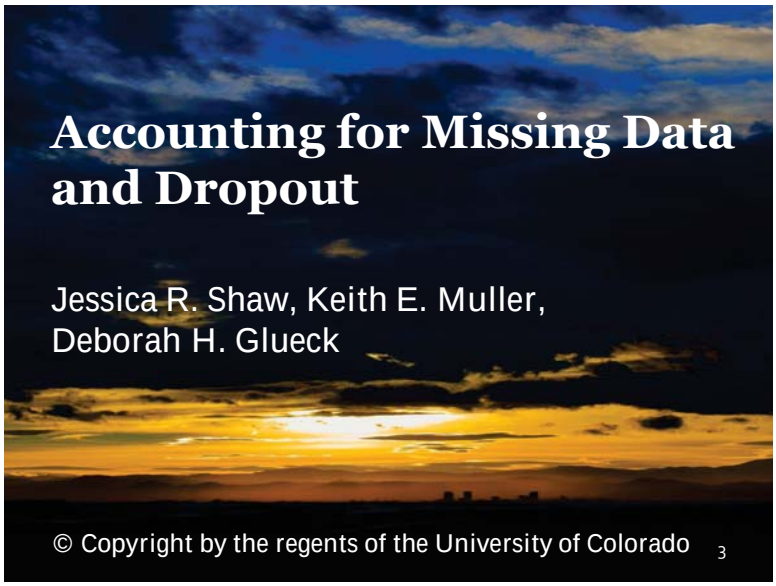
Day 2 Agenda

1 – 2 pm

Optional accessible walk and talk with faculty or break time

2 – 5 pm

14. Summarizing Your Design
15. Predicting Missing Data
16. Accounting for Missing Data and Dropout



Learning Objectives

Review the definition of missing data.

Review the definition of power.

Define dropout.

Understand how missing data causes problems.

Recall the longitudinal study of pain perceived after a root canal

Vignette

Researchers conducted a study to determine if patients who are instructed to use a sensory focus have a different pattern of long-term memory of pain than patients who did not. Patients were selected and randomly assigned to either intervention or no intervention.

Logan, Baron, and Kohout, 1995

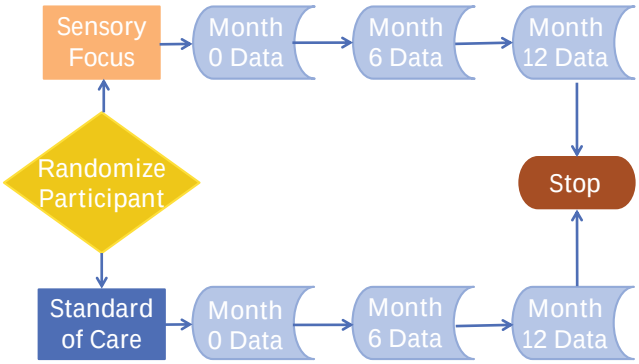
Recall the longitudinal study of pain perceived after a root canal

Vignette, continued

Patients in the intervention group listened to automated audio instructions to pay close attention only to the physical sensations in their mouth. Patients in the no intervention group listened to automated audio instruction on a neutral topic to control for media and attention effects.

Logan, Baron, and Kohout, 1995

Observed pain after root canal was measured at 0, 6, and 12 months



Logan, Baron, and Kohout, 1995

The scientists wished to compare the experience of the two groups

Measurements:

Patients were asked to rate their memories of pain at baseline, and at 3 and 6 months

Logan, Baron, and Kohout, 1995

The scientists wished to compare the experience of the two groups

Scientific goal:

The goal of the study was to compare the intervention group to the standard of care group in terms of their pattern of memory over time.

Logan, Baron, and Kohout, 1995

The scientists wished to compare the experience of the two groups

Null hypothesis:

There is *no difference* in the pattern of pain over time between groups. In other words, there exists no group by time interaction in patients' recall of pain.

The scientists planned a repeated measures analysis of variance

Planned analysis:

A *repeated measures* analysis of variance allows testing all trends across time and their interaction with treatment.

The scientists planned a repeated measures analysis of variance

Model:

The scientists planned to fit a general linear multivariate model, with the repeated measurements of memory of pain as the outcome, and indicator variables for group as the predictors.

The scientists planned a repeated measures analysis of variance

Statistical test:

The scientists planned to use a Hotelling-Lawley test to test interaction between group and time with an 0.05 Type I error rate.

The scientists understood the missing data pattern

Over 12 months, researchers expected a

25% loss to follow up.

Logan, Baron, and Kohout, 1995

The scientists understood the missing data pattern

They recognized that patients left the clinic because their insurance status changed or because their employers changed insurance.

Logan, Baron, and Kohout, 1995

The scientists understood the missing data pattern

Further, they discerned that leaving was unassociated with the memory of pain, the treatment, income, or any other factor they could measure.

Logan, Baron, and Kohout, 1995

FIRST DRAFT OF SAMPLE SIZE SUMMARY

Preliminary calculations found a required sample size of 44 participants

We plan a repeated measures ANOVA using the Hotelling-Lawley Trace to test for a time by treatment interaction. Based on previous studies, we predict that repeated measures of pain recall will have a standard deviation of 0.98.

Logan, Baron, and Kohout, 1995

Preliminary calculations found a required sample size of 44 participants

The correlation in pain recall between baseline and 6 months will be 0.5. Informed by clinical experience, we predict that the correlation will decrease slowly over time. Thus, we anticipate a correlation of 0.4 between pain recall measures at baseline and 12 months.

Logan, Baron, and Kohout, 1995

Preliminary calculations found a required sample size of 44 participants

For a desired power of 0.90 and a Type I error rate of 0.01, we need to enroll 44 participants to detect a mean difference of 1.2.

Logan, Baron, and Kohout, 1995

ACCOUNTING FOR MISSING DATA

The required sample size must be adjusted in anticipation of missing data

Inflate sample size by the 25% anticipated loss to follow up.

$$\frac{44}{1-0.25} = 58.6$$

Round up to 60 for equal division into two treatment groups

$$\frac{44}{1-0.25} = 58.6$$

$$\approx 60$$

Revise the sample size summary to include the adjusted sample size

Over 12 months, we expect 25% loss to follow up. We will inflate the sample size by 25% to account for the attrition, for a total enrollment goal of 60 participants, or 30 participants per treatment arm.

Revise the sample size summary to include the adjusted sample size

Over 12 months, we expect 25% loss to follow up. We will inflate the sample size by 25% to account for the attrition, for a total enrollment goal of 60 participants, or 30 participants per treatment arm.

Recently published results give better adjustments

Ringham et al. (2015, *Statistics in Medicine*) gave new and more accurate approximations to adjust for missing data.

Code available at www.SampleSizeShop.org

REVIEW OF LEARNING OBJECTIVES

Sample size calculation should account for which of the following?

- Drop-out
- Data missing at random
- Number of study groups
- Recruitment time
- All of the above

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

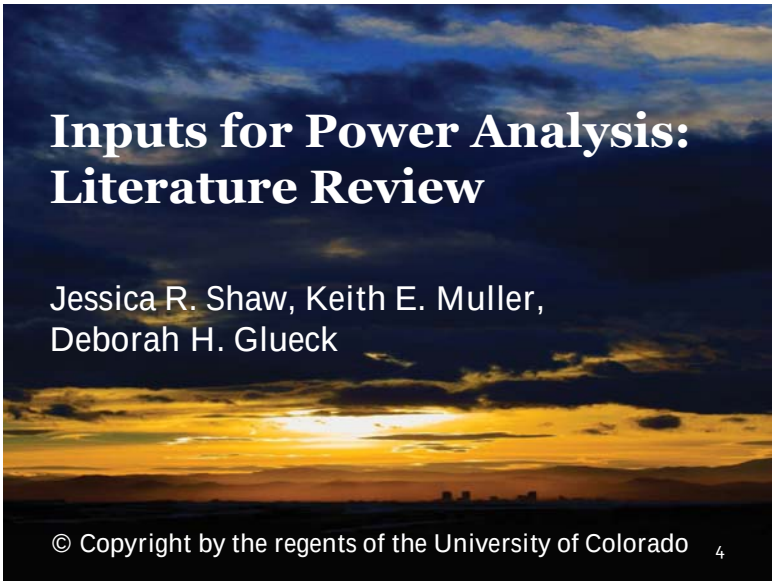
1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded



Learning Objectives

Understand the inputs for power or sample size analysis.

Understand how to search literature for inputs.

UNDERSTAND KEY INPUTS

To perform calculation, specify four types of information

1. Design
2. Statistical test
3. Criterion
4. Five key inputs for power or sample size analysis

Kreidler et. al., 2013
Munjal et. al, 2013

Some inputs for power analysis are derived from the experimental setup

1. Design
2. Statistical test
3. Criterion
4. Five key inputs for power or sample size analysis

Kreidler et. al., 2013
Munjal et. al, 2013

Other inputs for power analysis are collected from external sources

1. Design
2. Statistical test
3. Criterion
4. Five key inputs for power or sample size analysis

Kreidler et. al., 2013
Munjal et. al, 2013

To perform calculation, specify the design, test, criterion, and inputs

Design describes the experiment.

Power, and thus sample size, depends on the choice of statistical test.

Criterion could be power or sample size.

Inputs provide information about the expected results of the experiment.

Kreidler et. al., 2013
Munjal et. al, 2013

Study design involves five inputs

1. Clustering
2. Predictors
3. Covariates
4. Repeated Measures
5. Response

Kreidler et. al., 2013
Munjal et. al, 2013

Depending on the planned analysis, there are multiple statistical tests

Multivariate approach to repeated measures tests

1. Hotelling-Lawley Trace
2. Pillai-Bartlett Trace
3. Wilks' Lambda
4. Roy's largest root

Univariate approach to repeated measures tests

5. Uncorrected
6. Box
7. Geisser-Greenhouse
8. Huynh-Feldt

An analyst needs to specify a criterion

The criterion for sample size is power.

The criterion for power is sample size.

It may be necessary to specify the smallest group size to calculate the correct total sample size

Examples:

In a 1:1 randomization, if the smallest group size is 10, then the total sample size is 20.

In a 2:1 randomization design, if the smallest group size is 10, then the total sample size is 30.

Five key inputs are needed for power or sample size analysis

1. Hypothesis
2. Predictors in model
3. Desired Type I error
4. Hypothesized difference in means, or slopes and intercepts of model
5. Standard deviations and correlation between measurements

The last two inputs are numeric

1. Hypothesis
2. Predictors in model
3. Desired Type I error
4. Hypothesized difference in means, or slopes and intercepts of model
5. Standard deviations and correlation between measurements

In this lecture we will focus on the use of literature review for finding values for the key numeric inputs

Literature review is the systematic collection and examination of past research studies relevant to the present scientific goal.

UNDERSTAND HOW TO SEARCH LITERATURE FOR NUMERIC INPUTS

Literature review can help estimate unknown key numeric inputs for power and sample size calculation

Frequently unknown numeric inputs:

- Standard deviation
- Correlation between measurements
- Scientifically meaningful detectable difference

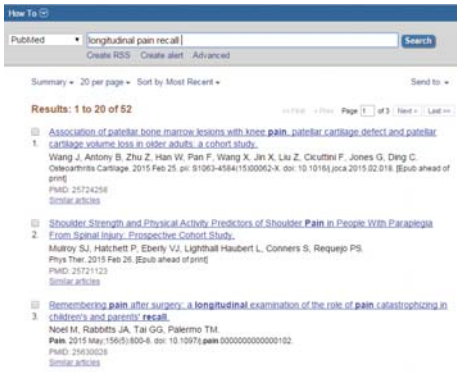
Literature review can also reveal standards for desired power and Type I error

Standards may vary by field of study

- Power:
  - 0.80 common
  - 0.90 usually recommended by us
- Type I error
  - 0.05 (5%)
  - 0.01 (1%) for more conservative analysis

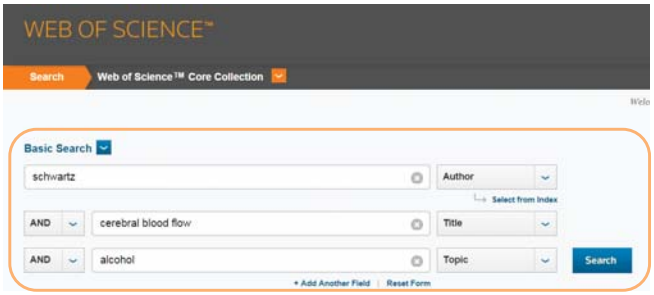
TOOLS FOR LITERATURE REVIEW

PubMed is most used database for literature review for NIH proposals



Web of Science is another useful tool for literature review

If you find a useful paper, you can search for it, and then find other papers that cited it.

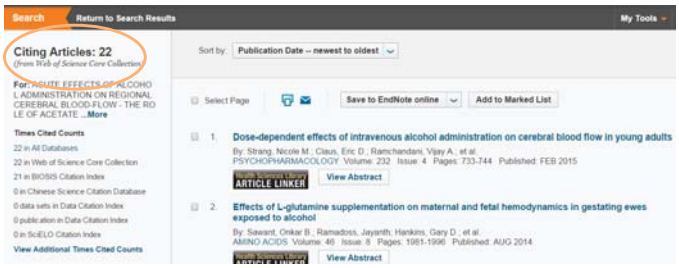


Web of Science allows discovery of manuscripts similar to the planned study

Click on the Times Cited number to the right of the screen.

Examine review papers that have cited the original paper of interest

Review papers may be useful in your search for inputs for power analysis.



Plan your literature review and document your progress

Identify questions of interest, such as unknown variances or standards of design in a given field.

Identify search terms.

Example: field of study + study design

Search for selected terms.

Carefully document the results of your search

You may choose to estimate inputs based on a group of studies in your field of research or one highly representative study.

Study	Author	Title	Field of study	Study design	Standard deviation	Correlation between measurements		
						3 months apart	6 months apart	12 months apart
1	Logan, *et al.*	...	Pain recall	Longitudinal	0.98		Slightly higher than 0.4	0.4
2	...	...	Pain recall	Longitudinal	0.87	...	...	...
3	...	...	Pain recall	Longitudinal	1.02	...	...	...

We show an example literature review summary for numeric inputs

Vignette

Researchers propose a randomized controlled trial to compare pattern of weight gain over pregnancy for two treatment groups. The trial will enroll pregnant women at risk of gestational diabetes.

Vignette

The women will be randomized to either a dietary intervention designed to prevent excess weight gain, or to the standard of care. Weight will be measured at 6 time points during pregnancy.

Vignette

Analysis plan: repeated measures analysis of variance

Planned test: Hotelling-Lawley test

Type I error rate: 0.05

Covariate: pre-pregnant BMI

After finding relevant studies, summarize key inputs found in each

Study	Trt. Group (X, SE in kg)	Control Group (X, SE in kg)	N (treat/ control)	Treatment
Quinlivan *et al.* (2007)	7 (0.65)	13.8 (0.67)	63/61	Continuity of care + weight tracking + 5 min diet counseling + mental health assessment
Wolff *et al.* (2008)	6.6 (1.1)	13.3 (1.4)	23/27	10 (1hr) diet consults throughout pregnancy
Thornton *et al.* (2007)	5.0 (0.6)	14.06 (0.7)	116/116	1 visit with RD + daily food log throughout pregnancy

Summarize unpublished data from one of your own similar studies

The unpublished study provided an estimate of correlation between measurements.

Correlation is rarely published, which makes estimating correlation for power analysis difficult.

Summarize unpublished data from one of your own similar studies

	HS obese women Mean (SD)
Pre-pregnant weight (kg)	95.23 (16.76)
Last weight (kg)	105.05 (17.65)
Correlation	0.87
Predicted weight for 39 weeks (kg)	106.22 (17.50)
Correlation	0.89

Notice that there are multiple differences in means and standard errors cited in the table

	HS obese women Mean (SD)
Pre-pregnant weight (kg)	95.23 (16.76)
Last weight (kg)	105.05 (17.65)
Correlation	0.87
Predicted weight for 39 weeks (kg)	106.22 (17.50)
Correlation	0.89

To account for uncertainty in the power and sample size calculation, a researcher may:

Use a power curve.

Consider many different values of mean differences close to the expected difference.

Consider different values of variance.

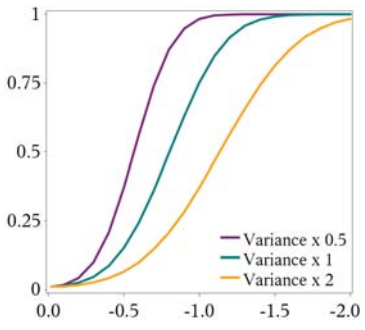
Use the experimental situation to guide the choice of a conservative or liberal estimate.

The next several images show how to incorporate uncertainty from estimates in power curves

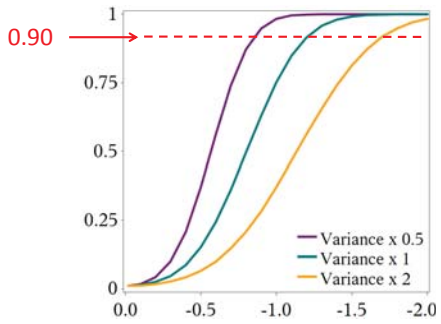
One can draw several curves, one for each variance estimate.

For a given power, one can look at the range of mean differences one could detect for difference variance estimates.

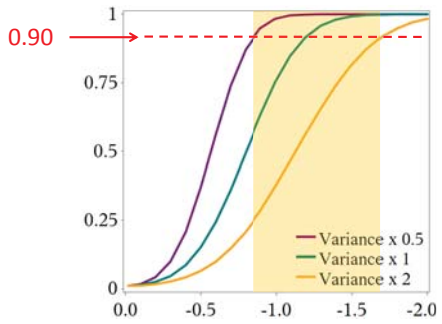
Accounting for uncertainty



Accounting for uncertainty



Accounting for uncertainty



Be careful about using published numbers

The previous table included standard errors, not standard deviations.

Standard errors are smaller, and power analysis will be wrong if one mistakes standard errors for standard deviations.

PRACTICE

For practice, we obtain inputs for sample size analysis for the memory of pain study

Vignette

Researchers conducted a study to determine if dental patients who are instructed to use a sensory focus have a different pattern of long-term memory of pain than participants who did not.

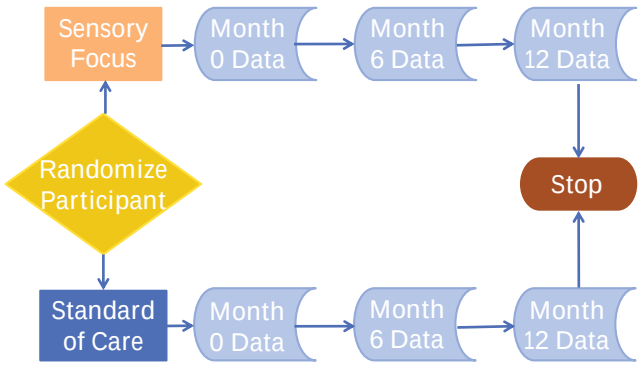
Logan, Baron and Kohout, 1995

Vignette, continued

Participants were selected and randomly assigned to either intervention or no intervention. Those in the intervention group listened to automated audio instructions to pay close attention only to the physical sensations in their mouth. Paticipants in the no intervention group listened to automated audio instruction on a neutral topic to control for media and attention effects.

Logan, Baron and Kohout, 1995

Observed pain after root canal was measured at 0, 6, and 12 months



Logan, Baron and Kohout, 1995

Here, the interest is in an interaction hypothesis: a between-by-within hypothesis

No time by treatment interaction

Some of the information for the study is known already

Design: repeated measures analysis of variance

Analysis plan: Fit general linear multivariate model with three repeated measurements of memory of pain as outcome, and indicator variables for treatment as predictors

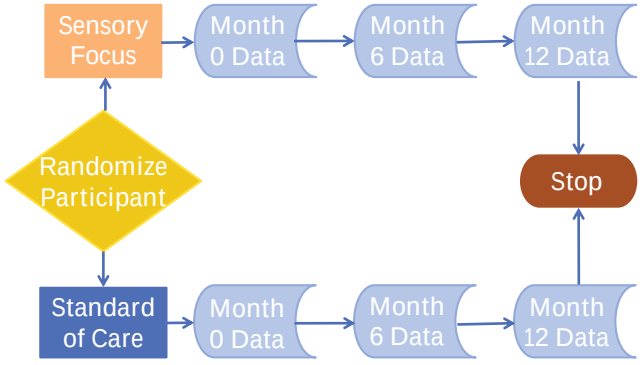
Some of the information for the study is known already

Statistical test: Hotelling-Lawley test

Criterion: sample size

Type I error: 0.01

It remains to summarize the numeric inputs for the sample size calculation



Logan, Baron and Kohout, 1995

We must search the literature to obtain the following numeric inputs for the sample size calculation

Hypothesized difference in means

Standard deviations and correlation between measurements

Kreidler et. al., 2013
Munjal et. al, 2013

Published articles contain hints for numeric inputs

Correlation Between Outcomes Over Time

Gedney, Logan, and Baron (2003) identified predictors of the amount of experienced pain recalled over time…One of the findings was that memory of pain intensity at 1 week and 18 months had a correlation of 0.4. …assume that the correlation between measures 18 months apart will be similar to the correlation between measures 12 months apart. Likewise, the correlation between measures 6 months apart will be only slightly greater than the correlation between measures 18 months apart.

Logan et al, 2009

Published articles contain hints for numeric inputs

Correlation Between Outcomes Over Time

Gedney, Logan, and Baron (2003) identified predictors of the amount of experienced pain recalled over time…One of the findings was that memory of pain intensity at 1 week and 18 months had a correlation of 0.4. … assume that the correlation between measures 18 months apart will be similar to the correlation between measures 12 months apart. Likewise, the correlation between measures 6 months apart will be only slightly greater than the correlation between measures 18 months apart.

Logan et al, 2009

Published articles contain hints for numeric inputs

Correlation Between Outcomes Over Time

Gedney, Logan, and Baron (2003) identified predictors of the amount of experienced pain recalled over time…One of the findings was that memory of pain intensity at 1 week and 18 months had a correlation of 0.4. … assume that the correlation between measures 18 months apart will be similar to the correlation between measures 12 months apart. Likewise, the correlation between measures 6 months apart will be only slightly greater than the correlation between measures 18 months apart.

Logan et al, 2009

Using the previous paragraph, we can identify correlation

Correlation at 6 months apart

Correlation at 12 months apart

Logan et al., 2009

Using the previous paragraph, we can identify our power inputs

Correlation at 6 months apart

Correlation at 12 months apart

0.4

We must estimate certain inputs based on information available

Correlation Between Outcomes Over Time

Gedney, Logan, and Baron (2003) identified predictors of the amount of experienced pain recalled over time...One of the findings was that memory of pain intensity at 1 week and 18 months had a correlation of 0.4... assume that the correlation between measures 18 months apart will be similar to the correlation between measures 12 months apart. Likewise, the correlation between measures 6 months apart will be only slightly greater than the correlation between measures 18 months apart.

We must estimate certain inputs based on information available

Correlation Between Outcomes Over Time

Gedney, Logan, and Baron (2003) identified predictors of the amount of experienced pain recalled over time...One of the findings was that memory of pain intensity at 1 week and 18 months had a correlation of 0.4... assume that the correlation between measures 18 months apart will be similar to the correlation between measures 12 months apart. Likewise, the correlation between measures 6 months apart will be only slightly greater than the correlation between measures 18 months apart.

Interpret the phrase "slightly greater" to estimate the correlation at 6 months

Correlation at 6 months apart

Correlation at 12 months apart

0.4

Interpret the phrase "slightly greater" to estimate the correlation at 6 months

Correlation at 6 months apart

0.5

Correlation at 12 months apart

0.4

Reading further, we identify the standard deviation of measurements

Standard Deviation of the Outcome

Logan, Baron, and Kohout (1995) examined whether sensory focus therapy during a root canal procedure could reduce a patient's experienced pain. The investigators assessed experienced pain on a 5 point scale both immediately and at one week following the procedure. The standard deviation of the measurements was 0.98.

Reading further, we identify the standard deviation of measurements

Standard Deviation of the Outcome

Logan, Baron, and Kohout (1995) examined whether sensory focus therapy during a root canal procedure could reduce a patient's experienced pain. The investigators assessed experienced pain on a 5 point scale both immediately and at one week following the procedure. The standard deviation of the measurements was 0.98.

Information from published articles provides numeric inputs

Standard deviation of memory of pain

0.98

Published means give hints for the treatment difference

Treatment	Baseline	6 Months	12 Months
Sensory Focus (SF)	3.6	2.8	0.9
Standard of Care (SOC)	4.5	4.3	3.0
Treatment Difference (SF - SOC)	-0.9	-1.5	-2.1
Net Difference Over Time (12 Months - Baseline)			-1.2

Logan et al., 2009

Conduct simple subtraction to calculate the treatment difference

Treatment	Baseline	6 Months	12 Months
Sensory Focus (SF)	3.6	2.8	0.9
Standard of Care (SOC)	4.5	4.3	3.0
Treatment Difference (SF - SOC)	-0.9	-1.5	-2.1
Net Difference Over Time (12 Months - Baseline)			-1.2

Logan et al., 2009

Conduct simple subtraction to calculate the treatment difference

Treatment	Baseline	6 Months	12 Months
Sensory Focus (SF)	3.6	2.8	0.9
Standard of Care (SOC)	4.5	4.3	3.0
Treatment Difference (SF - SOC)	-0.9	-1.5	-2.1
Net Difference Over Time (12 Months - Baseline)			-1.2

Logan et al., 2009

Conduct simple subtraction to calculate the treatment difference

Treatment	Baseline	6 Months	12 Months
Sensory Focus (SF)	3.6	2.8	0.9
Standard of Care (SOC)	4.5	4.3	3.0
Treatment Difference (SF - SOC)	-0.9	-1.5	-2.1
Net Difference Over Time (12 Months - Baseline)			-1.2

Logan et al., 2009

Net difference is the difference between treatment differences over 12 months; interaction effect

Treatment	Baseline	6 Months	12 Months
Sensory Focus (SF)	3.6	2.8	0.9
Standard of Care (SOC)	4.5	4.3	3.0
Treatment Difference (SF - SOC)	-0.9	-1.5	-2.1
Net Difference Over Time (12 Months - Baseline)			-1.2

Having identified all inputs, specify each within GLIMMPSE

treatment	memory of pain
sensory focus	-1.2
standard of care	0

Select the time (location, etc.) from the list(s) below. *

time 3

Specify a mean difference for each time point

treatment	memory of pain
sensory focus	-1.2
standard of care	0

Select the time (location, etc.) from the list(s) below. *

time 3

Choose a time point

Specify a mean difference for each time point

This will allow you to edit the means at the selected time or location.

treatment	memory of pain
sensory focus	-1.2
standard of care	0

Select the time (location, etc.) from the list(s) below.

time 3

Choose a time point

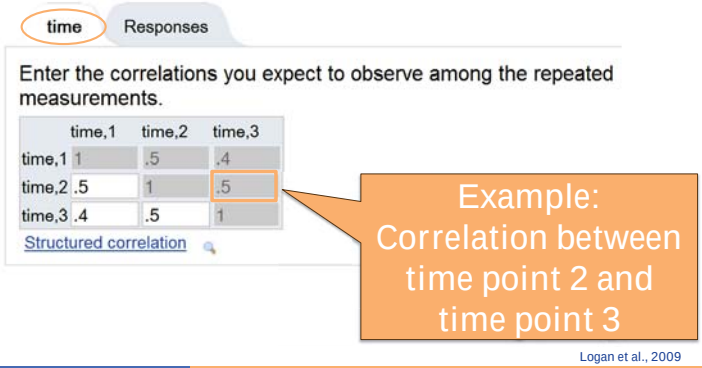
Enter the expected net mean difference

Input your expected standard deviation for each response

Input your expected standard deviation for each response

Input your expected standard deviation for each response

Specify correlations between repeated measures



REVIEW OF LEARNING OBJECTIVES

Briefly discuss why a literature review is a key step in power and sample size calculation

Literature review is an important step in power and sample size calculation because it provides insight into unknown parameters such as standard deviation, correlation between measurements, and detectable difference.

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

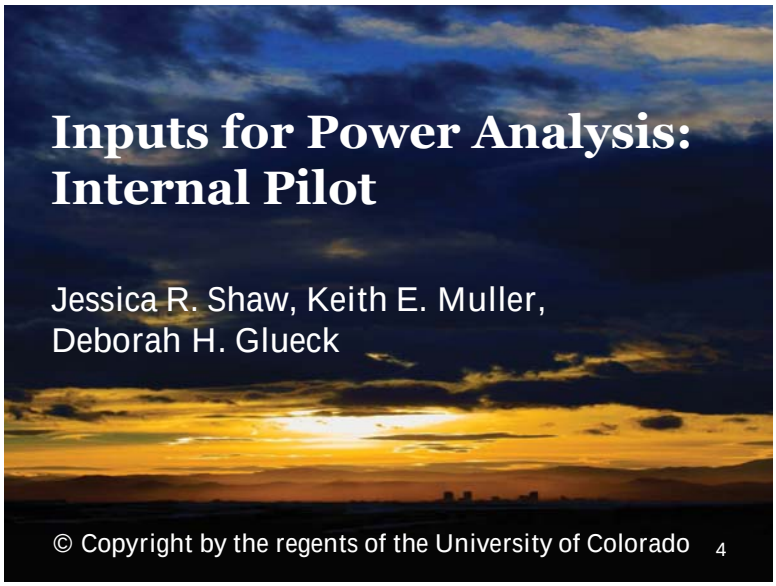
1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded



Learning Objectives

Define internal pilot studies.

Understand the benefits of internal pilot studies.

Understand the limitations of internal pilot studies.

An internal pilot design can be a strategy for identifying inputs for power and sample size analysis

An internal pilot estimates parameters using data from a portion of the main study.

Coffey and Muller, 1999, 2000, 2003
Coffey et al., 2007
Gurka et al., 2007

Internal pilots are conducted in five sequential steps

1. Specify study characteristics
   - design dimensions
   - mean differences of a priori interest
   - planning value standard deviation
   - Recruiting time
2. Collect half of the data (typically)

Coffey and Muller, 1999, 2000, 2003
Coffey et al., 2007
Gurka et al., 2007

Internal pilots are conducted in five sequential steps

3. Re-estimate variance based on pilot
4. Re-calculate sample size
5. Collect remaining data and conduct test

Coffey and Muller, 1999, 2000, 2003
Coffey et al., 2007
Gurka et al., 2007

Internal pilots are useful for interim power analyses

Using a portion of the main study's data enables better estimation of inputs for power and sample size calculation.

Unfortunately, internal pilots have limited applications

Internal pilots are usually not intended for:

- Estimating means or confidence intervals
- Interim data analysis

See Kairalla et al., 2010 for ways to combine internal pilots and interim data analysis.

Kairalla et al., 2010

Internal pilots can also inflate the Type I error rate

Recall, Type I error is the probability of *incorrectly* rejecting the null hypothesis.

Type I error inflation occurs in small studies when the variance is estimated.

For large studies (>~50 participants), no Type I error rate inflation occurs.

Coffey and Muller, 1999, 2000, 2003
Coffey et al., 2007
Gurka et al., 2007

REVIEW OF LEARNING OBJECTIVES

An internal pilot aids in power and sample size analysis in which of the following ways?

- Improve estimation of variance
- Validate parameters by using another researcher's data

Internal pilot studies can be used for which of the following?

- Estimation of variance for power and sample size calculation
- Interim data analysis
- Estimation of means and confidence intervals

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

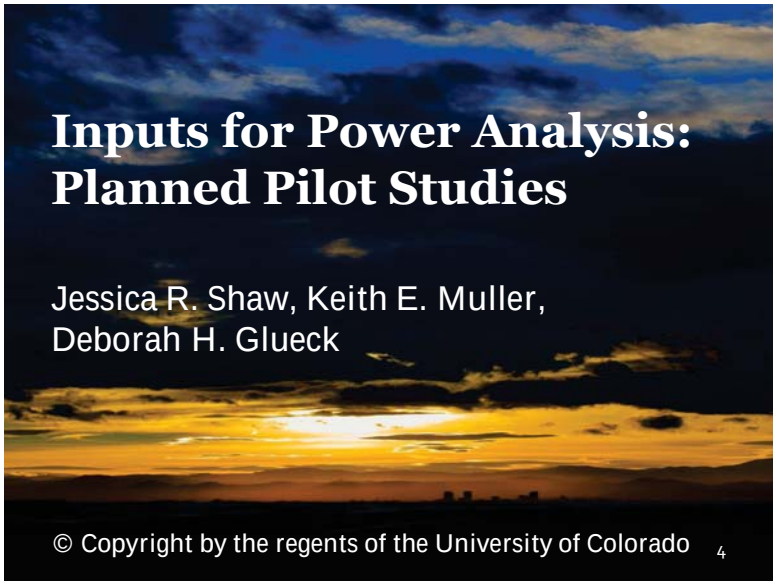
1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded



Learning Objectives

Define a planned pilot study.

Understand the difference between planned pilot studies and internal pilot studies.

A planned pilot is a preview of the actual study, with a similar design

A planned pilot precedes the actual study.

A planned pilot provides proof of concept.

A planned pilot demonstrates recruitment and retention feasibility.

A planned pilot is a preview of the actual study, with a similar design

A planned pilot provides estimates of means, variances and correlation structure.

A planned pilot provides initial data for grant writing.

Data collected in a pilot is *not* included in the final data analysis

Unlike internal pilot data, hypothesis testing should be done *only* on data collected after the end of the pilot, *excluding* pilot data.

Planned pilot data has limitations when it comes to hypothesis testing

Researchers may conduct separate hypothesis tests in pilot.

Unlike internal pilot data, planned pilot data is not used in hypothesis testing in the final study.

Researchers may only test hypotheses in a pilot study separate from the primary analysis

Inappropriate use of pilot data creates bias and increases Type I error.

Many interim analysis papers address the problem.

A common, *inappropriate* use of pilot data involves the following process

1. Testing hypotheses using pilot data
2. Interpreting the p-values
3. Collecting more data
4. Combining pilot data with subsequent study data
5. Testing hypotheses again

Investigators usually plan sequences of studies

As a result, power analyses based on pilot data may be optimistic or pessimistic.

Optimism is highly likely if the pilot analyses lead to changes in the analysis of the pooled data.

If the first study is *non-significant*, scientists may plan a larger study in an attempt to achieve significance

Power calculations are likely *pessimistic*, resulting in too large a sample size.

If the first study is *significant*, scientists may attempt to replicate the result

Power calculations are likely *optimistic*, resulting in too small a sample size.

Pilot study data are used to generate estimates, which are in turn used to calculate power and sample size

Data observed in pilot studies come from random sampling.

Any value based on data collected using random sampling is also random in nature

Values of means, variances and correlations from pilot studies are random; are *statistical estimates*.

Power and sample size values are also *statistical estimates* when based on (random) estimates from pilot studies.

Due to random inputs, power and sample size values have non-zero variance

Confidence bounds for power (statistical) *estimates* can account for variability.

Confidence bounds provide a range of estimates likely to contain the true value.

Taylor and Muller, 1996

REVIEW OF LEARNING OBJECTIVES

What is one purpose of a planned pilot study?

To provide estimates for power and sample size analysis for grants.

Also finds bugs in protocol needing fixed.

In a sentence or two, explain the difference between planned and internal pilot studies

Unlike internal pilot data, planned pilot data is used for grant writing but not for hypothesis testing.

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

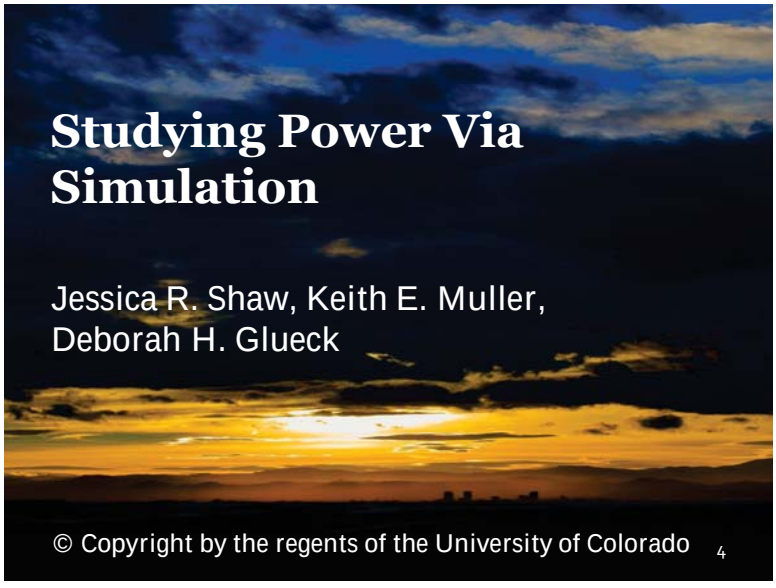
1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Statistical Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded



Learning Objectives

Conceptually understand how to simulate power.

Understand how long simulation takes.

Learn why simulations can be wrong.

We discuss the computational steps behind power simulation

We will not discuss programming or details of how to implement a power simulation.

A power simulation is a statistical program developed to estimate power under hypothetical conditions

Power simulations are frequently programmed in SAS or R.

A power simulation involves the following steps

1. Set the Type I error rate.

2. Specify the experimental design, data analysis, hypothesis.

3. Specify how many times to repeat the simulated experiment (i.e., 10,000 times, the number of replications).

A power simulation involves the following steps

4. Simulate data using an appropriate random number generator.
5. Conduct the hypothesis test.
6. Determine whether to reject the null hypothesis.

Repeat steps 4-6 for each replication.

A power simulation runs the following steps a specified number of times

7. Count the number of rejections
8. Divide the number of rejections by the total number of times the experiment was run.

The resulting fraction is the empirical power (simulated power).

Empirical power is a proportion

$$\text{Empirical power} = \frac{\text{\# of rejections}}{\text{\# of replications}}$$

The number of replications is chosen to limit the error in the empirical power

Margin of error for empirical power is largest when power = 0.50.

Margin of error is half width of the confidence interval.

Margin of error is no more than 0.01 with 10,000 replications

For 95% confidence interval,

$$\text{Margin of error} \approx 1.96\sqrt{\frac{p(1-p)}{\#\text{ of replications}}}$$

The goal of simulation may be to specify power +/- 0.01

Simulated power is 0.92 when true power is 0.93 is typically acceptable.

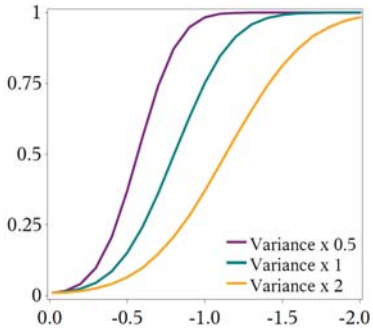
Simulated power of 0.80 when true power is 0.90 is unacceptable.

Simulations are time consuming

Example	Mean GLIMMPSE CPU Time (sec. × 10^{-4})	Mean Simulation CPU Time (sec.)
1	4.1	0.17
2	< 0.1	0.15
3	1.8	0.17
4	< 0.1	0.18
5	1.1	0.32
6	1.9	0.81
7	< 0.1	0.57
8	0.9	15.12
9 MB	< 0.1	0.61
9 MEST	< 0.1	0.61

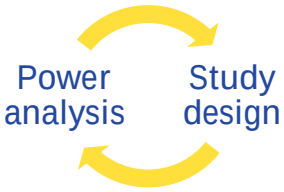
Kreidler et al., 2013

Creating complex power graphs takes much longer than calculating power at a single point



Power analysis is an iterative process

Design changes are made in response to initial power analysis, which requires new power analyses to be conducted.



The total time required for a power analysis is even longer

Total time = Time to write the simulation

+ Time to debug the simulation

+ Time to run the simulation to generate an entire curve

Kreidler et al.,

Each round of power simulations requires extensive effort to validate

Validation efforts may include code review, writing another simulation to validate the original, and having another person write a simulation to provide a comparison.

Further errors can occur at any step in the validation process, making it difficult to confidently reach a conclusion.

Simulations can produce incorrect empirical power

Reasons include

- Mistakes in coding,
- Mistakes in initial assumptions,
- Mistakes in calculations, or
- Mistakes in output.

Kreidler et al., 2013

Mathematical calculation of power also requires extensive validation

Strategies include analytic derivation, comparison to simulation, comparison to other calculations, and peer review.

All apply to GLIMMPSE.

GLIMMPSE has been validated by comparison to published values and to simulations

MAD for **GLIMMPSE** vs. Published Values ($\times 10^{-2}$)	MAD for **GLIMMPSE** vs. Simulation
0.004	0.047
0.003	0.047
0.063	0.032
0.061	0.033
7.300	0.095
7.300	0.094
7.300	0.071
7.300	0.072

Maximum absolute deviation (MAD) is the maximum difference observed over many values.

Kreidler et al., 2013

What four ways can a simulation be wrong?

1. Mistakes in coding.
2. Mistakes in initial assumptions.
3. Mistakes in calculations.
4. Mistakes in output.

REVIEW OF LEARNING OBJECTIVES

Kreidler et al., 2013

Compared to power calculations, how long do simulations take?

1. Simulations usually take longer
2. Simulations take the same amount of time.
3. Simulations are faster.

Kreidler et al., 2013

Day 3 Agenda

9 – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded

Learning Objectives

Define enrollment feasibility.

Understand the importance of enrollment feasibility in a sample size calculation.

Enrollment feasibility is the ability to accrue the planned sample size

Time constraints and funding constraints often limit enrollment potential.

Few trials attain recruitment goals

A 1984 survey of randomized controlled trials in the NIH's inventory found that just **34%** reached their planned recruitment (Charlson, 1984).

Charlson et.al, 1984

Culturally competent strategies may improve recruitment of minority and ethnic subgroups

PubMed cites over 8,000 studies which discuss challenges and strategies for recruitment in clinical trials.

Different strategies are needed for different populations.

Otado et al., 2015

Investigators must consider characteristics of the study population that may affect recruitment

Health factors, socioeconomic factors, and demographic factors can be predictive of recruitment difficulty.

Patel et al., 2003

Factors indicative of poor health may foreshadow difficult recruitment

Examples:

- Recent or present illness
- Frequent use of medical care
- Smoking

Patel et al., 2003

Socioeconomic factors may predict recruitment challenges

Examples:

- Low educational status
- Low occupational status
- Low income

Patel et al., 2003

Demographic factors may also contribute to recruitment challenges

Examples:

- Greater age
- Male gender
- Non-white race
- Urban residence

Patel et al., 2003

Grants should explicitly discuss the practicality of enrollment goals

Investigators should address the following questions:

"Is the target population sufficiently large?"

"Can recruitment be completed in the proposed time period?"

We now demonstrate recruitment feasibility for the longitudinal study of pain recall

Vignette

Researchers conducted a study to determine if dental patients who are instructed to use a sensory focus have a different pattern of long-term memory of pain than patients who did not.

Vignette, continued

Patients were selected and randomly assigned to either intervention or no intervention. Those in the intervention group listened to automated audio instructions to pay close attention only to the physical sensations in their mouth. Patients in the no intervention group listened to automated audio instruction on a neutral topic to control for media and attention effects.

Vignette, continued

On average, the dentist office treated **30** patients per week with a high desire/low felt coping style. Researchers predicted that **40%** of eligible patients would consent to participation in the study.

As demonstrated in lecture 16, adjust the required sample size for anticipated loss to follow up

$$\frac{44}{1 - 0.25} = 58.6$$

$$\approx 60$$

Calculate the number of patients that can realistically be recruited per week

	30	Patients eligible per week
×	0.40	Consent rate
	12	Patients enrolled per week

Using estimated enrollment per week, forecast a realistic timeline for recruitment

$$\frac{60 \text{ patients total}}{12 \text{ patients per week}} = 5 \text{ weeks}$$

Required sample size	Duration of enrollment	Available sample size
60	5 weeks	60 ✓

Include an outline of enrollment timeline in the power and sample size section

The clinic treats 30 patients per week with the high desire/low felt coping style. Based on recruitment experience for previous studies, we expect a 40% consent rate. At an effective enrollment of 12 participants per week, we will reach the enrollment goal of 60 participants in 5 weeks time.

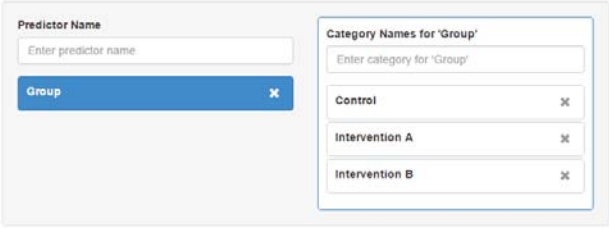
Include an outline of your enrollment timeline in your power and sample size section

The clinic treats 30 patients per week with the high desire/low felt coping style. Based on recruitment experience for previous studies, we expect a 40% consent rate. At an effective enrollment of 12 participants per week, we will reach the enrollment goal of 60 participants in 5 weeks time.

REVIEW OF LEARNING OBJECTIVES

Word problem

Investigators plan a trial to compare alcohol consumption after either no treatment or one of two proposed interventions, A and B.

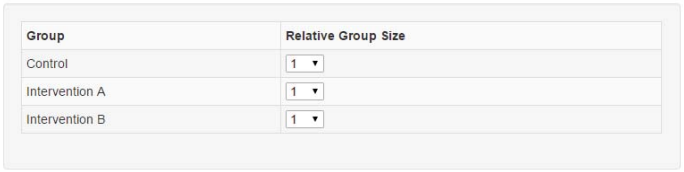


Word problem

Investigators have calculated a required sample size of 400 participants total to achieve 80% power.

Word problem

Participants will be assigned to intervention A, intervention B, or the control group using a 1:1:1 randomization scheme.



Word problem

Recruitment will take place at an outpatient health clinic. From previous experience, the investigators know that 30 people per week who fit the enrollment criteria appear at the clinic. Further, researchers know that approximately 70% of people invited to participate will enroll in the trial.

Word problem

Again from previous studies in the population of interest, researchers know that roughly 20% of participants are lost to follow up. The team needs to finish recruitment within eight months.

Summarize the facts

Required total sample size: 400 people

Number of randomization arms: 3

Size of eligible population: 30 per week

Consent rate: 70%

Expected loss to follow up: 20%

How many people total should the investigators recruit?

Step 1: Adjust for 20% loss to follow up

$$\text{Adjusted} = \left(\frac{400}{1 - 0.20}\right) = 500 \text{ total}$$

How many people total should the investigators recruit?

Step 2: Check that sample size divides equally into three randomization arms

$$\frac{500}{3} = 166.67 \text{ per arm}$$

How many people total should the investigators recruit?

Step 3: Round up to achieve three equal groups

$$\frac{500}{3} = 166.67 \text{ per arm} \approx 167 \text{ per arm}$$

How many people total should the investigators recruit?

Step 4: Report total sample size

$$167 \text{ per arm} \times 3 = 501 \text{ people total}$$

How long will the recruitment take?

Step 1: Calculate feasible recruitment per week

$$\begin{aligned} 30 &\quad \text{Patients eligible per week} \\ \times\ 0.70 &\quad \text{Consent rate} \\ \hline 21 &\quad \text{Patients enrolled per week} \end{aligned}$$

How long will the recruitment take?

Step 2: Given feasible recruitment per week, calculate recruitment time

$$501 \div 21 = 23.86 \text{ weeks} \approx 24 \text{ weeks}$$

Is the recruitment feasible?

Yes. With an estimated recruitment time of 24 weeks, the investigators will accrue their required sample size before their deadline.

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded

Learning Objectives

Define multiple aims.

Understand the challenges that arise from multiple aims.

Understand how to plan for multiple aims.

Studies often have multiple aims

Aims typically represent different hypotheses.

Each aim requires its own power analysis.

Multiple aims can present challenges in power and sample size calculation

Sample size may be adequate for one aim but too large or too small for others.

Power may be adequate for one aim but too large or too small for others.

The goal is to balance harms and benefits of each sample size.

Choosing a sample size that accommodates all aims in the study

1. Calculate the sample size required for each aim individually
2. Choose the **largest sample size** of all calculated

   Warning: the approach ignores any ethical concerns, if present.

For practice, we discuss a grant that includes three aims

Aim 1

Determine whether community gardening leads to increased intake of fruits and vegetables and thus increased intake of fiber, lower total energy intake, and higher Healthy Eating Indices.

For practice, we discuss a grant that includes three aims

Aim 2

Determine whether community gardening leads to reduced sedentary time, increased moderate-to-vigorous physical activity (MVPA), and reduced age-associated weight gain.

For practice, we discuss a grant that includes three aims

Aim 3

Elucidate the mechanisms underlying the differences found in diet, activity, BMI and waist circumference between gardeners and non-gardeners.

Each aim requires its own sample size section describing the following factors

1) Type of test

2) Type I error rate

3) Estimated power

Each aim requires its own sample size section describing the following factors

4) Description of clusters
   - nature of clusters
   - number of clusters
   - people per cluster

Each aim requires its own sample size section describing the following factors

5) Expected loss to follow-up
6) Total sample size, adjusted for loss
7) Sample size per randomization arm, adjusted
8) Correlations accounted for in analysis

When writing your grant, give a preface for your analyses

We conducted a separate power analysis for each aim. A detailed analysis for Aim 1 appears in Section 4.F.1, Aim 2 in 4.F.2, and Aim 3 in 4.F.3. The final sample size for the study is the maximum of the three sample sizes found to provide sufficient power for each aim.

Next, identify and describe your overall sample size

With 156 per randomization arm, and a total sample size of 312, the study will have greater than 80% power for each aim. This estimate reflects an adjustment for 30% loss to follow-up.

Then, provide a detailed sample size summary for each aim

Aim 1

"Estimates of means and variances for fruit and vegetable intake were based on our cross-sectional study in Denver. We conducted a power analysis for the overall test of time by treatment."

Then, provide a detailed sample size summary for each aim

Aim 1, continued

At a Type I error rate of 0.04, power for the Hotelling-Lawley test is estimated at 0.98 with 30 neighborhoods, and 109 people per randomization arm, for a total sample size of 218.

Then, provide a detailed sample size summary for each aim

Aim 1, continued

"Assuming loss to follow-up of 30%, the total sample size required will be 312 people in 39 neighborhoods, or 156 per randomization arm group. The power analysis accounts for correlation within gardens, and within neighborhoods."

Next, provide a detailed sample size summary for each aim

Aim 2

"Estimates for the power analysis were taken from a population-based Neighborhood Environments and Health Survey (NEHS) of 470 residents of Denver (Litt, PI).

Next, provide a detailed sample size summary for each aim

Aim 2, continued

Community gardeners in the study (N=63) reported an average of 146.6 (std. dev = 12.1) hours of sedentary time per week, while non-gardeners reported an average of 153 hours (std. dev. = 9.6).

Next, provide a detailed sample size summary for each aim

Aim 2, continued

Community gardeners in the study (N=63) had an average BMI of 24, while non-gardeners had an average BMI of 27.

Power analysis was conducted both for sedentary time and for BMI.

Next, provide a detailed sample size summary for each aim

Aim 2, continued

With a Type I error rate of 0.04, we calculated the power for a Hotelling-Lawley test of the group by time interaction reflecting a decrease in sedentary time of 6.4 hours per week in midsummer for gardeners (std. dev. = 12).

Next, provide a detailed sample size summary for each aim

Aim 2, continued

Under these conditions, power will be more than 99% for 30 neighborhoods and 130 people per randomization arm, for a total sample size of 240 people.

Next, provide a detailed sample size summary for each aim

Aim 2, continued

Assuming loss to follow-up of 30%, the total sample size required to achieve that power will be 312 in 39 neighborhoods, 156 per randomization arm.

Next, provide a detailed sample size summary for each aim

Aim 2, continued

The power analysis accounts for correlation within gardens and within neighborhoods. The power analysis proposed is conservative, since it is used with the larger of two possible variance estimates.

Next, provide a detailed sample size summary for each aim

Aim 2, continued

For the outcome of moderate-to-vigorous physical activity, we estimate adequate power, based on the difference in means from the NEHS (Litt, PI).

Next, provide a detailed sample size summary for each aim

Aim 2, continued

Community gardeners reported an average of 2.5 more hours a week of moderate to vigorous physical activity, compared to non-gardeners."

Next, provide a detailed sample size summary for each aim

Aim 3

"Power calculations considered the magnitude of individual paths, as well as power to test the mediated effect. Power was calculated in Mplus using Monte Carlo simulation.

Next, provide a detailed sample size summary for each aim

Aim 3, continued

Estimates driving the power analyses are shown in Figure 4 (presented as standardized beta coefficients) and are based on data from the GGHC study.

Next, provide a detailed sample size summary for each aim

Aim 3, continued

This example estimate used social involvement as the mediator between assuming a neighborhood level ICC of 0.012 for physical activity and 0.0886 for fruit and vegetable intake.

Next, provide a detailed sample size summary for each aim

Aim 3, continued

Assuming a two-sided alpha=0.05, the sample size of 240 participants will provide 80% power to detect the mediated effect for fruit and vegetable intake and 95% in all instances."

Your final grant submission should describe the sample size required for each aim individually

Aim	N total	N per randomization arm
1	312	156
2	312	156
3	240	120

Your grant should then identify the largest sample size per aim as the overall sample size

Aim	N total	N per randomization arm
1	312	156
2	312	156
3	240	120

REVIEW OF LEARNING OBJECTIVES

What two steps should an investigator take to plan for multiple aims?

1. Conduct power and sample size analysis for each aim individually.
2. Choose the highest sample size needed for any aim, as long as there are no ethical concerns from doing so.

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded

Learning objectives

Learn how to structure the sample size section of a grant.

We first review the six components required in the sample size section of a grant

1. Align power analysis with data analysis
2. Justify the power analysis
3. Account for uncertainty
4. Plan for missing data
5. Demonstrate enrollment feasibility
6. Plan for multiple aims

For grant-writing practice, we now refer back to our longitudinal pain trial example

Vignette

Researchers conducted a study to determine if patients who are instructed to use a sensory focus have a different pattern of long-term memory of pain than patients who did not.

Adapted from Logan et al., 1995

Vignette, continued

Patients were selected and randomly assigned to either intervention or no intervention. Those in the *intervention group* listened to automated audio instructions to pay close attention only to the physical sensations in their mouth.

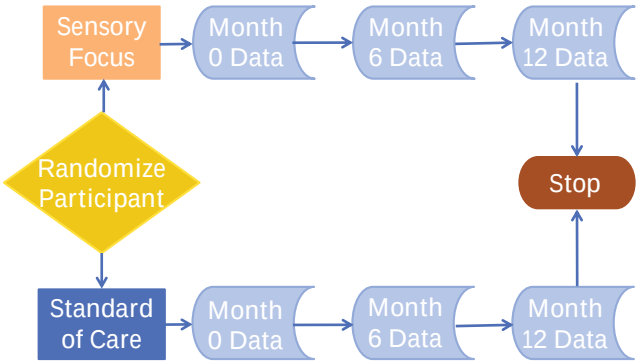
Adapted from Logan et al., 1995

Vignette, continued

Patients in the *no intervention group* listened to automated audio instruction on a neutral topic to control for media and attention effects.

Adapted from Logan et al., 1995

Observed pain after root canal was measured at 0, 6, and 12 months



Adapted from Logan et al., 1995

Researchers analyzed three distinct hypotheses

1 Between-independent sampling unit null hypothesis:

Mean pain experienced by patients in the sensory focus treatment group does not significantly differ from that experienced by patients in the standard of care group.

Adapted from Logan et al., 1995

Researchers analyzed three distinct hypotheses

2 Within-independent sampling unit null hypothesis:

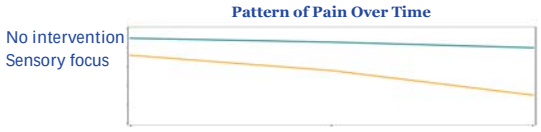
Mean pain experienced does not vary significantly over time.

Adapted from Logan et al., 1995

Researchers analyzed three distinct hypotheses

3 Between by within-independent sampling unit null hypothesis:

The **pattern** of pain perception over time is different for the sensory-focus group than for the control group.



Adapted from Logan et al., 1995

1. ALIGN POWER ANALYSIS WITH DATA ANALYSIS

In order to align power analysis, you must first indicate the type of data analysis you will employ

"We plan a repeated measures ANOVA using the Hotelling-Lawley Trace to test for a time by treatment interaction."

2. JUSTIFY THE POWER ANALYSIS

Begin by outlining your power inputs

Give all the values needed to recreate the power analysis.

Provide appropriate citation.

Describe and justify your parameter choices (preference, estimate, …)

Based on previous studies, we predict memory of pain measures will have a standard deviation of 0.98 and the correlation between baseline and 6 months will be 0.5. Based on clinical experience, we believe the correlation will decrease slowly over time, for a correlation of 0.4 between pain recall measures at baseline and 12 months.

Describe and justify your parameter choices

Based on previous studies, we predict memory of pain measures will have a standard deviation of 0.98 and the correlation between baseline and 6 months will be 0.5. Based on clinical experience, we believe the correlation will decrease slowly over time, for a correlation of 0.4 between pain recall measures at baseline and 12 months.

Next document your power goals and inputs

Based on the parameter choices, for a desired power of 0.90 and a Type I error rate of 0.01, we estimated that we would need 44 participants to detect a mean difference of 1.2.

3 & 4. ACCOUNT FOR UNCERTAINTY & MISSING DATA

Adjust your sample size for anticipated missing data and revise your grant section

Over 12 months, we expect 25% loss to follow up. We will inflate the sample size by 25% to account for the attrition, for a total enrollment goal of 60 participants, or 30 participants per treatment arm.

5. DEMONSTRATE ENROLLMENT FEASIBILITY

Include an outline of your enrollment timeline in your power and sample size section

The clinic treats 30 patients per week with the high desire/low felt coping style. Based on recruitment experience for previous studies, we expect a 40% consent rate. At an effective enrollment of 12 participants per week, we will reach the enrollment goal of 60 participants in 5 weeks time.

6. PLAN FOR MULTIPLE AIMS

Plan for multiple aims

Recall that aims typically represent different hypotheses.

Select the maximum of the sample sizes calculated for each aim as the overall required sample size for the study.

Your final sample size section of your grant should resemble the following paragraphs

We plan a repeated measures ANOVA using the Hotelling-Lawley Trace to test for a time by treatment interaction. Based on previous studies, we predict measures of pain recall will have a standard deviation of 0.98. The correlation in pain recall between baseline and 6 months will be 0.5.

Your final sample size section of your grant should resemble the following paragraphs

Based on clinical experience, we predict that the correlation will decrease slowly over time. Thus, we anticipate a correlation of 0.4 between pain recall measures at baseline and 12 months. For a desired power of 0.90 and a Type I error rate of 0.01, we need to enroll 44 participants to detect a mean difference of 1.2.

Your final sample size section of your grant should resemble the following paragraphs

Over 12 months, we expect 25% loss to follow up. We will inflate the sample size by 25% to account for the attrition, for a total enrollment goal of 60 participants, or 30 participants per treatment arm.

Your final sample size section of your grant should resemble the following paragraphs

The clinic treats 30 patients per week with the high desire/low felt coping style. Based on recruitment experience for previous studies, we expect a 40% consent rate. At an effective enrollment of 12 participants per week, we will reach the enrollment goal of 60 participants in 5 weeks time.

REVIEW OF LEARNING OBJECTIVES

We discussed six components that should be in the sample size section of a grant. Name as many as you can.

1. Align power analysis with data analysis
2. Justify the power analysis
3. Account for uncertainty
4. Plan for missing data
5. Demonstrate enrollment feasibility
6. Plan for multiple aims

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

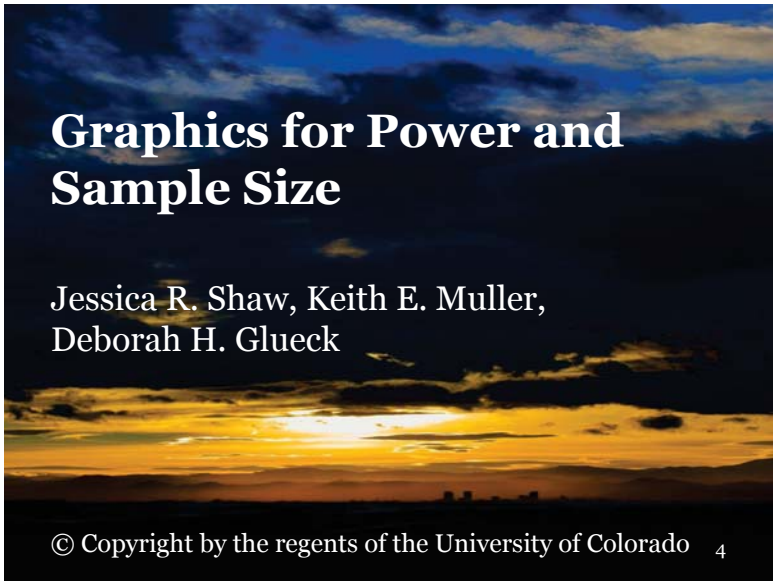
1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded



Learning Objectives

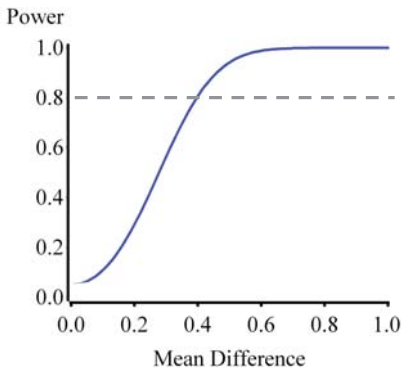
Understand power curves.

Decide how to choose a graphic that best tells the desired story.

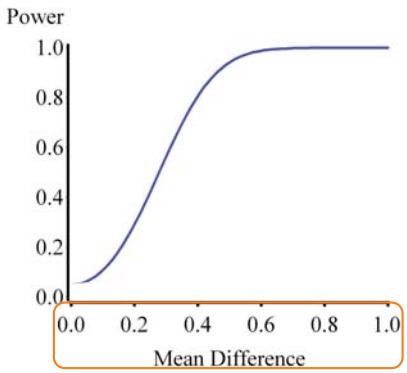
Understand how design inputs affect power.

POWER CURVES

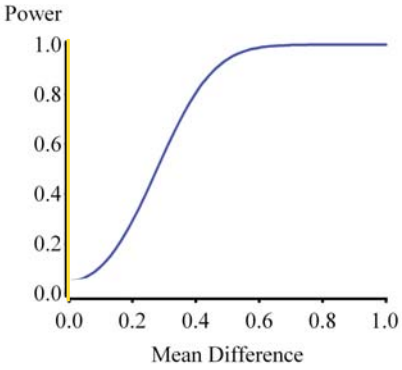
Recall the power curve



Null and alternative hypotheses are described in terms of mean difference

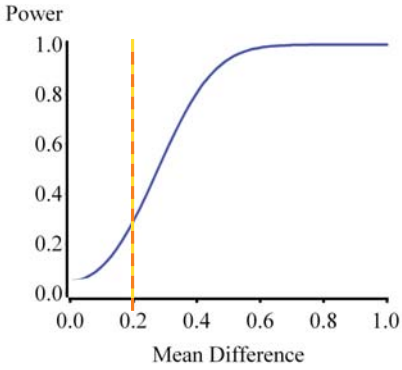


A mean difference of zero represents the null hypothesis being true



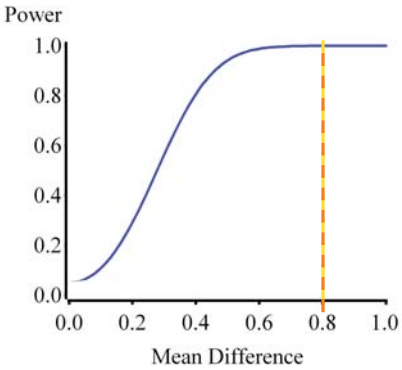
	True H_0	False H_0
Fail to reject	Correct decision	Type II error
Reject	Type I error	Correct decision

A mean difference not equal to zero represents an alternative hypothesis



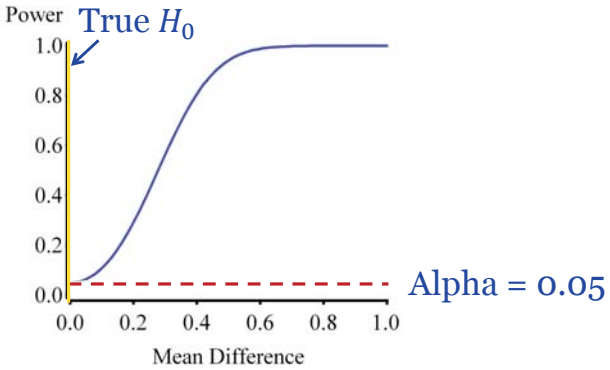
	True H_0	False H_0
Fail to reject	Correct decision	Type II error
Reject	Type I error	Correct decision

The x axis of a power curve represents many possible alternative hypotheses

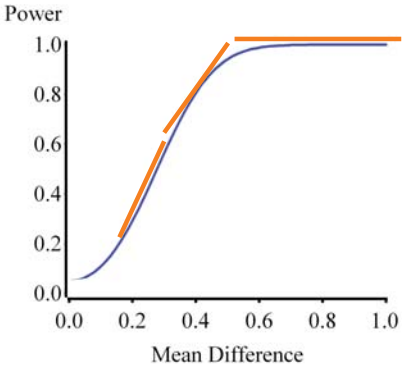


	True H_0	False H_0
Fail to reject	Correct decision	Type II error
Reject	Type I error	Correct decision

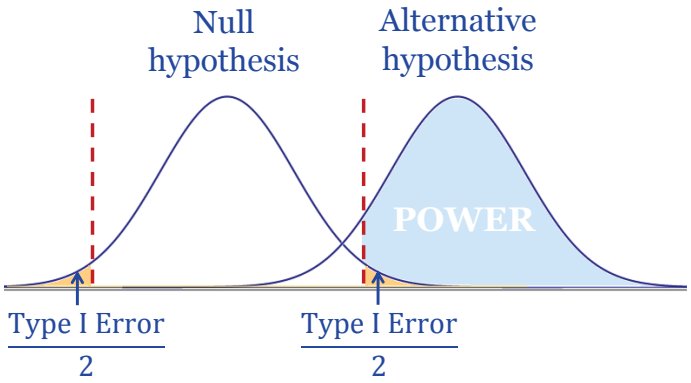
Power is equal to the Type I error rate when the null hypothesis is true (for an exact test)



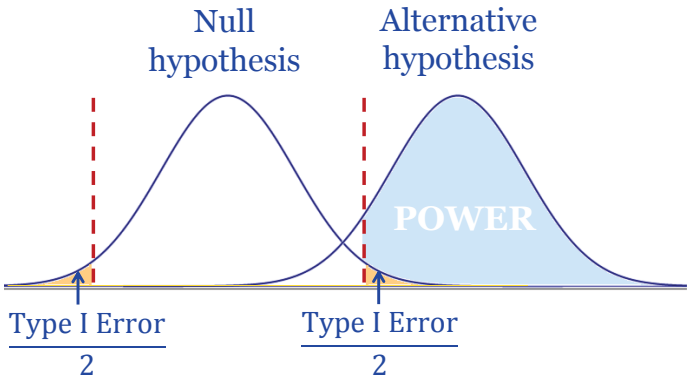
The power curve eventually flattens out as the mean difference gets larger



As the Type I error rate decreases, power also decreases



A lower Type I error requires a larger sample size to achieve the same power



GRAPHICS WHICH TELL STORIES

Graphical representations of power calculations tell stories

Each graph should convey one idea.

We will look at several example graphs.

The examples will serve as templates for presenting power analysis in grants and manuscripts.

Recall the group randomized trial example

Vignette

A single level study examined the efficacy of a workplace training program designed to reduce alcohol consumption. Researchers randomized workplaces to one of two treatment groups.

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

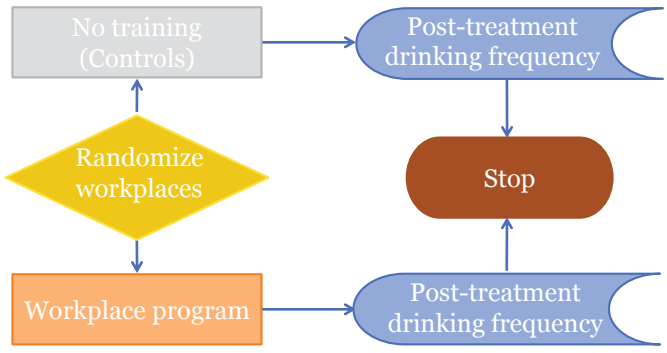
Workplaces were randomized to two different treatments

Vignette, continued

The first treatment included a workplace training program and the second treatment included no training. Post-treatment drinking frequency for each worker was measured as the outcome of interest.

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

Employees were surveyed after training or no training

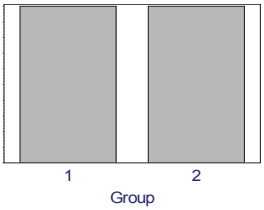


Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

Researchers hoped to observe an association between treatment and reduction in drinking

Null hypothesis:

There is no difference in post-treatment drinking frequency between workers who receive no training and workers who receive the workplace program.



This was a single level design with one outcome measurement in time

Independent sampling unit: Workplace

Unit of observation: Drinking frequency for each employee after treatment

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

This study evaluated the impact of treatment on employee behaviors

Between-independent sampling unit factor:

Intervention (standard of care or workplace drinking program)

Within-independent sampling unit factor:

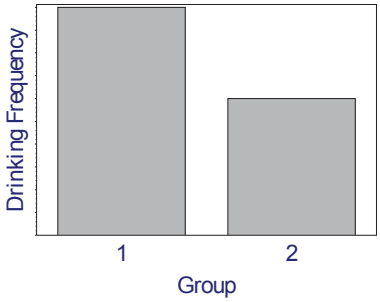
Cluster membership

Interest in factor depends on setting

Adapted from Reynolds, G. Shawn and Joel B. Bennett, 2015

Analysis compared post-treatment drinking frequency between the two treatment groups

Scientific goal: **REJECT** the null hypothesis



Measurements of drinking frequency of workers within a workplace are correlated

Some workers drink together.

Some workers attend sobriety programs together.

Workers discuss drinking habits with each other.

Randomization allowed researchers to make two important assumptions

1. The average correlation between workers in the intervention groups **is equal to** the average correlation between workers in the control groups.

2. Pre-existing employee factors did not bias study outcomes.

What is the best way to present a power analysis?

The choice of graphic depends on the question one is trying to answer.

We present multiple questions and associated graphs.

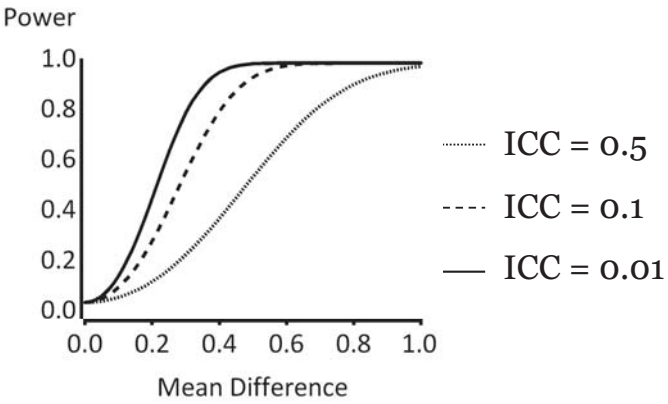
What is the effect on power of correlation within a cluster (ICC)?

Suppose we have 20 workplaces assigned to no intervention, and 20 assigned to a workplace treatment program.

Suppose we have 10 workers in each workplace.

Assume we have the Type I error rate set at 0.05, and that the variance of the outcome is 1.

Power decreases as correlation within a cluster (ICC) increases



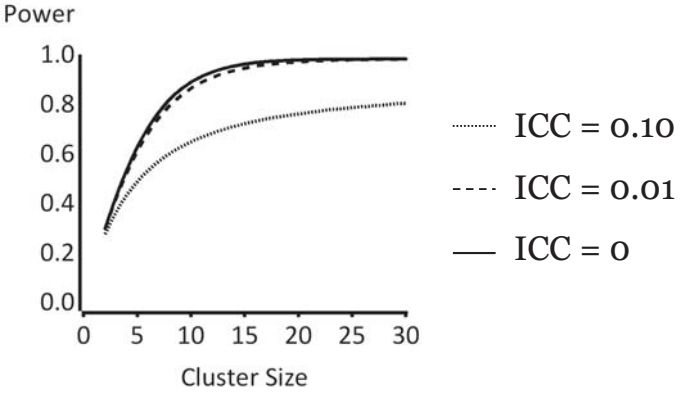
For a given cluster size, what is the effect of the intracluster correlation (ICC) on power?

Suppose we have 5 workplaces assigned to no intervention, and 5 assigned to a workplace treatment program.

Assume that the variance is 1, the mean difference is 0.75, the workplace size is 10, and the Type I error rate is 0.05

Vary the ICCf rom 0 to 0.1.

Power decreases as correlation within the cluster (ICC) increases



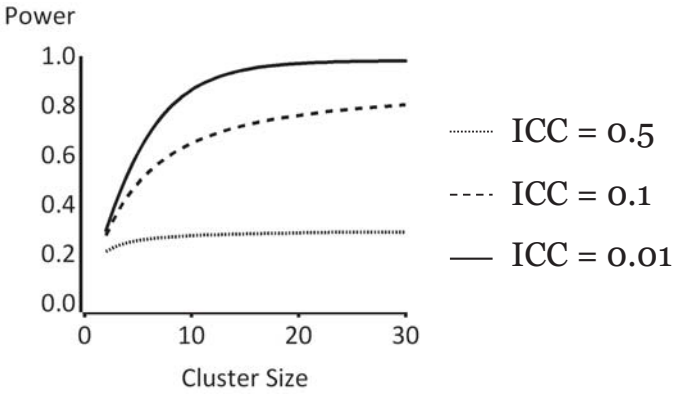
What is the effect on power of cluster size?

Again suppose we have 20 workplaces assigned to no intervention, and 20 assigned to a workplace treatment program.

Assume we have the Type I error rate set at 0.05.

Vary the cluster size from 10 to 30.

Power increases as cluster size increases



What is the effect on power of variance?

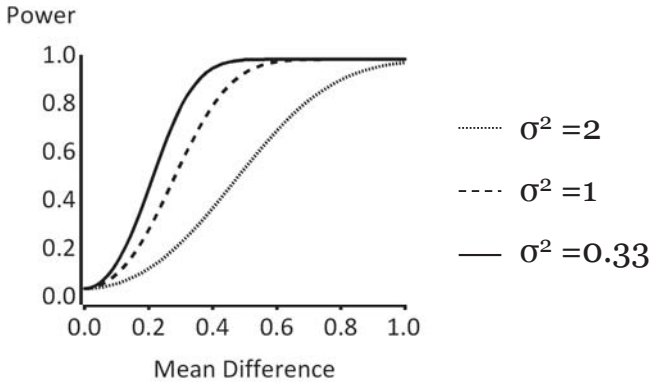
Again suppose we have 20 workplaces assigned to no intervention, and 20 assigned to a workplace treatment program.

Assume that the cluster size is 10.

Assume we have the Type I error rate set at 0.05.

Change the variance from 0.33 to 2.

Power decreases as variance increases



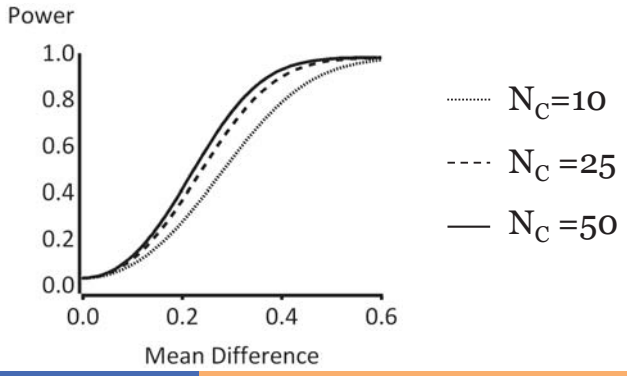
For a given mean difference, what cluster size should I choose?

Again suppose we have 20 workplaces assigned to no intervention, and 20 assigned to a workplace treatment program.

Assume that the variance is 1, the Type I error rate is 0.05 and the intracluster correlation (ICC) is 0.1.

Change the cluster size from 10 to 50.

For a fixed number of clusters, power increases as cluster size (N_C) increases



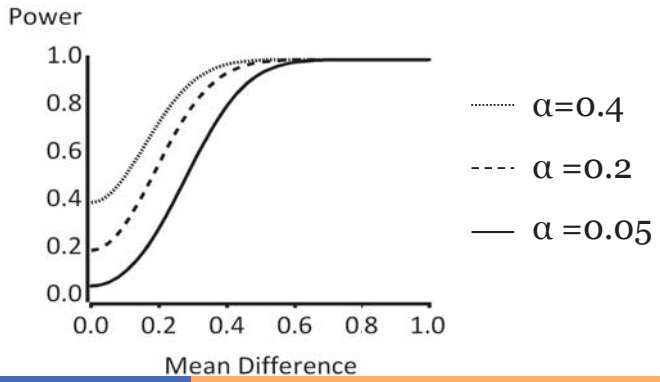
For a given mean difference, how does Type I error rate affect power?

Again suppose we have 20 workplaces assigned to no intervention, and 20 assigned to a workplace treatment program.

Assume that the variance is 1, the workplace size is 10, and the intracluster correlation (ICC) is 0.1.

Change the Type I error rate from 0.05 to 0.4.

Power increases as Type I error rate(α) increases



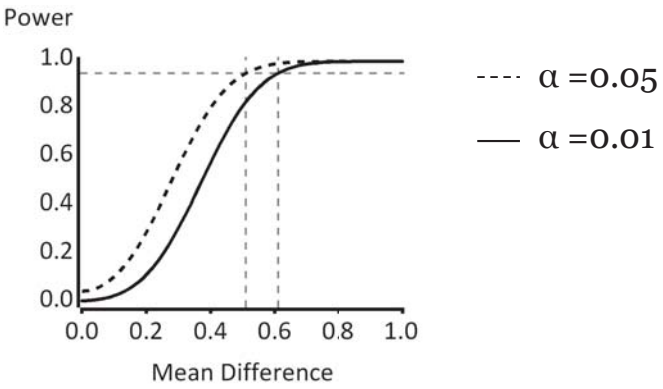
Is there a big difference in power between two designs with different Type I error rates?

Again suppose we have 20 workplaces assigned to no intervention, and 20 assigned to a workplace treatment program.

Assume that the variance is 1, the workplace size is 10, and the intracluster correlation (ICC) is 0.1.

Vary the Type I error rate from 0.01 to 0.05.

There is little change in power for different Type I error rates



REVIEW OF LEARNING OBJECTIVES

What is the relationship between the following design inputs and power?

Intracluster correlation — Power

What is the relationship between the following design inputs and power?

Cluster size — Power

What is the relationship between the following design inputs and power?

Variance Power

What is the relationship between the following design inputs and power?

Type I error rate Power

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

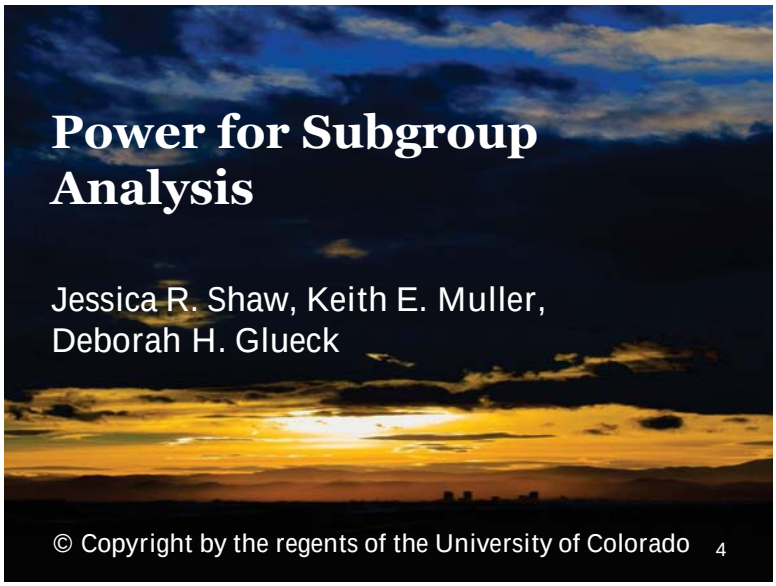
1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded



Learning Objectives

Define subgroup analysis.

Understand how to conduct power and sample size calculations for subgroup analysis.

Subgroups may be defined by a variety of characteristics

Examples include race, ethnicity, and gender.

Researchers are often interested in evaluating hypotheses within different groups

A subgroup analysis is a planned analysis that can help researchers understand whether effects differ between subgroups (subgroup by effect interaction or subgroup heterogeneity)

Described as assessing heterogeneity of treatment effects in randomized trials.

Researchers are often interested in evaluating hypotheses within different groups

Subgroup analysis can allow estimation of effects within each subgroup.

Subgroup analysis is often important in the study of disparities.

Studies often have multiple subgroups

Example: A study with both genders and three age groups results in a study with six subgroups.

	Gender	
Age	**Male**	**Female**
10-39	Males, ages 10-39	Females, ages 10-39
40-49	Males, ages 40-49	Females, ages 40-49
50+	Males, ages 50+	Females, ages 50+

We will discuss two ways of conducting power for subgroup analyses

1. Traditional
2. Non-traditional

TRADITIONAL APPROACH

The traditional approach to subgroup analysis requires splitting data into separate sets for analysis

After stratifying data by subgroup, the traditional approach requires a separate hypothesis test within each stratum.

The traditional approach for subgroup analysis has drawbacks

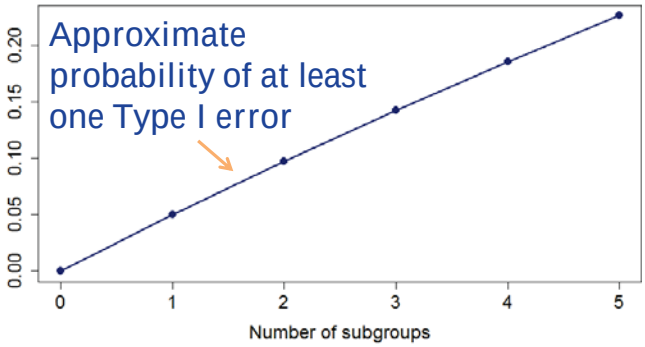
Stratifying the sample size results in small sample size in each individual stratum.

With the traditional approach, the variance estimate comes from each stratum; unbiased but more uncertainty.

In studies with multiple subgroups, conducting a hypothesis test in each one can inflate Type I error rate

Type I error is considered inflated when the true error rate is higher than the planned rate.

The chance of having at least one Type I error increases with the number of subgroups



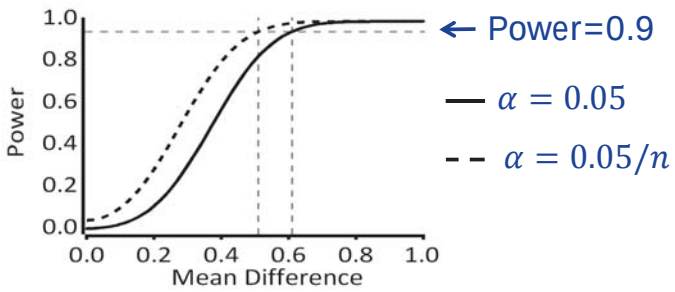
A Bonferroni correction (alpha splitting) divides the Type I error by the number of independent groups

$$\alpha_{\text{original}} = 0.05$$

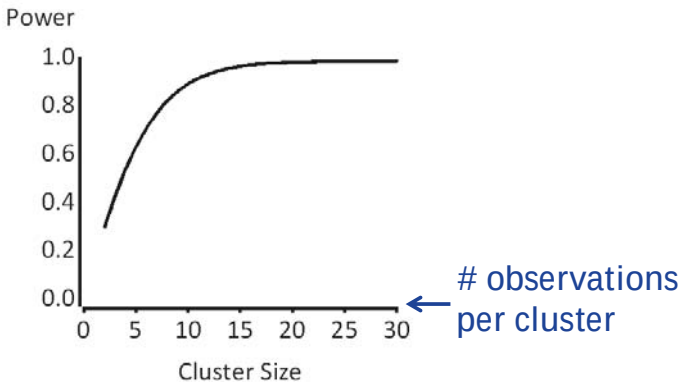
$$\alpha_{\text{corrected}} = \frac{0.05}{n}$$

A Bonferroni correction (alpha splitting) reduces Type I error, but has a modest impact on the detectable mean difference

Example: n=5 subgroups



In contrast, a reduction in sample size has a significant impact on power



Because standard errors are larger in smaller samples, stratification reduces power for subgroup analyses

Traditional subgroup analysis

↓

Sample size stratification

↓

Increased standard errors of means

↓

Reduced power

NON-TRADITIONAL APPROACH

The non-traditional approach to subgroup data analysis does *not* stratify by subgroup

The non-traditional approach to subgroup analysis provides greater power than the traditional approach.

The non-traditional approach yields more power by pooling the degrees of freedom across subgroups

Recall, larger samples have smaller standard errors.

Smaller standard errors give greater power.

We will now briefly describe a data analysis method

Although aligned power analysis requires a basic understanding of the planned data analysis methodology, data analysis is *not* a focus of this course.

Procedure for non-traditional subgroup analysis

1. Rather than stratify the data, fit a model using subgroup definition variables as predictors (between-ISU factors).
2. If X is the effect of interest, add subgroup and subgroup by X interaction effects to the model.
3. Test the subgroup by X interaction.

Procedure for non-traditional subgroup analysis

4. If the interaction is significant, report interaction effect and estimates within subgroups.
5. Otherwise, test the subgroup and X effects.
6. Report tests and estimates.

The non-traditional approach controls experiment-wise Type I error by testing in all subgroups at once

The (single) interaction test looks at all differences between subgroups at once.

The (single) main effect of subgroup test looks at all subgroup differences at once.

Alpha-splitting is only needed for any stepdown tests.

Power is increased by the non-traditional (pooled analysis) approach

Increases power if there are two or more small differences (diffuse effect)

More power for either
a diffuse or a concentrated effect due to increasese error degrees of freedom

REVIEW OF LEARNING OBJECTIVES

What is subgroup analysis?

A subgroup analysis is a planned analysis that allows investigators to test for heterogeneity of effects across subgroups (subgroup by effect interaction), estimate effects within each subgroup, and search for disparities.

Day 3 Agenda

9 am – 12 pm

17. Inputs for Power Analysis: Literature Review
18. Inputs for Power Analysis: Internal Pilot Studies
19. Inputs for Power Analysis: Planned Pilot Studies
20. Studying Power via Simulation
21. Demonstrating Recruitment Feasibility

Day 3 Agenda

12 – 1 pm

Working lunch with faculty or optional break

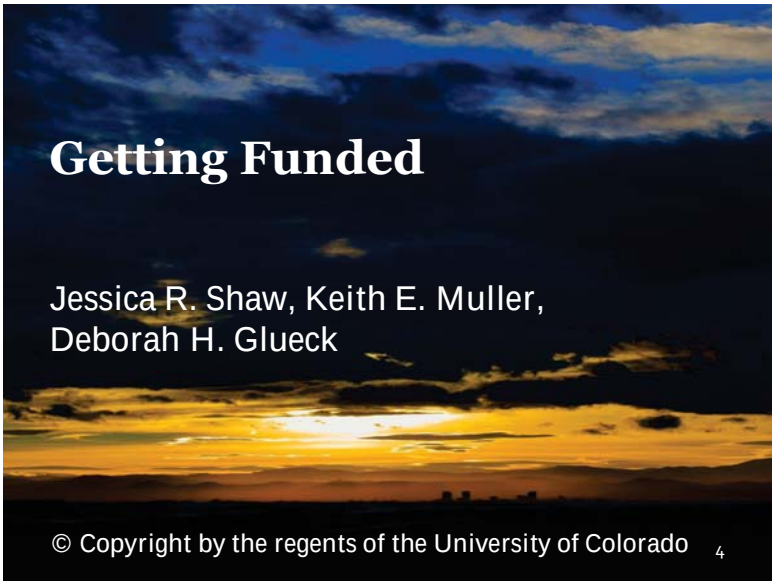
1 – 2 pm

Optional accessible walk and talk with faculty or break time

Day 3 Agenda

2 – 5 pm

22. Handling Multiple Aims
23. Writing the Sample Size Section for your Grant
24. Graphics for Power and Sample Size
25. Power for Subgroup Analysis
26. Getting Funded



Learning Objectives

Understand how to search for funding opportunities.

Understand how to apply for funding opportunities.

Know what to expect when applying for funding.

Recognize that clear design and analysis plans improve chances of funding.

We discuss steps to getting research funded

The process of getting research funded consists of two main phases of preparation:

1. Preparing grant infrastructure
2. Writing the grant

PREPARING GRANT INFRASTRUCTURE

Preparing grant infrastructure involves ten key tasks

1. Define the research idea
2. Identify a funding target
3. Choose a type of grant
4. Research the funding announcement
5. Match your topic to a specific funding announcement

Preparing grant infrastructure involves ten key tasks

6. Explore previously funded research
7. Identify collaborators
8. Write specific aims
9. Talk with program officer
10. Choose a study section

1. Define the research idea

Identify clear research goals.

Consider: What question or questions would you like to answer? What tools would you like to leverage? Which theories or perspectives will guide your inquiry?

2. Identify funding target

Funding targets can be public, such as the NIH, or private, such as the Lundbeck Foundation.

3. Choose a type of grant

In choosing a type of grant, consider:

- The possible amount of funding offered relative to the funding required for your project,
- Which career stage the grant opportunity is targeted,
- The competitiveness of the grant,
- Opportunity costs.

4. Research the funding announcement

Note whether each grant is intended for specific groups of individuals.

NIH provides funding for people from groups historically underrepresented in research, individuals with disabilities, individuals from disadvantaged backgrounds.

http://grants.nih.gov/grants/guide/pa-files/PAR-15-064.html

4. Research the funding announcement

Note the career stage intended for each grant.

Career stages:

- Undergraduate
- Pre-doctoral
- Post-doctoral
- Transition to independence
- Early career
- Mid-career

4. Research the funding announcement

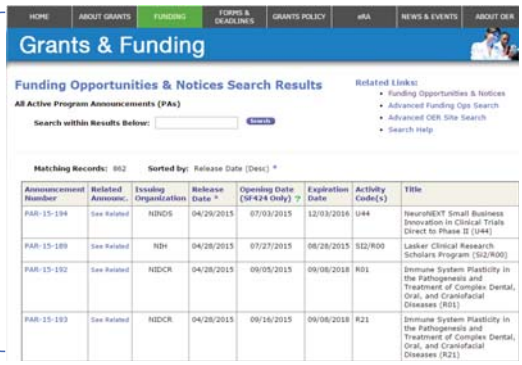
Note the monetary support provided by different grants

Award Mechanism	R01	R03	R21	R15
Title	Research Project Grant	NIH Small Grant Program (R03)	NIH Exploratory / Development Grant Award (R21)	Academic Research Enhancement Award (AREA)
Parent Announcement No.	PA-11-260	PA-11-262	PA-11-261	PA-12-006
Length of Award	Up to 5 years	Up to 2 years	Up to 2 years	Up to 3 years
Allowable Costs (Direct Costs)	Requires advance permission for $500,000 or more in any year	Up to $50,000 per year	Total not to exceed $275,000 for entire period	Up to $300,000 for entire 3-year period

NIH

5. Match your topic to a specific funding announcement

Google "All RFA and PA"



5. Match your topic to a specific funding announcement

Click on grants of interest and copy the program announcement number.

6. Explore previously funded research

Visit NIH RePORTER to investigate successful applications to the program announcements of interest.

6. Explore previously funded research

Read about funding awarded and decide if the grant is a good match for your topic.

If the grant is not a great fit, repeat the previous steps.

6. Explore previously funded research

Confirm that no-one else is funded on exactly the same topic.

Consult at least three resources:

Resource	Information provided
NIH RePORTER	Funding in progress
Conference presentations	Soon to be published research
PubMed	Related publications

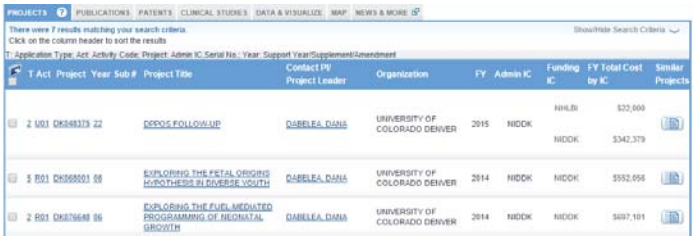
7. Identify collaborators

Aim to partner with established experts who are productive.

Use PubMed to identify researchers with publications in the same field, complementary skillsets, and high levels of productivity and impact.

7. Identify collaborators

Use NIH RePORTER to identify researchers with history of receiving funding.



http://projectreporter.nih.gov/reporter.cfm

7. Identify collaborators

Seek collaborators with whom you are likely to have a positive working relationship.

Use word-of-mouth to find congenial collaborators who are kind, ethical, interested and available for collaboration or mentoring.

8. Write specific aims

Write specific aims to match research goals.

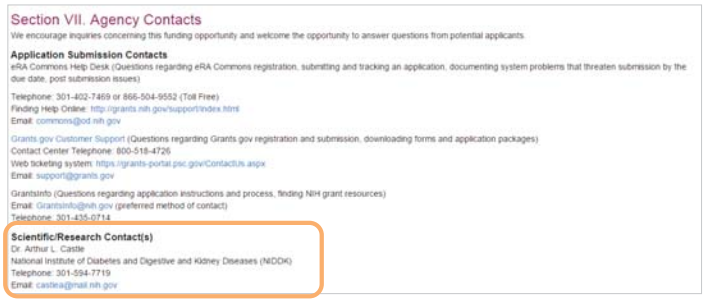
Write aims as testable hypotheses, or provide separate testable hypotheses.

9. Talk with program officer

Ask if your aims are “responsive” to the goals of the announcement

9. Talk with program officer

Program officer contact information can be found at the bottom of RFA and PA.



Section VII. Agency Contacts

We encourage inquiries concerning this funding opportunity and welcome the opportunity to answer questions from potential applicants.

Application Submission Contacts

eRA Commons Help Desk (Questions regarding eRA Commons registration, submitting and tracking an application, documenting system problems that threaten submission by the due date, post submission issues)

Telephone: 301-402-7469 or 866-504-9552 (Toll Free)
Finding Help Online: http://grants.nih.gov/support/index.html
Email: commons@od.nih.gov

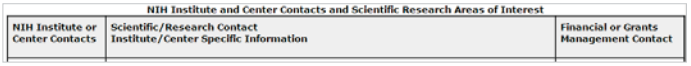
Grants.gov Customer Support (Questions regarding Grants.gov registration and submission, downloading forms and application packages)
Contact Center Telephone: 800-518-4726
Web ticketing system: https://grants-portal.psc.gov/ContactUs.aspx
Email: support@grants.gov

GrantsInfo (Questions regarding application instructions and process, finding NIH grant resources)
Email: GrantsInfo@nih.gov (preferred method of contact)
Telephone: 301-435-0714

Scientific/Research Contact(s)
Dr. Arthur L. Castle
National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)
Telephone: 301-594-7719
Email: castlea@mail.nih.gov

9. Talk with program officer

Parent announcements will offer full lists of contacts

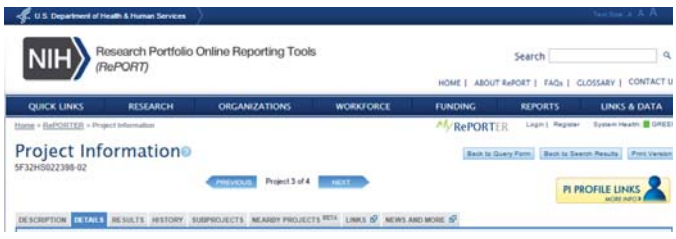


NIH Institute and Center Contacts and Scientific Research Areas of Interest

NIH Institute or Center Contacts	Scientific/Research Contact Institute/Center Specific Information	Financial or Grants Management Contact

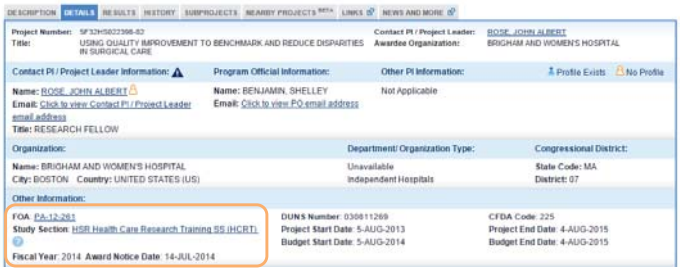
10. Choose a study section

Use NIH RePORTER to look at which study sections provided reviews that led to funding for similar projects.



10. Choose a study section

Identify which study sections reviewed successful grants.



10. Choose a study section

For further information on NIH Study sections, consult the NIH Center for Scientific Review.

The Study Section Search page provides links to all NIH Study Sections and special emphasis panels.

WRITING THE GRANT

Writing the grant involves seven steps

1. Write the aims
2. Write the analysis plan
3. Conduct an aligned power and sample size analysis
4. Draft the budget

Writing the grant involves six steps

5. Coordinate with grants management administrator
6. Write, revise, repeat
   (Internal reviewers can help)
7. Submit the grant and set reasonable expectations

1. Write the analysis plan

Write a data analysis plan that is aligned with the power analysis.

Remember to identify:

1. Clustering
2. Predictors
3. Covariates
4. Repeated Measures
5. Response

2. Conduct an aligned power and sample size analysis

Recall the procedure for projects with multiple aims:

1. Conduct an aligned analysis for each individually
2. Select the maximum sample size for all aims *if equally important*
3. Account for missing data

3. Draft the budget

The budget should include expenses such as materials, staff, recruitment of staff, data collection, and recruitment of participants.

The grants administrator can help estimate costs.

3. Draft the budget

Sample size may need to be recalculated in light of budget and feasibility constraints.

Examples:

It will take too long to recruit the required sample size.

It will cost too much to conduct original study design.

3. Draft the budget

Include a budget justification that explains the rationale for each item in the budget.

Other Direct Costs

Travel ($5,000 in Years 1-5). Two five-day/four-night trips ($2500 each) per year are requested for the PI and a Co-Investigator to attend scientific meetings to present the project data. The costs are estimated as follows:

Airfare	$ 1,000
Hotel	$ 920 ($230 per night)
Rental Car	$ 200 ($40 per day)
PerDiem	$ 190 ($38 per day)
Incidentals	$ 190
Total	$ 2,500

Supplies ($2,185 Yr ; $2,070 Yr 2; $1,955 Yr 3; $1,840 Yr 4; $1,725 Yr 5). Project supplies included in the budget will be used in the preparation and distribution of study communications and progress reports, and the compilation of research data and publications. Supplies include stationary, envelopes, resource and reference materials, computer diskettes, toner, and paper. The estimated cost is based on supply costs for similar studies and is calculated at a rate of $2300 X personnel FTE. Actual costs will be charged to this study.

4. Coordinate with grants management administrator

Meet with the grants management official at your home institution to discuss facilities, data sharing, and delegation of grant-writing tasks.

Grants management will draft many pieces of the grant independently, including the Resources and Facilities sections.

5. Write, revise, repeat

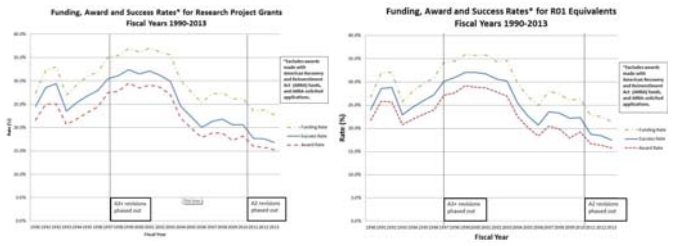
Finally, draft your grant and prepare for many rounds of review and revisions.

Plan for both internal and external reviewers.

Develop a time schedule for reviews.

6. Submit your grant and set reasonable expectations

The NIH accepts a very low percentage of grants submitted.



https://nexus.od.nih.gov/all/2014/03/05/comparing-success-award-funding-rates/

6. Submit your grant and set reasonable expectations

Why bother with good power analysis in the face of almost certain failure?

6. Submit your grant and set reasonable expectations

Sisyphus, in Greek mythology, was doomed to endlessly roll a boulder up a mountain.

Every time he came to the top, it rolled down again.

6. Submit your grant and set reasonable expectations

"I leave Sisyphus at the foot of the mountain. One always finds one's burden again... The struggle itself toward the heights is enough to fill a man's heart. One must imagine Sisyphus happy."

- Albert Camus

6. Submit your grant and set reasonable expectations

Despite the possibility of failure, every grant is a promise to do good science.

As researchers, we have an ethical obligation to study participants, other scientists, and society to plan studies that can accomplish their scientific goals without undue burden.

Accurate power and sample size analysis is an ethical obligation.