

Research and Evidence Webinar Series:

Power Analysis for Program Evaluation III: Applied Power Analysis

April, 2021

Welcome!

Power Analysis for Program Evaluation III: Applied Power Analysis



Dr. Lily Zandniapour

Research and Evaluation Manager

Office of Research and Evaluation, AmeriCorps

Introductory Remarks

Power Analysis for Program Evaluation III: Applied Power Analysis



Dr. Carrie Markovitz

Principal Research Scientist

NORC

Speaker

Power Analysis for Program Evaluation III: Applied Power Analysis



Dr. Eric Hedberg

Senior Data Scientist

NORC

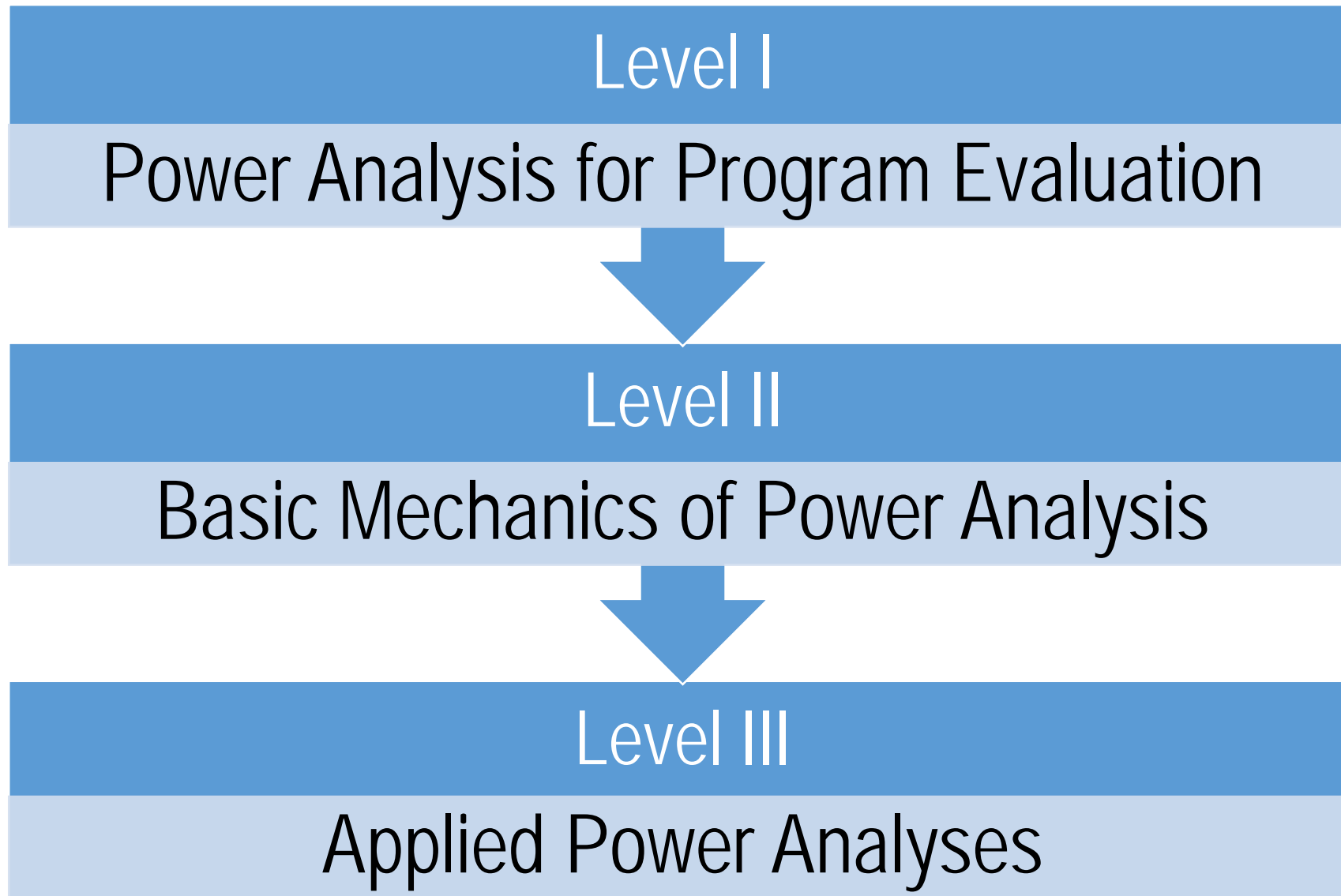


AmeriCorps



Power Analysis for Program Evaluation Level III: Applied Power Analysis

Overview of Courses (Three Levels)



- Intended for program staff and evaluators working with statisticians
- How to conduct power analyses
 - Parameter values for a power analysis
 - Estimating effect sizes
 - Examples for common designs
- How to write-up (or read) a power analysis

Level I: Defining and Understanding Statistical Power

- Intended for All Audiences: program staff, funders, Program Officers, internal or external evaluators, and third-party evaluation/evidence reviewers
- Introduces the concept of power
- Explains why power is important for evaluation planning

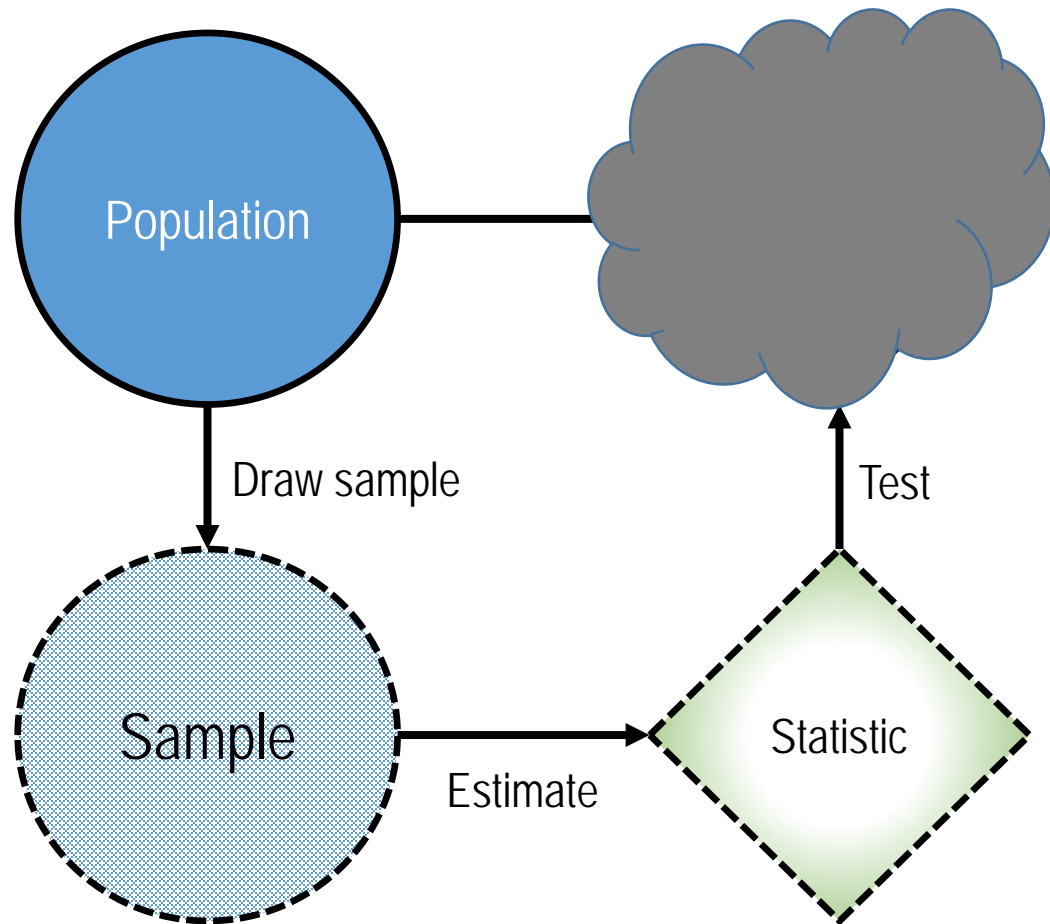
Level II: Basic Mechanics of Power Analysis

- Intended for program and evaluation staff
- Type I and Type II errors
 - How each error threatens a study
 - Power is complement of Type II error

Review: What is power?

- Power helps determine how large a sample size you need in order to obtain reliable evidence of program impacts
- *The power of a statistical test is the **probability** that it will yield statistically significant results, Cohen (1988)*
- *“statistically significant results”* are based on:
 - The design of the study (how groups are formed, how observations are selected, and how the data are analyzed)
 - The amount of data
 - How the data are analyzed
 - The achieved size of the impact
 - The desired significance level

Statistics for Impact Studies



- Impact studies want to infer the existence and size of the impact for the **population** based on **statistics** produced from **sample** data
- We will not know for sure whether a program will be successful in the future or for non-study participants
- We make an informed decision based on our sample from which we infer to the population

Type I and Type II errors

In deciding whether there is an impact, we are making an inference about something we do not ever know. There are four possible outcomes:

We can be right in two ways:

1. There is no impact, and we conclude so
2. There is an impact, and we conclude so

But we can be wrong in two ways:

3. There is no impact, BUT we conclude there is an impact (Type I error; α)
4. There is an impact, BUT we conclude there is not an impact (Type II error, β)

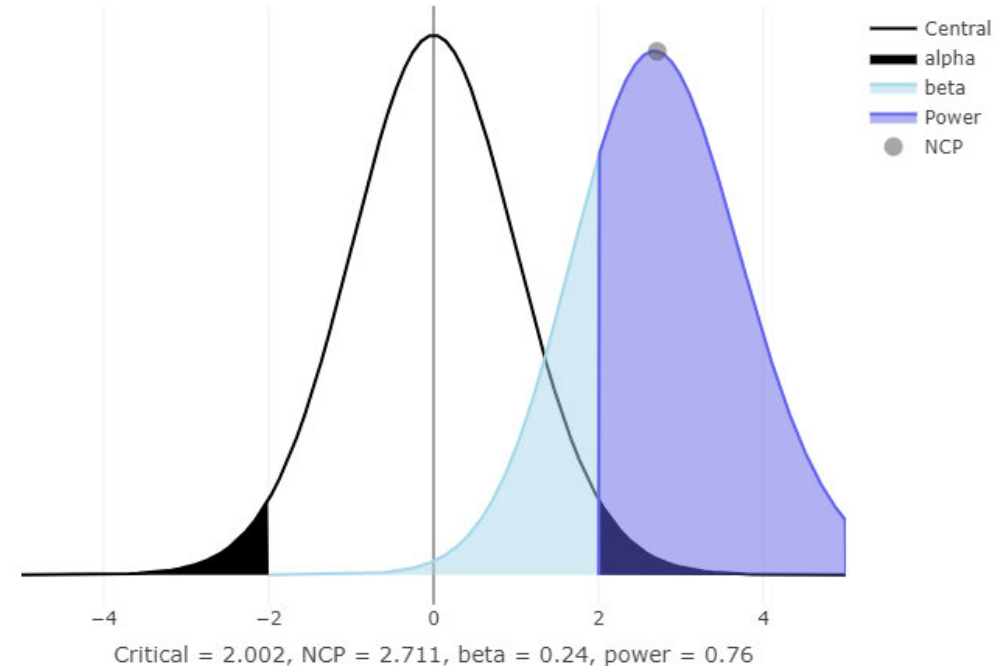
<u>What is True?</u>	<u>Study Conclusion</u>	
	Conclude Impact Exists	Conclude Impact Does Not Exist
Impact Exists	Conclusion is Correct	Type II Error β
Impact Does Not Exist	Type I Error α	Conclusion is Correct

- Power will increase as you increase Type I error (α) (increase your tolerance for wrongly detecting an impact)
- More likely to have a “significant” result if you increase your Type I error
- Finding a balance:
 - Want Type I error rate to be reasonably lower (typically $\alpha = .05$)
 - Want power to be high enough to minimize Type II error (β), so won't fail to detect an effect



Type I and Type II error work together to estimate power based on possible outcomes from your one sample draw

- The black curve centered on 0 is the null distribution= possible results if there is no impact
- The black shade is Type I error
- The blue/purple curve is another distribution based on the expected result
 - Blue are possible samples that are not significant (Type II error)
 - Purple are possible samples that are significant (Power)



Why conduct a power analysis for your study?



- Power analysis gives researchers a chance to determine if their sample design is adequate to detect the expected impact before the study (Minimize a Type II error)
- Power analysis encourages research teams to think critically about and explore
 - Expected Impacts of their intervention
 - The design of their study (forming groups and selecting/sampling observations) and how it impacts the analysis plan
 - Whether the analysis plan is feasible (does the design produce enough data for the analysis?)
 - Evaluation budget
- A power analysis can help programs effectively use and target resources

BEFORE DATA COLLECTION!

- Power analyses are only informative and helpful prior to data collection
- When there is no power analysis and the results are not statistically significant:

Perhaps there is no impact of the intervention
OR

Perhaps the study was underpowered to detect the actual effect

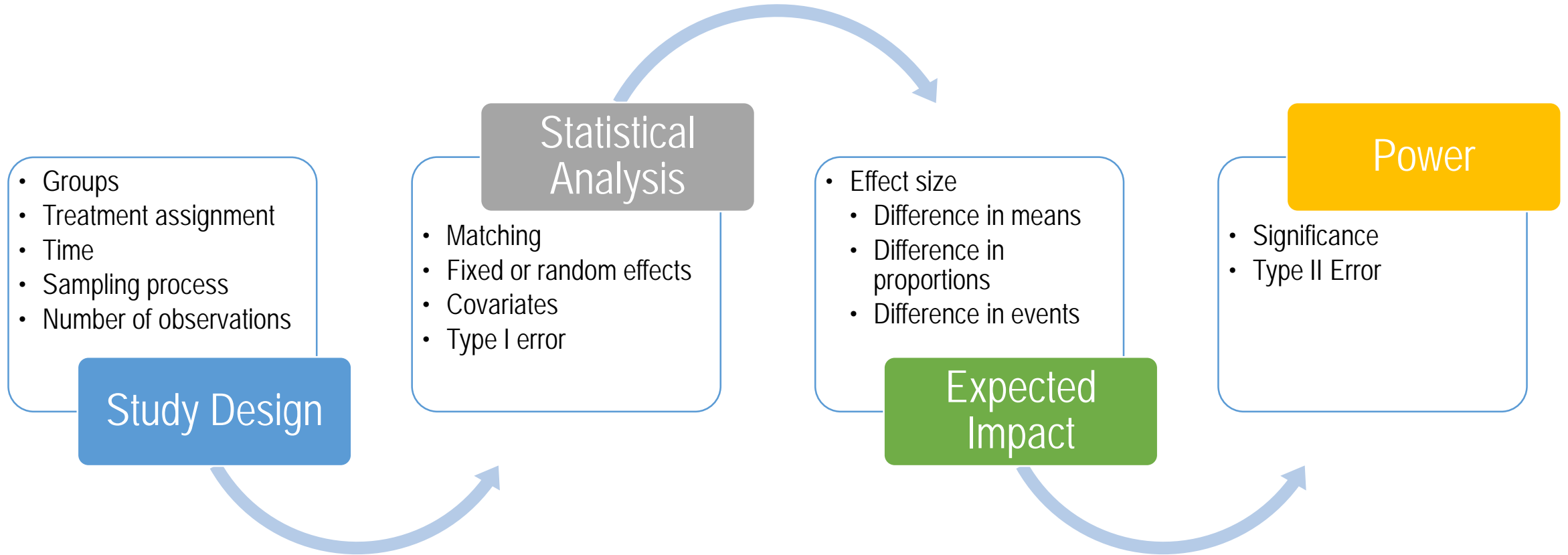
Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of educational and behavioral statistics*, 30(2), 141-167.

How do I design a high powered study?



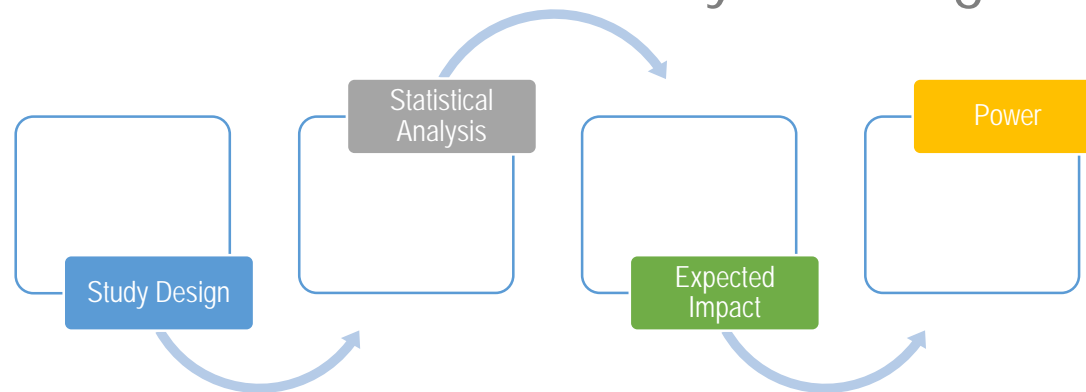
- Plan to collect the appropriate amount of data that satisfies the
 - Study design
 - Statistical analysis
 - Expected impact

Design, Analysis, and Impact → Power



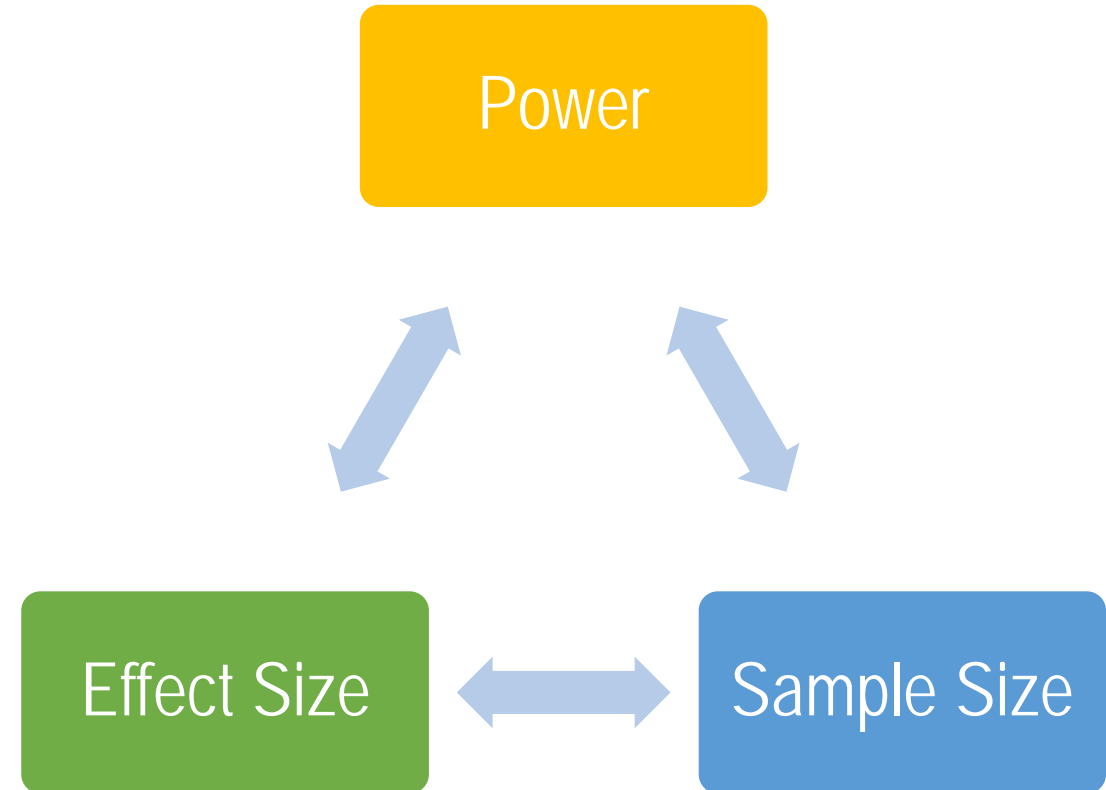
Well-powered studies coordinate

- Power is a direct function of several factors (all related):
 - Study design
 - Statistical analysis
 - Expected impact
- Power analysis examines this system and asks either:
How much **sample** do I need [given X power and Y impact]? OR
What is the smallest **impact** I can detect [given X power and Z sample]?
- Linking these questions is the Statistical Analysis Design



What is a power analysis?

- A calculation that helps determine if a study has an adequate chance to detect a statistically significant effect (if one truly exists)
- Power analysis is based on the relationship between power, sample size, and effect size (assume two of the elements and calculate the third)

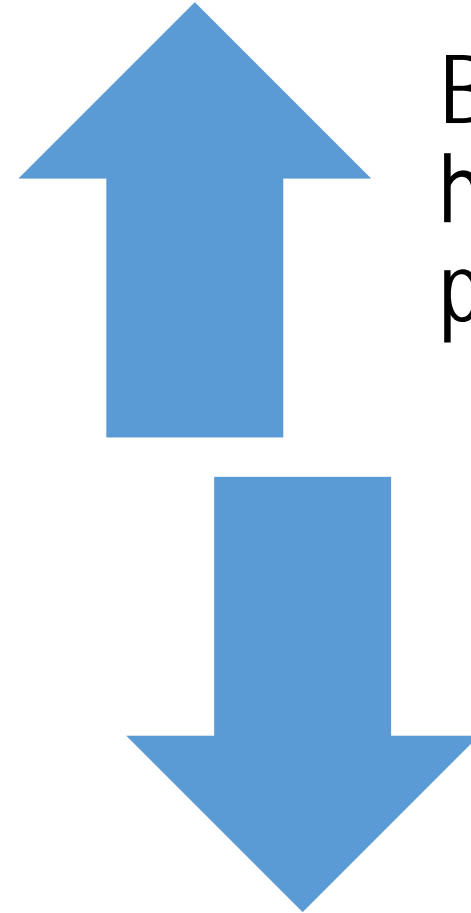


Power, sample size, and effect size



Big effect sizes have more power

Small effect sizes have less power



Big samples have more power

Small samples have less power

Type of power analysis depends on several factors:

- Type of outcome
 - Linear outcomes (continuous)
 - Categorical outcomes (limit number of values)
- Study Design (Organization of the study)
 - Group formation: How (and if) treatment and comparison groups are formed (random, matched, or natural)
 - Sampling: How study participants are selected from the population of interest
 - Sample size: How many times study participants are observed
 - Other measures: What other variables are collected or combined with outcome measures to be used as controls?
 - Analysis Plan: how the data are analyzed to test hypotheses

- Linear
 - Numbers with meaningful differences
 - If skewed, can be transformed to normalize numbers (e.g., logging the value)
 - Examples: test scores, money (e.g., income, wages, savings), biometrics (e.g., BMI, vision, cognition)
- Unordered Categories
 - Has a limited number of possible values
 - Examples:
 - “Yes/No” have two related, opposite values: above or below grade level; employed or not
 - Unordered categories have more than two independent values: matriculated, graduated, or dropped out

- Ordered categories
 - Analysis of ordered categories can be similar to linear or categorical
 - Examples: satisfaction questions; strongly agree....strongly disagree

NOTE: Typically, more data are needed to provide adequate power for categorical outcomes (ordered or unordered) than for linear outcomes

Group assignment is the process of assigning a treatment (or program) status to some study participants, and a control (or comparison) status to the rest. Group assignment can occur in many ways, e.g.,

- Random: a random process (usually computer) results in study participants being assigned to groups
- Matched: given a treatment group, a comparison group is formed from other population members that seeks to be similar, on average, to the treatment group
- Threshold: treatment is assigned based on the value of another assignment variable (e.g., being 200 percent of poverty line or lower)
- Natural: the population is comprised of some treated, some not, through other processes

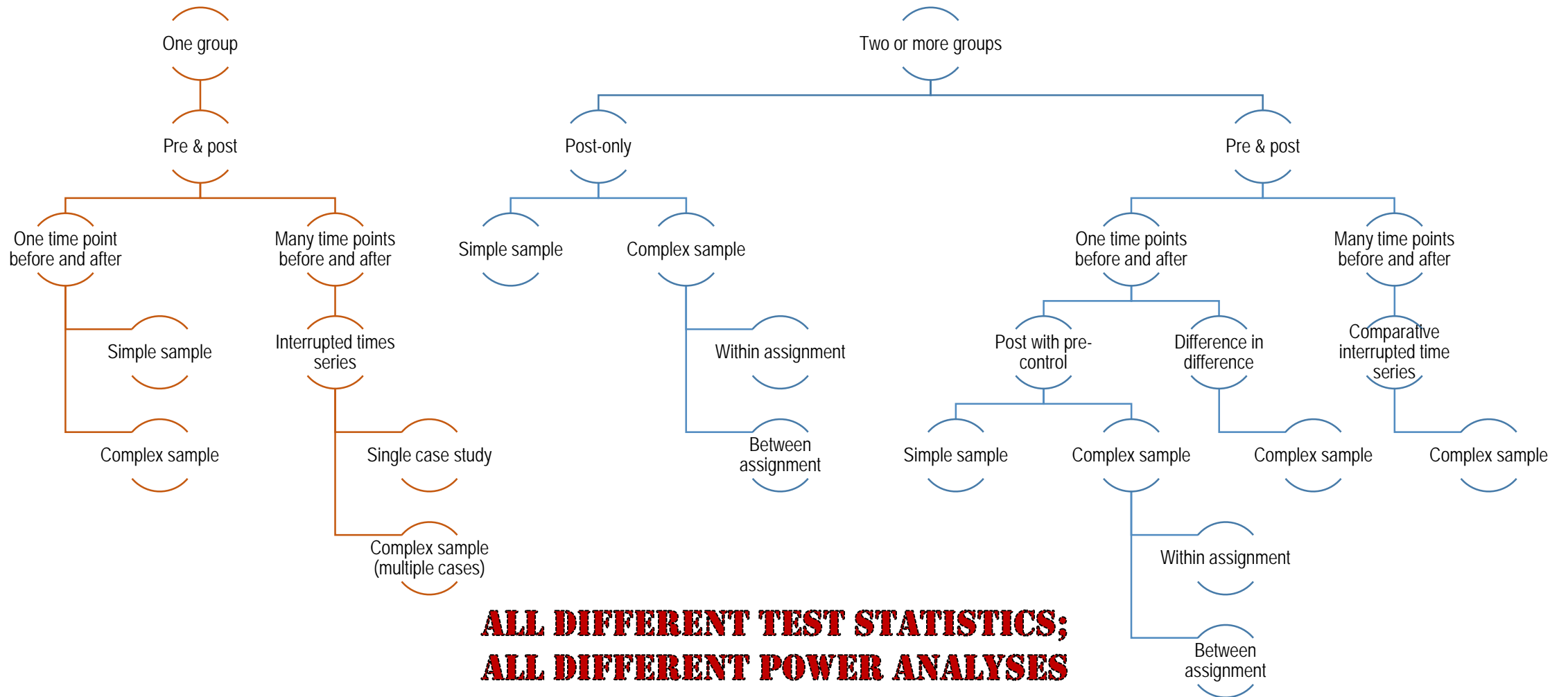
Sampling is a process that results in only a portion of your population being included in your data. There are many ways to pick study participants, broadly conceptualized as

- Simple (random) sample
 - Every population member has an equal chance of being selected, and their selection is unrelated to the selection of other population members
- Complex sample (multilevel)
 - Population members are organized in clusters (e.g., neighborhoods)
 - A first-stage random sample of clusters is selected, then
 - A second-stage sample of units within clusters may be selected.
- In complex samples, group assignment can happen at either level, which impacts analysis choices

Evaluators typically collect additional data beyond the outcome. These measures can include:

- Measures of the outcome variables prior to the treatment period
 - This creates a “pre” and “post” set of measures
 - When there is a comparison group, this allows for an analysis of post measures that hold constant “pre” values, and/or comparing the difference in the pre-post differences between treatment and control groups (so-called “difference-in-difference” analyses)
- Many measures of the outcome variables during the treatment period
 - This creates a set of “repeated” measures for more complex analysis of “growth”
- Measures of other related factors such as demographic group membership

Groups, Samples, and Measures



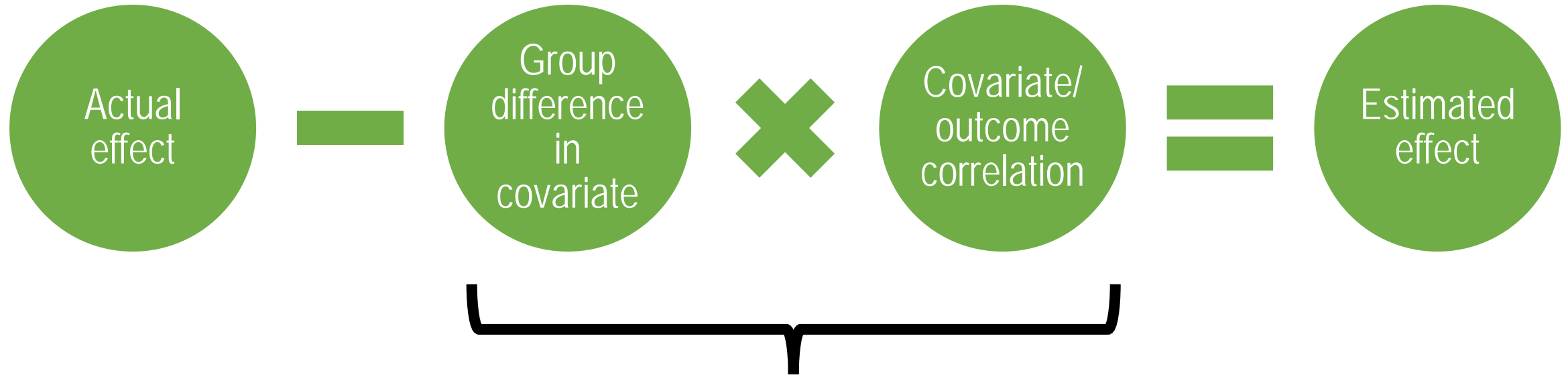
The analysis procedure is a natural progression from the other elements of the study design

- Type of outcome
 - Linear outcomes can use a least-squares-errors method like regression or ANOVA
 - Categorical outcomes can use a generalized model method like (ordered) logistic regression, probit, or count model
- Group formation
 - How treatment and control groups are formed determines other factors related to analysis choices, and what inferences can be made from the analysis
- Type of sample
 - The methods for outcome types each have extensions for complex sampling procedures, estimating fixed and random effects on parameters, or producing population average models
- Set of measures
 - Multiple measures across time (even pre-post) or other control variables extend the analysis options beyond t-tests or ANOVA to regression estimators

Using covariates in your analysis procedure (pre-tests, demographics, other factors) can either improve or degrade power

- Test statistics that summarize the data outcomes are based (in part) on the amount of outcome variation that is “left over” after being explained with control variables
- Thus, pre-tests and other information can reduce/explain “left over” variation, making test statistics larger when comparing groups
- How the groups are defined affects the usefulness of covariates
 - Randomization into groups or matching cases or clusters within groups can improve the chance that each group is similar (less unexplained variation)
 - Using control variables that are correlated with the outcome (i.e., multicollinearity) can reduce test statistic size and counteract any benefits to variance reduction

Covariates and effect estimation

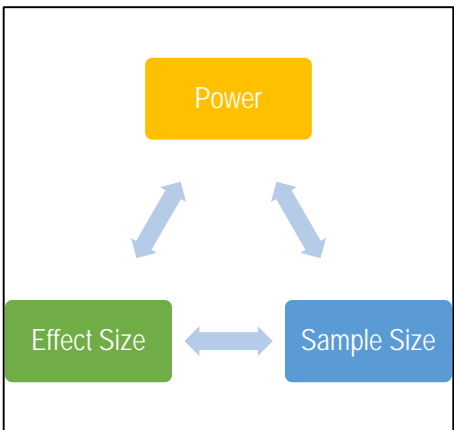
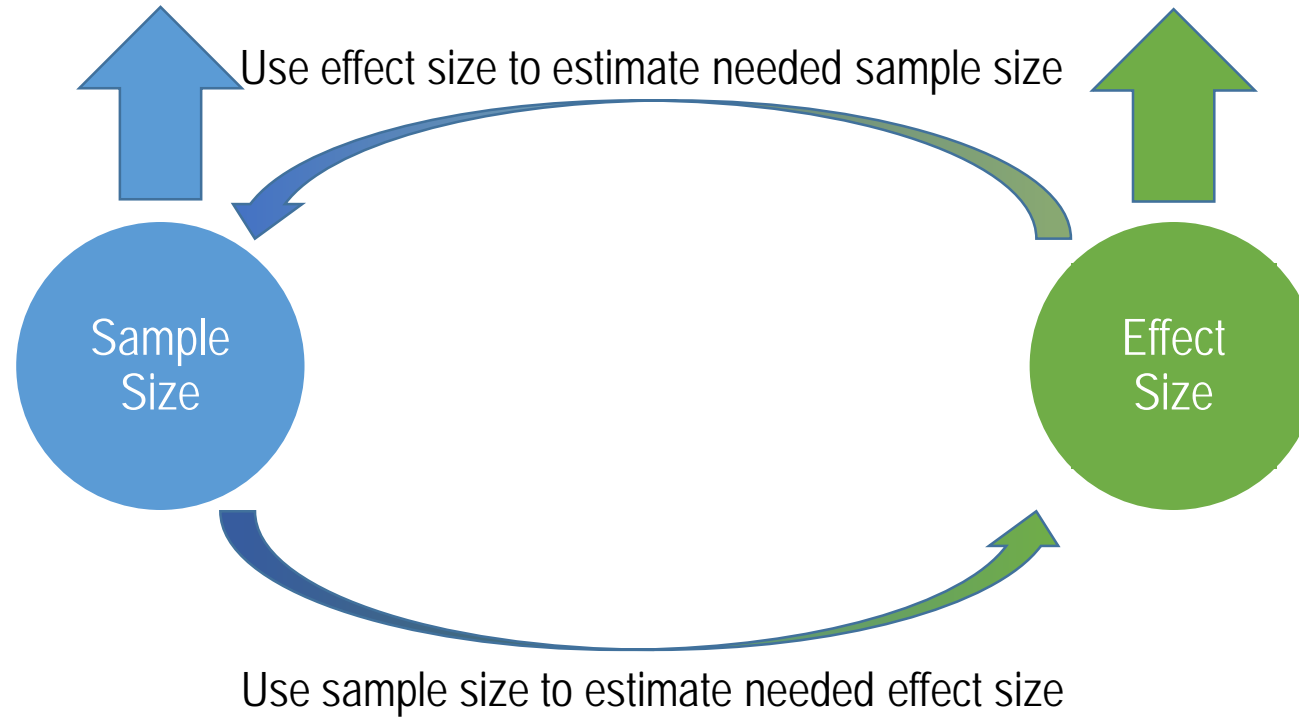
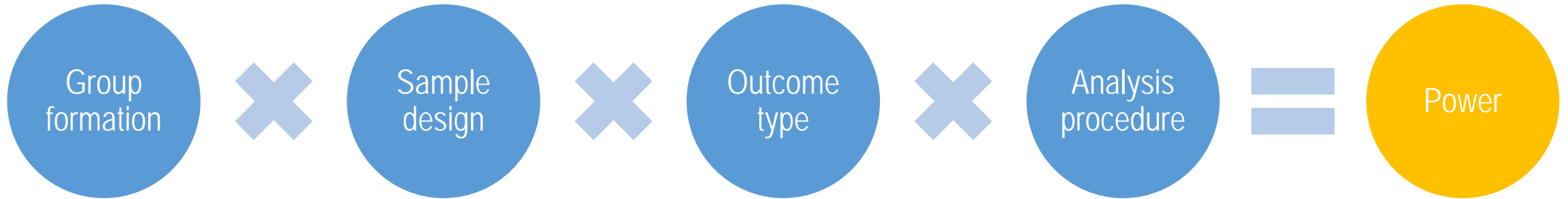


If either of these are zero, actual difference = estimated difference
Randomization or matching helps render group difference zero

Adapted from:

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology*, 34(4), 383.

Inputting sample size and effect size



How to interpret an effect size?

- Effect size expectations vary by outcome and research field
 - Social systems are vastly complex, making it much harder to demonstrate an effect; in these cases, even small impacts can be important and meaningful
 - An environmental program to remove invasive species will typically have larger effect sizes than a program to help prevent high school dropouts
- Cohen's effect size interpretation (.2 is "small," .5, is "medium," .8 is "big")
 - Sometimes is no better than guessing (analogous to "shirt sizes"), be sure to use the literature to determine what to expect
- Important to collect the data needed to measure impact with precision

How to perform a power analysis

- Computations for power analysis are straightforward for basic designs and analyses, but covariates, clustering, matching, weighting, and non-response adjustments all increase complexity
- There are software that provide power estimates based on your study design and analysis plan
 - Any major stats package will work with the correct formula (SPSS, Stata, SAS, R, Python)
 - Power analysis software (PowerUP! and G*Power)

- STUDY DESIGN: Number of groups, number of data points, sampling method, etc.
- POWER: Assume a conventional power level of .80 ($1 - \beta$, where β is Type II error rate)
- SIGNIFICANCE: Assume a Type I error rate for $\alpha = .05$ (two-tailed test)
- EFFECT SIZE: Estimate the effect size for the intervention study group based on previous studies or pilot data
- SAMPLE SIZE: Estimate the expected sample based on existing program data or estimate the sample needed to detect an expected effect size

- Strategies for using the literature to make effect size assumptions:
 - Find as many studies of similar programs as possible based on the type of intervention, population served, and outcomes measured; may consider other factors such as program location, staff (volunteer vs. professional), etc.
 - Select studies that include a high-quality comparison/control group (randomized controlled trials or quasi-experimental design studies with a matched comparison group)
 - Utilize meta-analyses– a systematic assessment of the results from multiple previous rigorous studies; often include effect sizes for similar studies

A school-based math intervention program for middle school age youth wants to determine if their program is having an impact on students' standardized math scores.

- POWER: Assume a conventional power level of .80 ($1 - \beta$, where β is Type II error rate)
- SIGNIFICANCE: Assume a Type I error rate for $\alpha = .05$ (two-tailed test)
- EFFECT SIZE: Based on a pilot test of the program, the expected effect size is 0.2
- SAMPLE SIZE: Number of students served by program is 500 students and expect another 500 students to be assigned to a control group

- Students in schools
 - Simple random sample
 - No covariates
 - With covariate
 - Complex sample (between randomization)
 - No covariates
 - With covariates
 - Complex sample (within randomization, random effects)
 - No covariates
 - With covariates
 - Complex sample (within randomization, fixed effects)
 - No covariates
 - With covariates

- PowerUp! Available at <https://www.causalevaluation.org/>
 - Excel
 - R
- Handout of similar examples and associated write-ups are available on the training session page
 - “Power Course Level III Handout”

Summary: Impact of Design Variations



Design	MDES (without Covariates)	MDES (with Covariates)
Simple Random sample	0.18	0.13
Complex sample with school assignment	0.54	0.33
Complex sample with student assignment (random effects)	0.21	0.17
Complex sample with student assignment (fixed effects)	0.18	0.13

How does the MDES from different randomized designs compare with the expected impact of .2 for studies powered to .8 for a two-tailed test with $\alpha = .05$?

A power analysis description should include

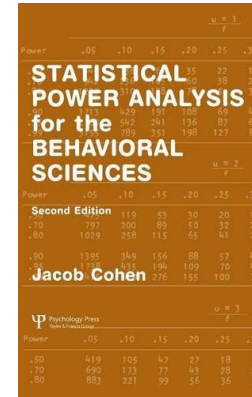
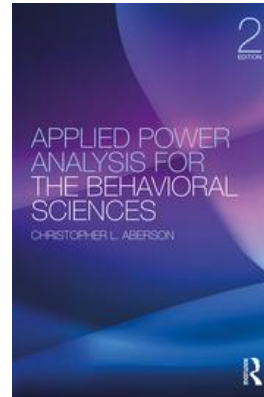


1. The study design (one group, two groups, number of data collection points)
 - The sample design: number of observations, clustering and observations per cluster
 - How groups are assigned (within/between clusters)
2. The statistical procedure that will be used to analyze the data
3. The assumptions and citations for them
 - The assumed effect size and where it came from (another study, pilot data)
 - Other parameters, as necessary
4. The power analysis procedure used (citation, software, formula)
 - Must match the analysis plan
5. Sensitivity to small changes to assumptions
6. Results of power analysis

- Power is the chance to find an estimated impact “statistically significant”
- Power analysis is an informed argument that the anticipated study has a high chance to yield the assumed effect
- Power analysis depends on several factors including, group formation, sampling method, data analysis, and control variables
- Power analysis must occur before the study is conducted
- Computations for power analysis can be straightforward for basic designs and analyses, but become more complex when including covariates, clustering, matching, and weighting
- There are several existing softwares that provide power estimates

- PowerUP! (with clustering available)
 - <https://www.causalevaluation.org/>
 - Excel and R versions
 - Impacts, moderation (subgroup), and mediation analyses
- Other designs and analyses (without clustering): G*Power
 - <https://www.gpower.hhu.de>
 - Regression/Correlation, ANOVA, Probability Models

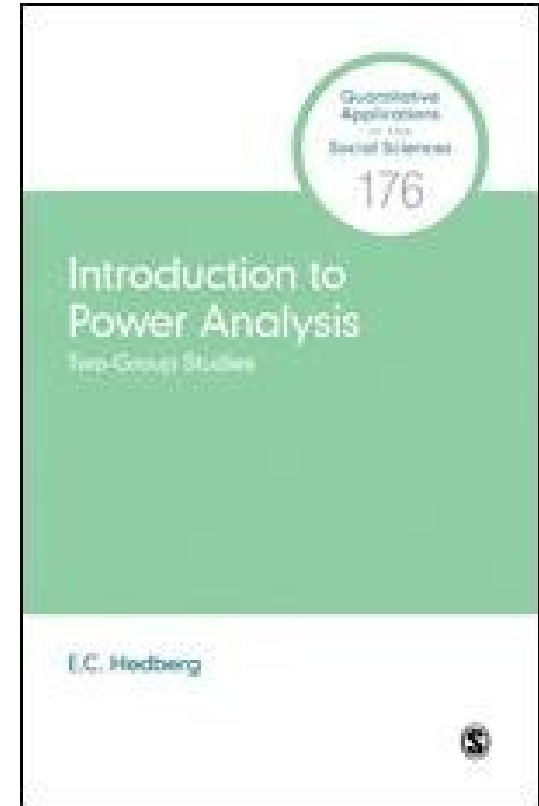
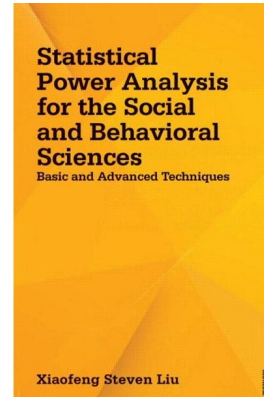
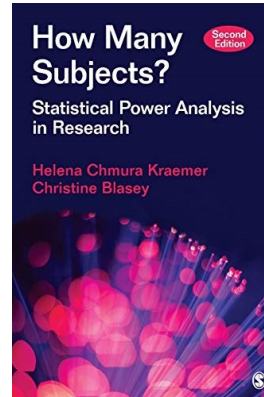
- Aberson, C. L. (2011). *Applied power analysis for the behavioral sciences*. Routledge.
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.
- Hedberg, E. C. (2017). *Introduction to Power Analysis: Two-group Studies* (Vol. 176). Sage Publications.
- Kraemer, H. C., & Blasey, C. (2015). *How many subjects?: Statistical power analysis in research*. Sage Publications.
- Liu, X. S. (2013). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. Routledge.
- Many more papers and blogs exist, likely specific to your outcome



		$\alpha = .1$					
		.05	.10	.15	.20	.25	.30
Power	.80	35	25	16	8	5	3
	.70	37	27	18	10	7	4
	.60	40	30	20	12	8	5
	.50	44	34	24	15	10	6
	.40	49	39	29	19	13	8
	.30	55	45	35	24	17	11
	.20	62	52	41	28	20	13
	.10	72	61	49	35	25	16
	.05	78	69	57	43	30	19
	.01	90	81	69	55	39	25
	.005	93	85	73	59	42	27
	.001	96	89	77	63	46	30
	.0005	97	91	79	65	47	31
	.0001	98	93	81	67	48	32
	.00005	98	93	81	67	48	32
	.00001	99	94	82	68	49	33
	.000005	99	94	82	68	49	33
	.000001	99	94	82	68	49	33
	.0000005	99	94	82	68	49	33
	.0000001	99	94	82	68	49	33
	.00000005	99	94	82	68	49	33
	.00000001	99	94	82	68	49	33
	.000000005	99	94	82	68	49	33
	.000000001	99	94	82	68	49	33

		$\alpha = .2$					
		.05	.10	.15	.20	.25	.30
Power	.80	115	83	70	50	35	24
	.70	121	89	76	55	39	27
	.60	129	95	81	60	43	29
	.50	139	102	87	65	47	31
	.40	150	110	93	70	50	33
	.30	163	119	100	75	54	35
	.20	178	129	108	80	58	37
	.10	195	140	117	85	62	39
	.05	214	152	127	90	66	41
	.01	235	165	137	95	70	43
	.005	248	177	147	100	74	45
	.001	263	189	157	105	78	47
	.0005	270	193	160	107	79	48
	.0001	273	195	161	108	80	48
	.00005	274	196	162	108	80	48
	.00001	275	197	163	109	80	48
	.000005	275	197	163	109	80	48
	.000001	275	197	163	109	80	48
	.0000005	275	197	163	109	80	48
	.0000001	275	197	163	109	80	48
	.00000005	275	197	163	109	80	48
	.00000001	275	197	163	109	80	48
	.000000005	275	197	163	109	80	48
	.000000001	275	197	163	109	80	48

		$\alpha = .3$					
		.05	.10	.15	.20	.25	.30
Power	.80	145	105	87	60	40	27
	.70	152	112	93	65	43	29
	.60	161	120	100	70	46	31
	.50	172	129	107	75	49	33
	.40	185	139	115	80	52	35
	.30	200	150	123	85	55	37
	.20	217	162	132	90	58	39
	.10	236	175	141	95	62	41
	.05	257	189	151	100	66	43
	.01	279	203	161	105	70	45
	.005	292	216	170	110	74	47
	.001	307	229	179	115	78	49
	.0005	313	234	182	117	79	48
	.0001	315	236	183	118	80	48
	.00005	316	237	184	118	80	48
	.00001	316	237	184	118	80	48
	.000005	316	237	184	118	80	48
	.000001	316	237	184	118	80	48
	.0000005	316	237	184	118	80	48
	.0000001	316	237	184	118	80	48
	.00000005	316	237	184	118	80	48
	.00000001	316	237	184	118	80	48



Closing Remarks

Power Analysis for Program Evaluation III: Applied Power Analysis



Dr. Carrie Markovitz

Principal Research Scientist

NORC