Research and Evidence Webinar Series:

# Power Analysis for Program Evaluation II: Basic Mechanics of Power Analysis

April, 2021

AmeriCorps

# Welcome!

Power Analysis for Program Evaluation II: Basic Mechanics of Power Analysis

**Dr. Lily Zandniapour**

***Research and Evaluation Manager***

***Office of Research and Evaluation, AmeriCorps***

# Introductory Remarks

Power Analysis for Program Evaluation II: Basic Mechanics of Power Analysis

**Dr. Carrie Markovitz**

***Principal Research Scientist***

***NORC***

# Speaker

Power Analysis for Program Evaluation II: Basic Mechanics of Power Analysis

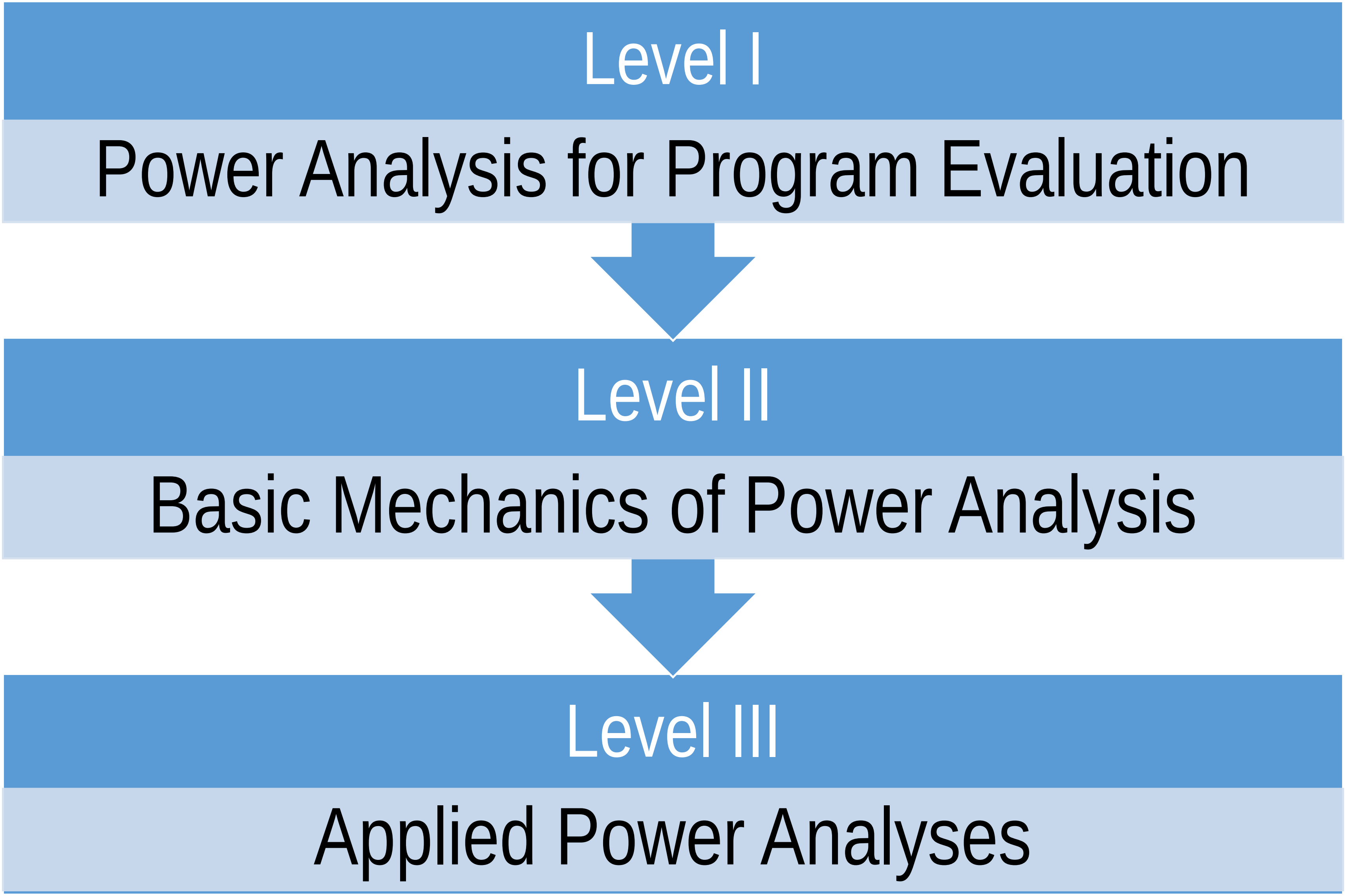
**Dr. Eric Hedberg**

***Senior Data Scientist***

***NORC***

# Power Analysis for Program Evaluation
# Level II: Basic Mechanics of Power Analysis

# Overview of Courses (Three Levels)



Level I

Power Analysis for Program Evaluation

Level II

Basic Mechanics of Power Analysis

Level III

Applied Power Analyses

# Level II Overview

- Intended for program and evaluation staff
- Introduction to power and power analysis
- Additional technical material for program and evaluation staff
- Type I and Type II errors
  - How each error threatens a study
  - Power is complement of Type II error

# Level II is the second part of a three-part series

AmeriCorps

## 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 III: Applied Power Analysis

- Intended for program staff or evaluators working with statisticians
- How to conduct power analyses
  - Parameter values for a power analysis
  - Examples for common designs
  - How to use the literature to inform specifications
- How to write-up (or read) a power analysis

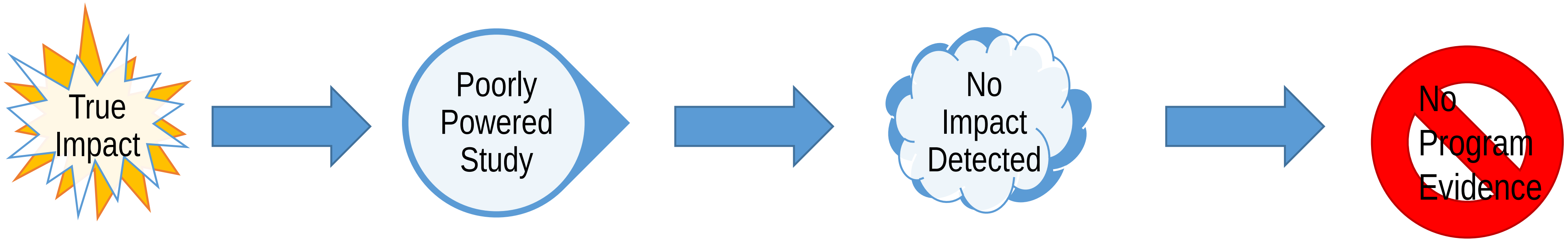
# Review: What is power?

Power helps determine how large a sample size you need in order to obtain reliable evidence of program impacts.

- Power is about planning
- Power analysis sections are estimates about the future
- Power is about using evaluation resources wisely
- Power should be estimated prior to conducting a study (i.e., power analysis)

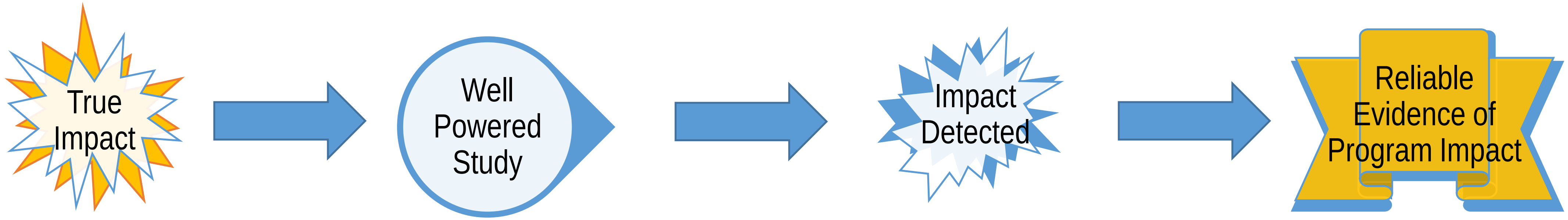
# Why power is important for program evaluations

## Results of a poorly powered study (i.e., Type II Error)



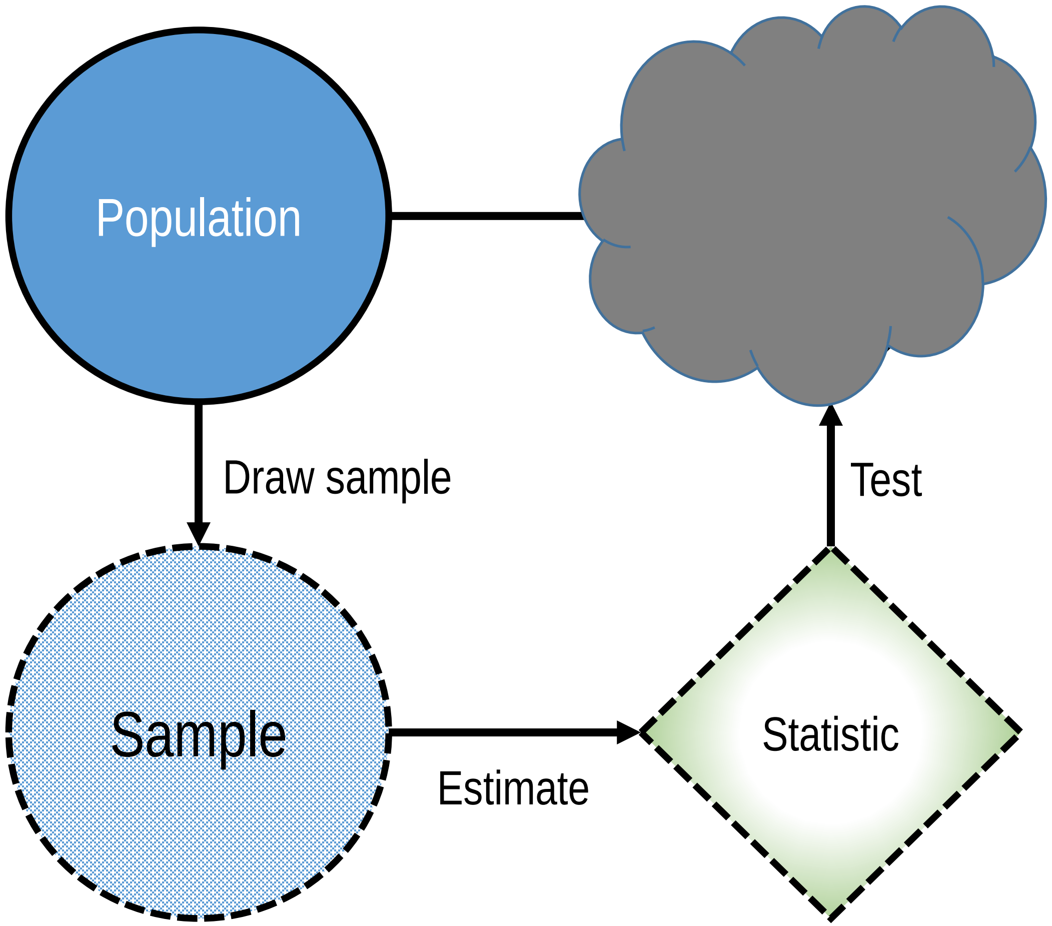
True Impact → Poorly Powered Study → No Impact Detected → No Program Evidence

## Results of a well powered study



True Impact → Well Powered Study → Impact Detected → Reliable Evidence of Program Impact

# 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; $\alpha$)
4. There is an impact, BUT we conclude there is not an impact (Type II error, $\beta$)

| What is True? | Study Conclusion | |
|---|---|---|
| | Conclude Impact Exists | Conclude Impact Does Not Exist |
| Impact Exists | Conclusion is Correct | Type II Error<br>$\beta$ |
| Impact Does Not Exist | Type I Error<br>$\alpha$ | Conclusion is Correct |

# What is Type I Error?

- Probability of wrongly detecting an impact (Concluding impact exists when it does NOT)
  - Often called $\alpha$ (alpha)
  - Convention is .05
- P-value is the chance of our test-statistic (or a more extreme one) occurring ($\alpha$)
  - If p-value is less than $\alpha$, result is considered "statistically significant"

Note: The *p*-value is **not** a measure of practical significance, so "statistically significant" is a statement about the probabilities of your conclusion, nothing more

NOTE: Active discussion underway for lowering $\alpha$ convention to .005 to combat *p*-hacking and other human errors (see Benjamin, and 71 other authors. (2018). Redefine statistical significance. *Nature; Human Behaviour*, *2*(1), 6-10).

# What $\alpha$ means for a set of studies

- The data we obtain only represents one possible sample
- $\alpha = .05$ means that 1 out of every 20 samples will be significant, EVEN WHEN THERE IS NO IMPACT!
- Accepting .05 chance of Type I error means that for a set of studies about a program that doesn't work, 1 in 20 will be "statistically significant"
- That means any one study has a 1 in 20 chance of being significant no matter what:
  - 1 in 20 even if there is no impact
  - 1 in 20 regardless of sample size (large or small)
  - 1 in 20 even if the data are analyzed poorly or well

# Role of Type I error in assessing the existence of impacts

Statistical tests focus on providing information about the existence of an impact through “null-hypothesis significance testing” (NHST)*

1. Assume there is no impact (null hypothesis)
2. Decide on an acceptable chance of being wrong (**Type I Error**)
3. Analyze/summarize data by calculating a single test statistic
4. Assess how likely it is to observe these results (data): the *p*-value ($\alpha$)
5. Conclusion: If the likelihood of observing these results (data) is less than the chance of being wrong, then reject null hypothesis that there is no impact.

*This approach is not without its issues, and other methods (Baysian) provide answers to different questions about impact. However, the bulk of evaluation research is performed under this paradigm. See, Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): what do the textbooks say?. *The Journal of Experimental Education*, *71*(1), 83-92.

# Critical values

- We think a test is “statistically significant” if the p-value is < .05
- Each p-value is associated with a “critical value” which marks the point on the distribution where the probability of error beyond this point is considered “small”
- Two things to note:
  1. As $\alpha$ (Type I error) reduces, critical values get larger
  2. The critical value for $\alpha$ on a one-tailed test is the critical value for $2 \times \alpha$ for a two-tailed test
     - i.e., one-tailed tests increase Type I error

Examples of critical values for a standard-z test

| $\boldsymbol{\alpha}$<br>(Type I error) | Critical value<br>(one-tail) | Critical value<br>(two-tails) |
|---|---|---|
| 0.1 | 1.28 | 1.64 |
| 0.05 | 1.64 | 1.96 |
| 0.01 | 2.33 | 2.58 |
| 0.001 | 3.09 | 3.29 |

# What is Type II error?

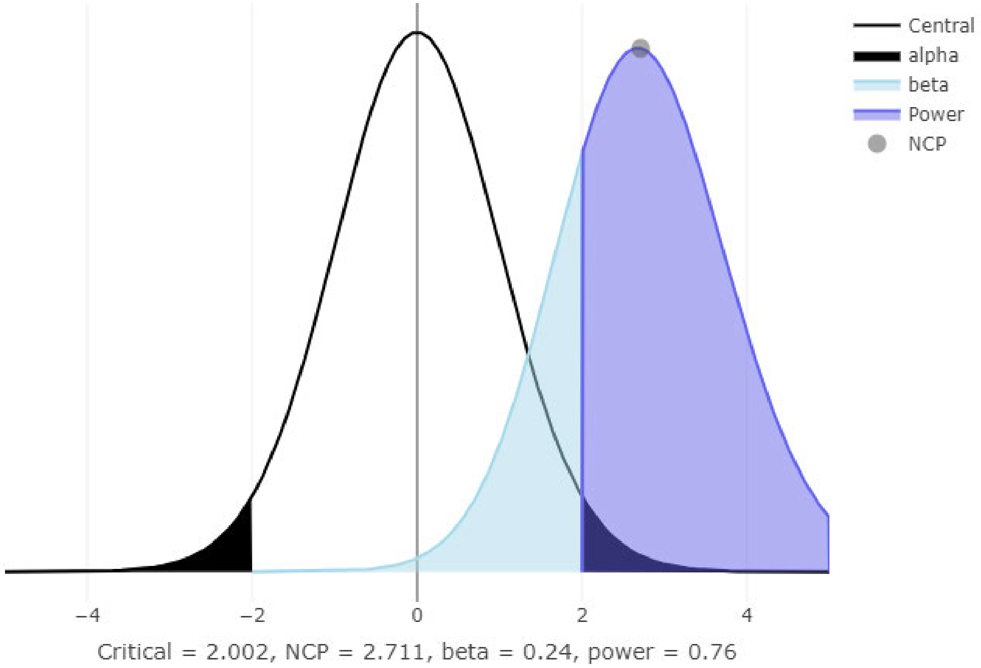
- In addition to Type I error, there is a second possible error, **Type II error**, symbol = $\beta$
  - Type II error is the probability that you fail to conclude that an impact exists when it is REAL
- This requires a different probability curve
  - This "alternative distribution" is centered on the expected impact
  - The area of the alternative distribution **before** the critical value represents possible samples that **do not** yield significant results
- Power = $\mathbf{1-\beta}$

# Power and errors

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

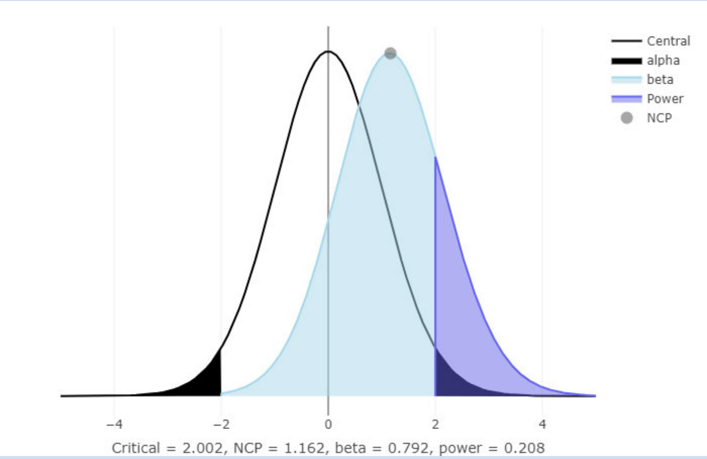
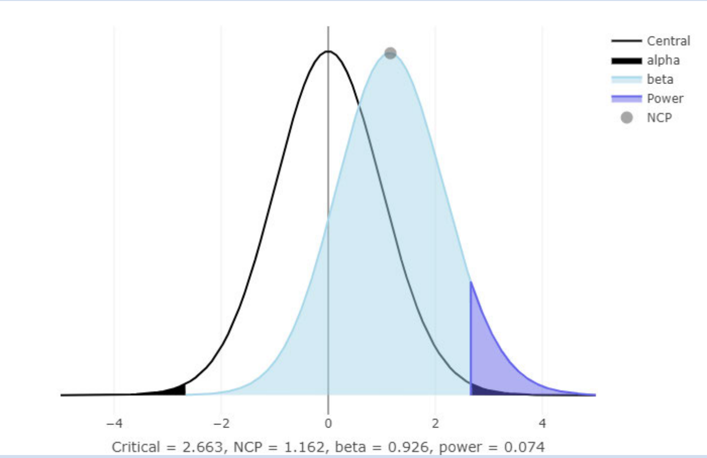
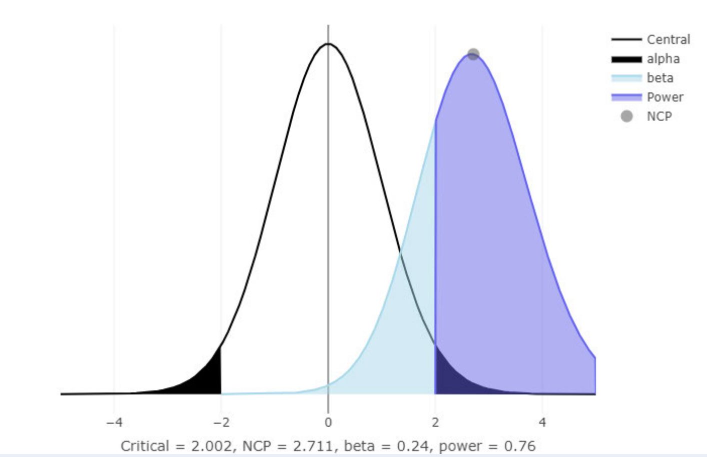
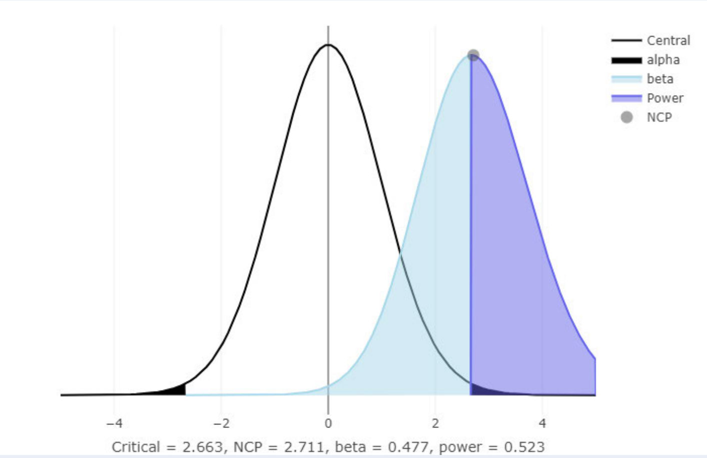
- The blue/purple curve is another distribution based on the expected result
  - Blue are possible samples that are not significant (Type II error)
  - Purple are possible samples that are significant (Power)



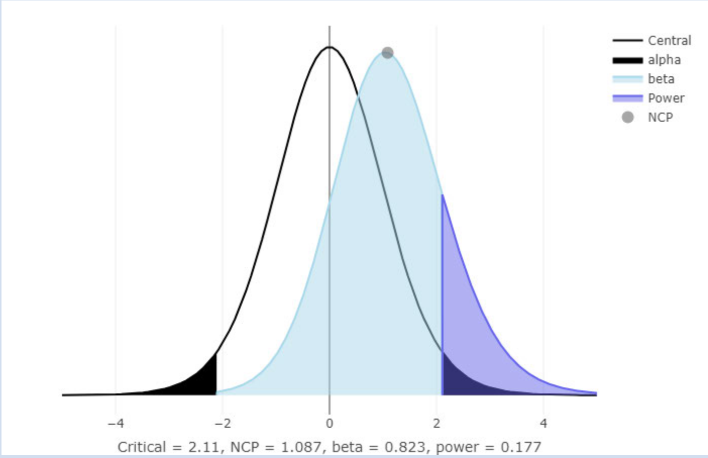
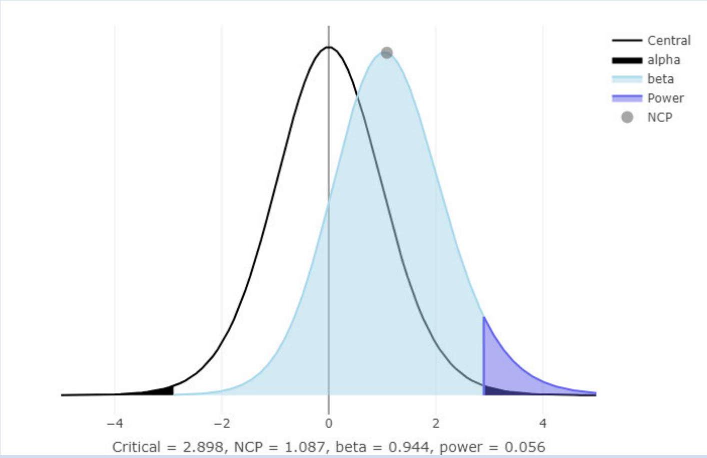
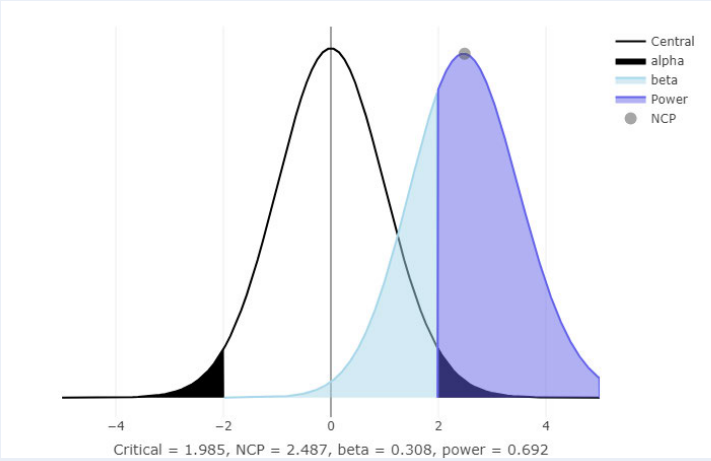
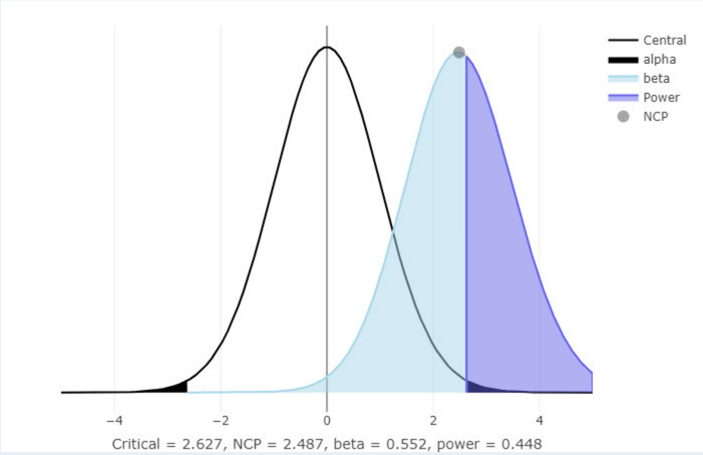
# Power and Type I error ($\alpha$)

- Power will increase as you increase Type I error ($\alpha$) (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 ($\beta$), so won’t fail to detect an effect

# Power (purple shade) for different impacts and $\alpha$ (but same sample size)

| Impacts | $\alpha = .05$ | $\alpha = .01$ |
|---|---|---|
| Small |  |  |
| Large |  |  |

# Power (purple shade) for different samples and $\alpha$ (but same impact)

| Sample | $\alpha = .05$ | $\alpha = .01$ |
|---|---|---|
| Small | <br> <br>Critical = 2.11, NCP = 1.087, beta = 0.823, power = 0.177 | <br> <br>Critical = 2.898, NCP = 1.087, beta = 0.944, power = 0.056 |
| Large | <br> <br>Critical = 1.985, NCP = 2.487, beta = 0.308, power = 0.692 | <br> <br>Critical = 2.627, NCP = 2.487, beta = 0.552, power = 0.448 |

# 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 sample and how it impacts the analysis plan
  - Whether the analysis plan is feasible (Is there enough data?)
  - Evaluation budget
- A power analysis can help programs effectively use and target resources

# When to conduct a power analysis

AmeriCorps

## 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

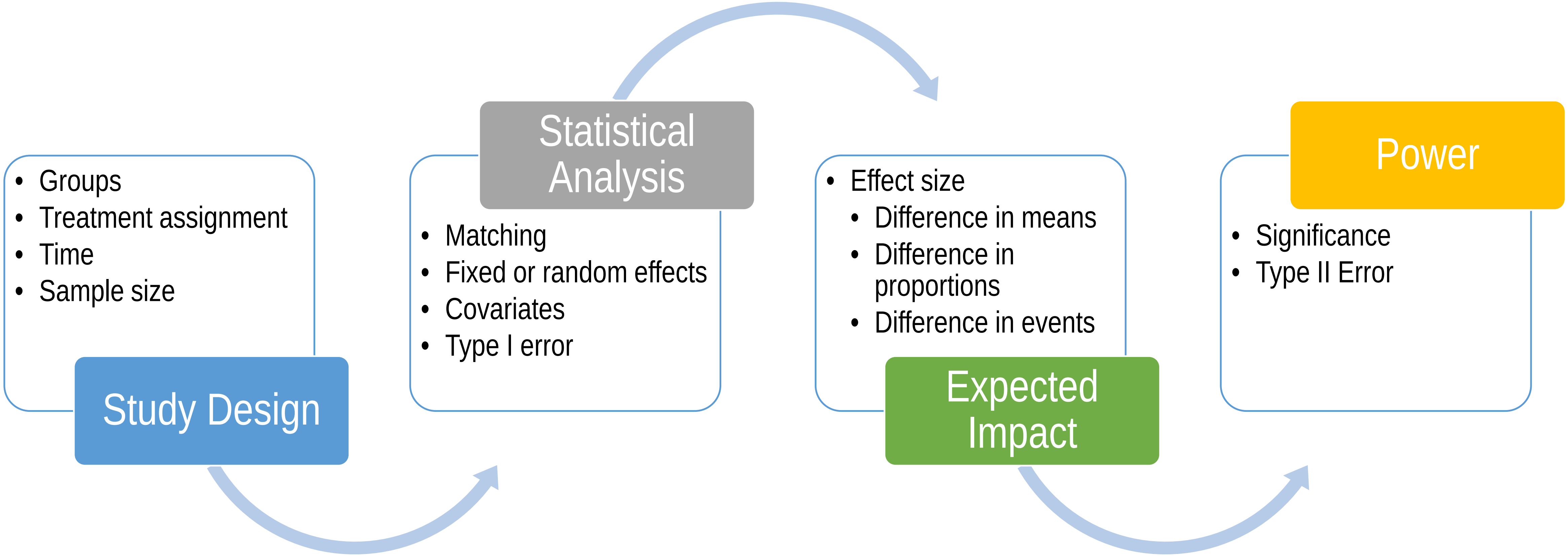
Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of educational and behavioral statistics*, *30*(2), 141-167.

# How do I design a high powered study?

AmeriCorps

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

# Design, Analysis, and Impact → Power



## Study Design

- Groups
- Treatment assignment
- Time
- Sample size

## Statistical Analysis

- Matching
- Fixed or random effects
- Covariates
- Type I error

## Expected Impact

- Effect size
  - Difference in means
  - Difference in proportions
  - Difference in events

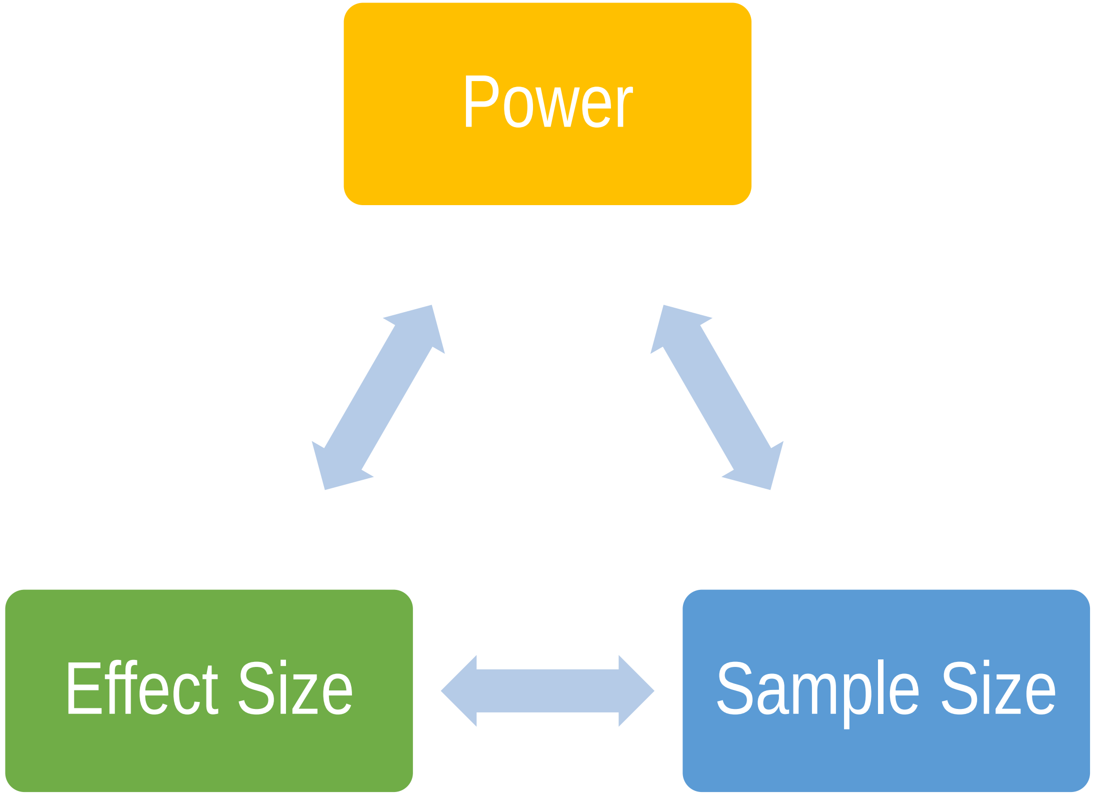
## Power

- Significance
- Type II Error

# What is a power analysis?

AmeriCorps

- 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)

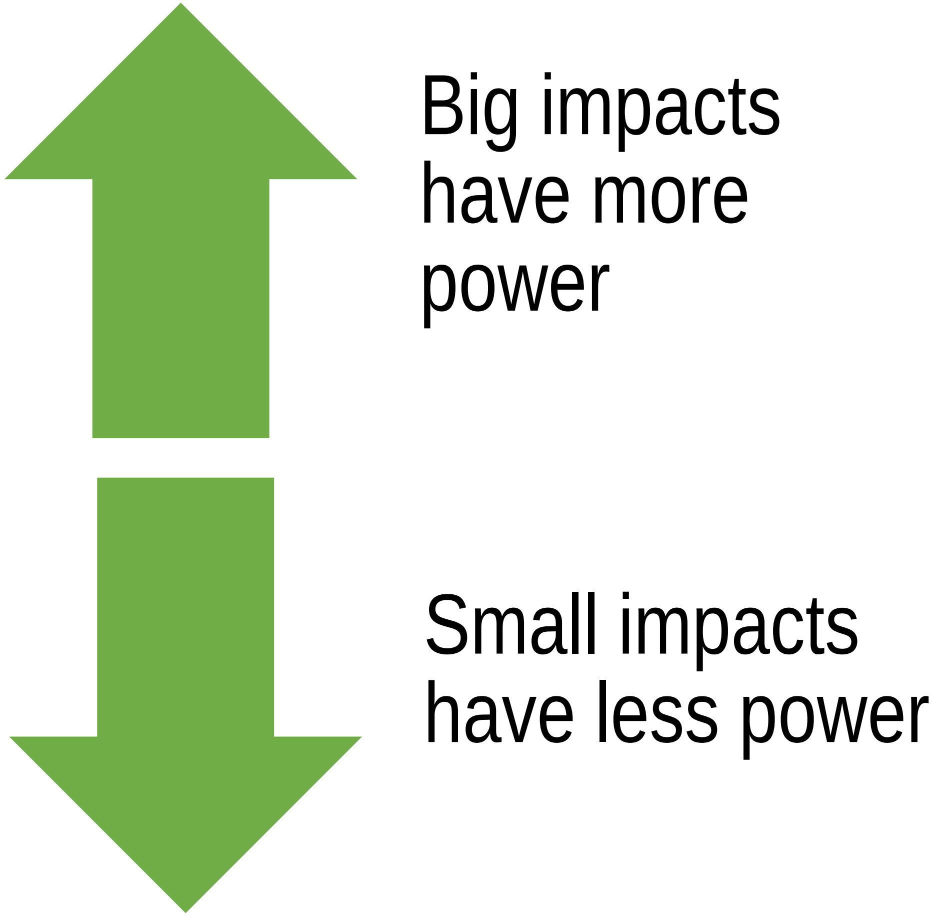


# Power and impact size

Power increases with impact size

Holding sample size/design constant:

- Bigger impacts have more power
- Smaller impacts have less power



Big impacts have more power

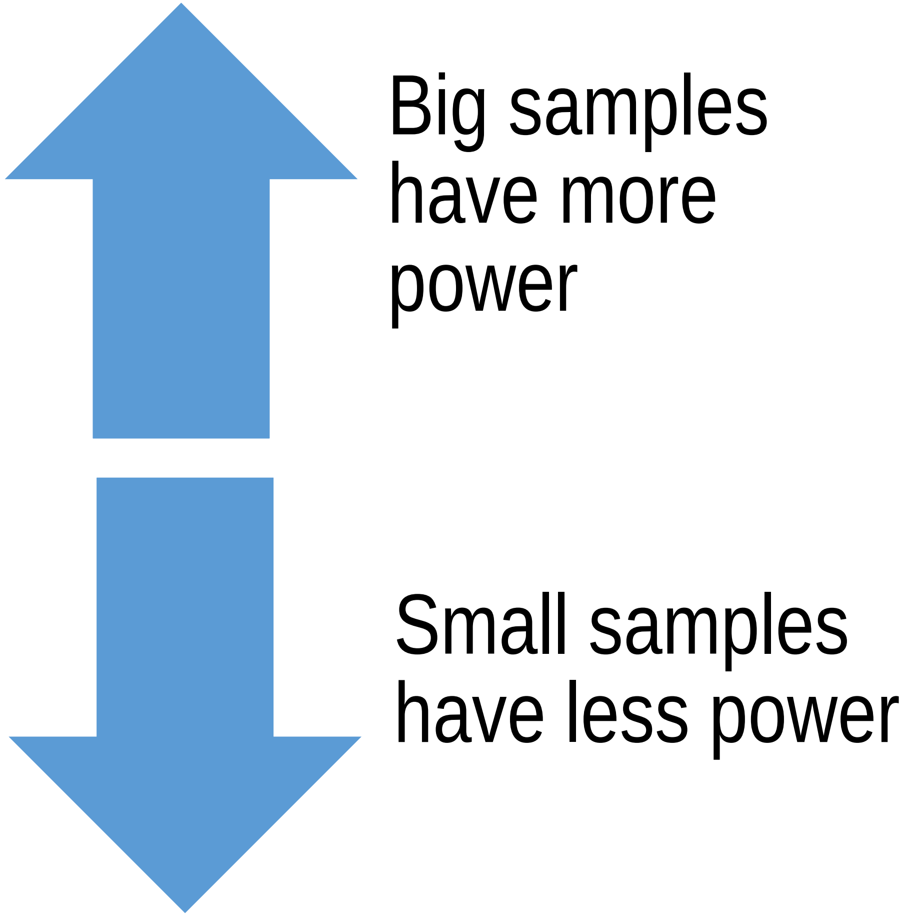
Small impacts have less power

# Power and sample size

Power increases with sample size

Holding impact constant:

- Bigger samples, depending on design, have more power
- Smaller samples, depending on design, have less power



Big samples have more power

Small samples have less power

# How to perform a power analysis

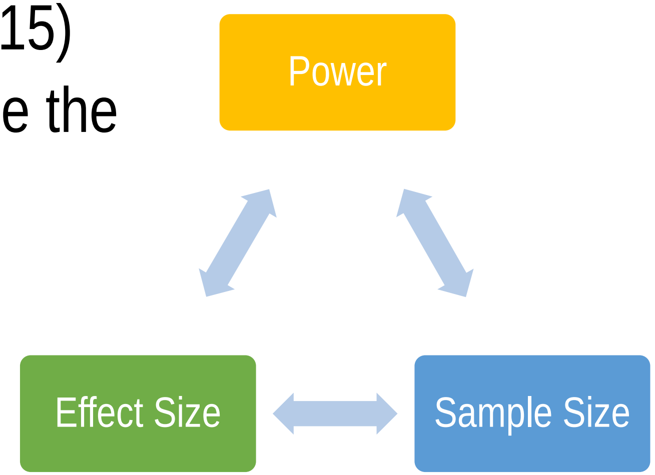
AmeriCorps

- 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)

# Example #1: Power analysis

A mentoring intervention program for high school age youth wants to determine if their program is having an impact on students’ school attendance.

- POWER: Assume a conventional power level of .80 $(1 - \beta, \mathrm{where}\ \beta$ is Type II error rate)
- SIGNIFICANCE: Assume a Type I error rate for $\alpha = .05$ (two-tailed test)
- EFFECT SIZE: Estimate the sample needed to detect an effect size for the intervention study group based on previous studies of mentoring programs (.15)
- SAMPLE SIZE: Estimate the effect size needed for a study where the number of students in the program is 250 students with another 250 students in the control group.



Power

Effect Size

Sample Size

# Example #1: Power analysis (Option #1)

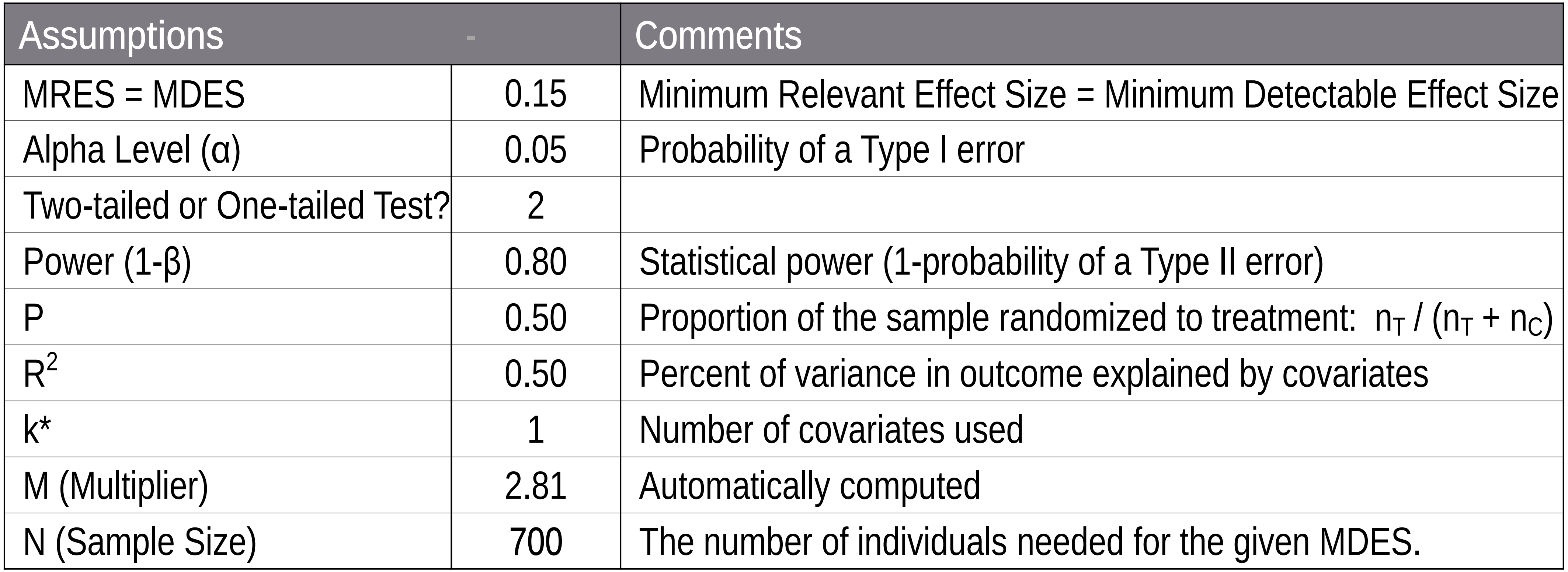
AmeriCorps

KEY QUESTION: How many students are needed to detect an effect size of .15?

- To have an 80% chance of detecting an effect size of .15 with statistical significance at $\alpha = .05$:
  - Assume a standard power level of .80 $(1 - \beta, \text{where } \beta$ is Type II error rate) and
  - Assume a Type I error rate for $\alpha = .05$ (two-tailed test)
  - Set effect size at .15

ANSWER: Without controlling for extraneous variables, 350 students are needed in the treatment group and 350 students are needed in the control group, for a total number of 700 students.

# PowerUp! Findings (Option #1)



| Assumptions | - | Comments |
|---|---|---|
| MRES = MDES | 0.15 | Minimum Relevant Effect Size = Minimum Detectable Effect Size |
| Alpha Level (α) | 0.05 | Probability of a Type I error |
| Two-tailed or One-tailed Test? | 2 | |
| Power (1-β) | 0.80 | Statistical power (1-probability of a Type II error) |
| P | 0.50 | Proportion of the sample randomized to treatment: $n_T / (n_T + n_C)$ |
| $R^2$ | 0.50 | Percent of variance in outcome explained by covariates |
| k* | 1 | Number of covariates used |
| M (Multiplier) | 2.81 | Automatically computed |
| N (Sample Size) | 700 | The number of individuals needed for the given MDES. |

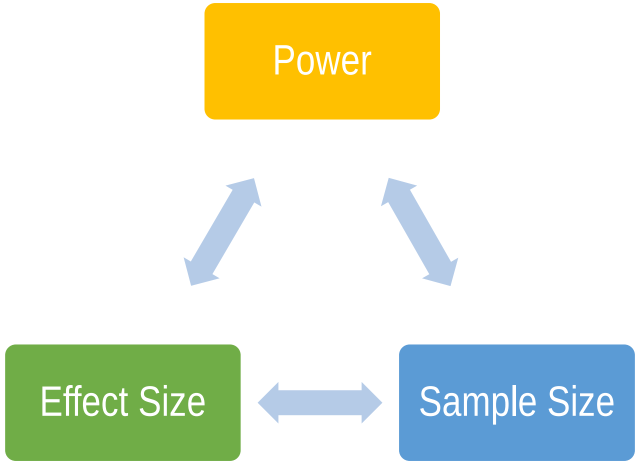
Source: www.causalevaluation.org

# Example #1: Power analysis (Option #2)

<u>KEY QUESTION</u>: What effect size must the program achieve to detect a statistically significant effect, given the known sample sizes?

- To have an 80% chance of a statistically significant study result at $\alpha = .05$:
  - Assume a standard power level of .80 ($1 - \beta$, where $\beta$ is Type II error rate) and
  - Assume a Type I error rate for $\alpha = .05$ (two-tailed test)
  - Set sample sizes at 250 for program/treatment group and 250 for control/comparison group

<u>ANSWER</u>: Without controlling for extraneous variables, an effect size of .18 standard deviations must be produced by the intervention

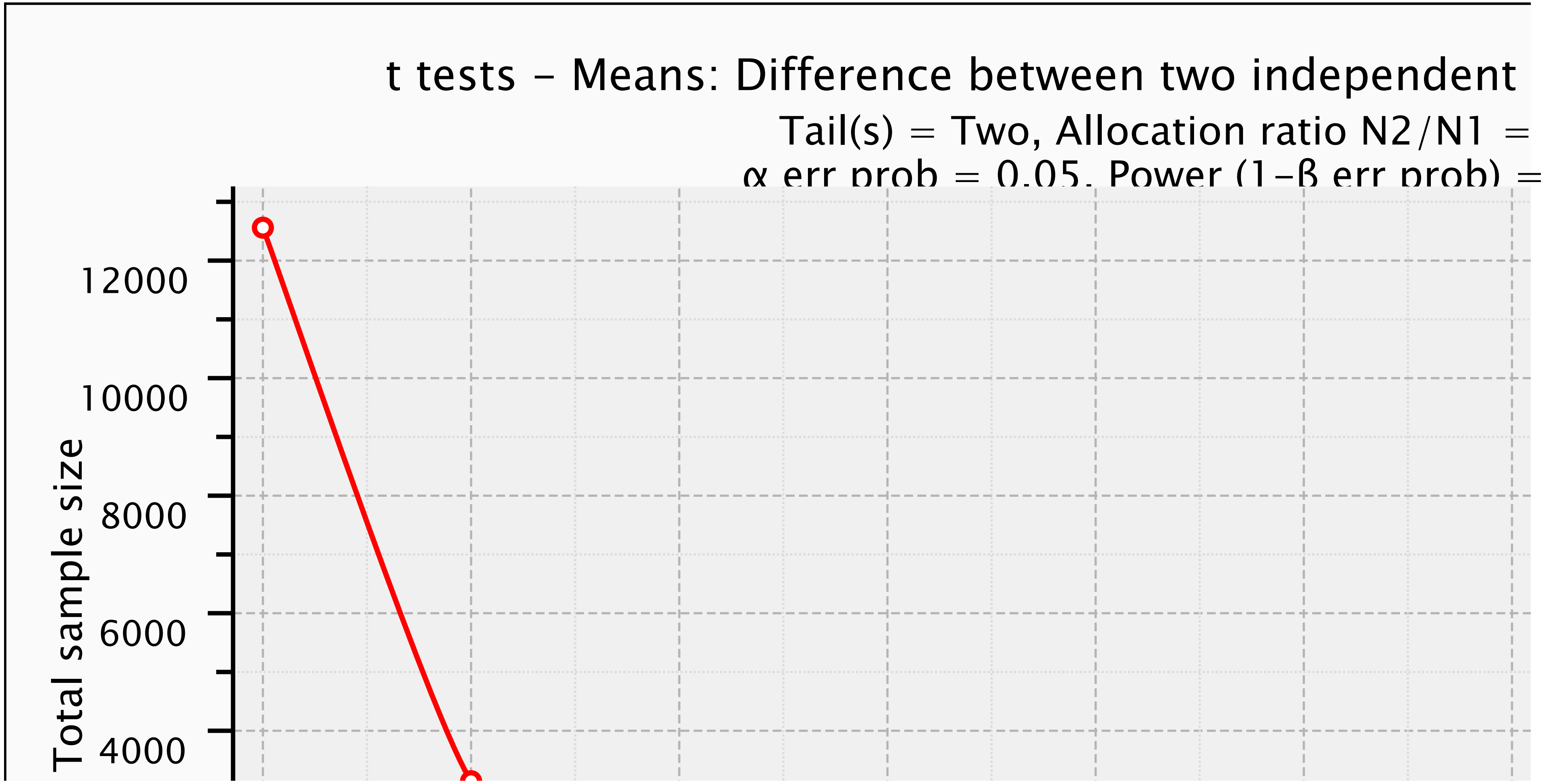


# PowerUp! Findings (Option #2)

| Assumptions | - | Comments |
|---|---|---|
| Alpha Level (α) | 0.05 | Probability of a Type I error |
| Two-tailed or One-tailed Test? | 2 | |
| Power (1-β) | 0.80 | Statistical power (1-probability of a Type II error) |
| P | 0.50 | Proportion of the sample randomized to treatment: $n_T$ / ($n_T$ + $n_C$) |
| $R^2$ | 0.50 | Percent of variance in outcome explained by covariates |
| k* | 1 | Number of covariates used |
| n (Total Sample Size) | 500 | |
| M (Multiplier) | 2.81 | Computed from $T_1$ and $T_2$ |
| $T_1$ (Precision) | 1.96 | Determined from alpha level, given two-tailed or one-tailed test |
| $T_2$ (Power) | 0.84 | Determined from given power level |
| MDES | 0.178 | Minimum Detectable Effect Size |

Source: www.causalevaluation.org

# Sample needed by effect size (no pretests)



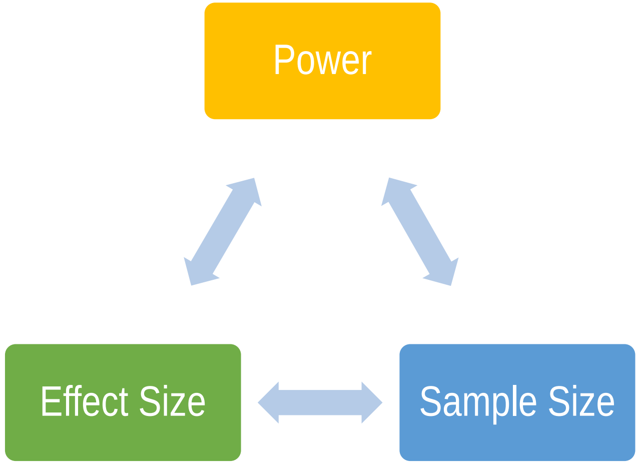
Plot produced with GPOWER; Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior research methods, instruments, & computers*, *28*(1), 1-11.

# Example #2: Power analysis

AmeriCorps

A job training intervention program for veterans wants to determine if their program is increasing the average wage growth for program participants.

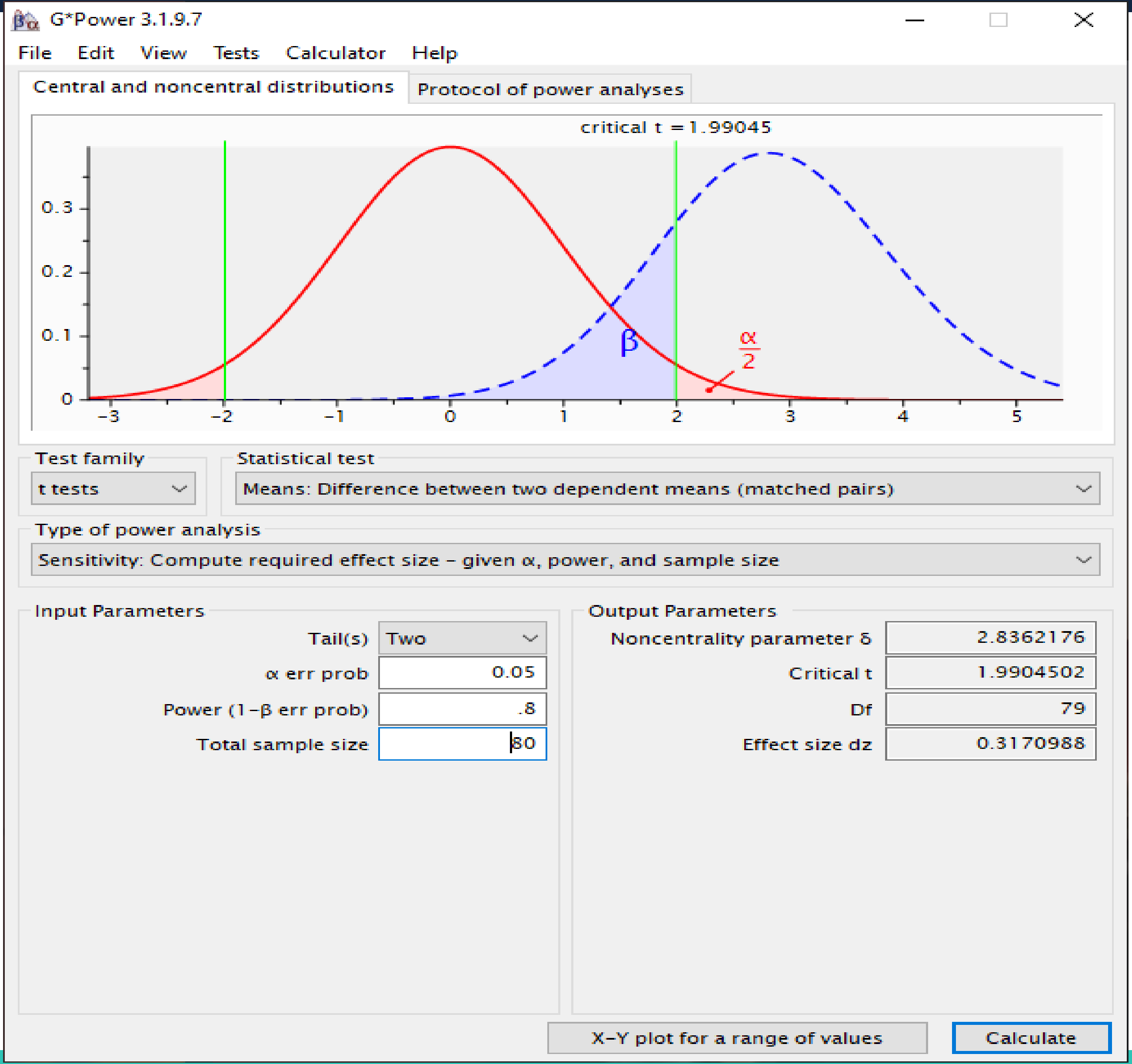
- POWER: Assume a conventional power level of .80 $(1-\beta, \text{where } \beta$ is Type II error rate)
- SIGNIFICANCE: Assume a Type I error rate for $\alpha = .05$ (two-tailed test)
- SAMPLE SIZE: Estimate the effect size needed for a study where the number of trainees served in a program year is 80 veterans.



ANSWER: The effect size needed for a study group of 80 veterans is 0.32.

# GPower

AmeriCorps



Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: Sensitivity: Compute required effect size – given α, power, and sample size

Input Parameters

| | |
|---|---|
| Tail(s) | Two |
| α err prob | 0.05 |
| Power (1−β err prob) | .8 |
| Total sample size | 80 |

Output Parameters

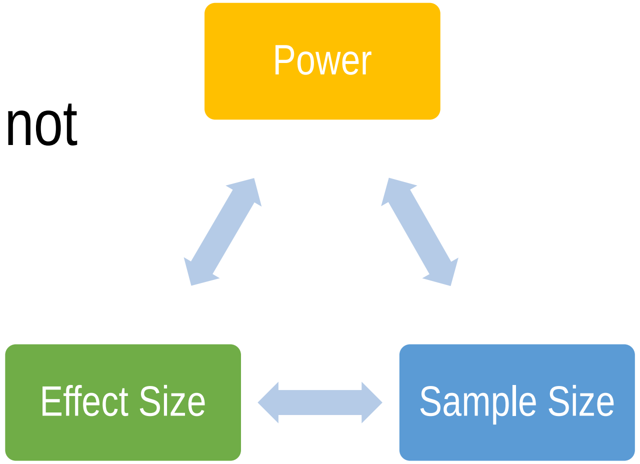
| | |
|---|---|
| Noncentrality parameter δ | 2.8362176 |
| Critical t | 1.9904502 |
| Df | 79 |
| Effect size dz | 0.3170988 |

Source:
https://www.gpower.hhu.de

# Example #3: Power analysis

A health education program for low-income residents wants to determine if their program is having an impact on basic knowledge levels in health screenings.

- POWER: Assume a conventional power level of .80 $(1 - \beta, \text{where } \beta$ is Type II error rate)
- SIGNIFICANCE: Assume a Type I error rate for $\alpha = .05$ (two-tailed test)
- EFFECT SIZE: Previous studies of health education programs among low-income populations show an effect size of 0.12.
- SAMPLE SIZE: The number of attendees served by the program is 500 residents. The study would identify 500 other residents who do not participate in the program to serve as a comparison group.



# Example #3: Power analysis

KEY QUESTION: What effect size must the program achieve to detect a statistically significant effect, given the known sample sizes?

- To have an 80% chance of a statistically significant study result at $\alpha = .05$:
    - Assume a standard power level of .80 ($1 - \beta, \text{where } \beta$ is Type II error rate) and
    - Assume a Type I error rate for $\alpha = .05$ (two-tailed test)
    - Set sample sizes at 500 for program/treatment group and 500 for control/comparison group

ANSWER: Without controlling for extraneous variables, an effect size of .13 standard deviations would need to be produced by the intervention. Therefore, the program would have to demonstrate a higher level of effectiveness than most previous studies of health education programs working with low-income populations in order to show a statistically significant effect.

# PowerUp! Findings (Example #3)

| Assumptions | - | Comments |
|---|---|---|
| Alpha Level (α) | 0.05 | Probability of a Type I error |
| Two-tailed or One-tailed Test? | 2 | |
| Power (1-β) | 0.80 | Statistical power (1-probability of a Type II error) |
| P | 0.50 | Proportion of the sample randomized to treatment: $n_T / (n_T + n_C)$ |
| $R^2$ | 0.50 | Percent of variance in outcome explained by covariates |
| k* | 1 | Number of covariates used |
| n (Total Sample Size) | 1000 | |
| M (Multiplier) | 2.80 | Computed from $T_1$ and $T_2$ |
| $T_1$ (Precision) | 1.97 | Determined from alpha level, given two-tailed or one-tailed test |
| $T_2$ (Power) | 0.84 | Determined from given power level |
| MDES | 0.125 | Minimum Detectable Effect Size |

Source: www.causalevaluation.org

# Power, impact size and sample size

AmeriCorps

- Data are like a microphone trying to detect a sound
- Small samples (microphone) can only detect large impacts (loud sounds)
- Big samples (microphone) can detect impacts both large and small (loud sounds and pin drops)

# Level II Summary

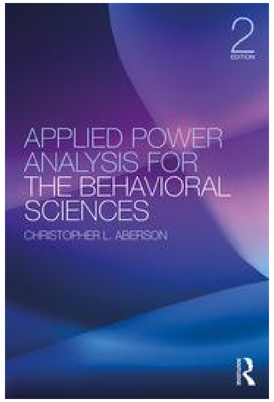
- Power is the chance to find an estimated impact “Statistically Significant”
- Power is the result of Type II Error relative to Type I Error
- Well powered studies can increase confidence in evaluation findings
- Power is the result of a process involving study design, analysis, and impacts
- Larger samples can detect smaller or larger impacts (are more sensitive) compared to smaller samples that can only detect larger impacts (are less sensitive)
- Power analysis helps you minimize Type II error
- Power analysis must occur before the study is conducted

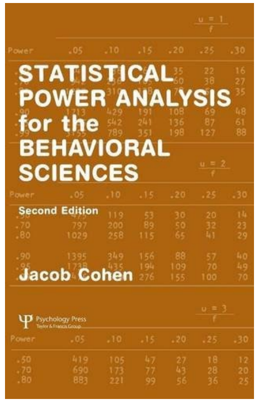
# Free power analysis software

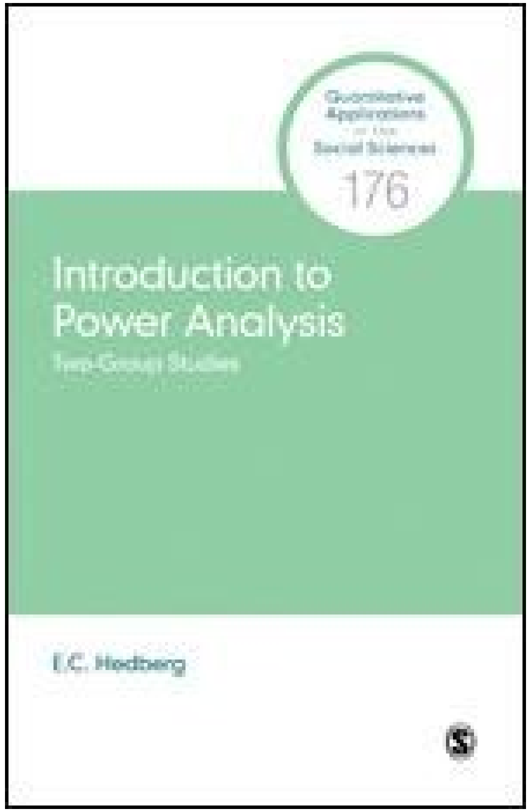
- PowerUP! (with clustering available)
  - https://www.causalevaluation.org/
  - Excel and R versions
  - Impacts, moderation (subgroup), and mediation analyses
- Other designs and analyses (without clustering): G*Power
  - https://www.gpower.hhu.de
  - Regression/Correlation, ANOVA, Probability Models

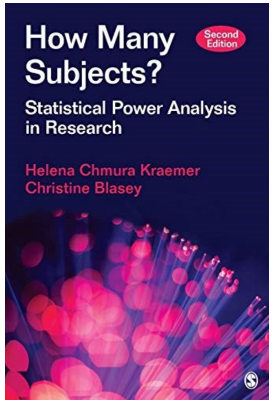
# Resources

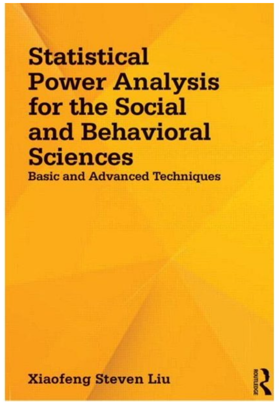
- Aberson, C. L. (2011). *Applied power analysis for the behavioral sciences*. Routledge.
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.
- Hedberg, E. C. (2017). *Introduction to Power Analysis: Two-group Studies* (Vol. 176). Sage Publications.
- Kraemer, 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


2 EDITION
APPLIED POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES
CHRISTOPHER L. ABERSON


STATISTICAL POWER ANALYSIS for the BEHAVIORAL SCIENCES
Second Edition
Jacob Cohen
Psychology Press


Quantitative Applications in the Social Sciences
176
Introduction to Power Analysis
Two-Group Studies
E.C. Hedberg


How Many Subjects?
Second Edition
Statistical Power Analysis in Research
Helena Chmura Kraemer
Christine Blasey


Statistical Power Analysis for the Social and Behavioral Sciences
Basic and Advanced Techniques
Xiaofeng Steven Liu

# Closing Remarks

Power Analysis for Program Evaluation II: Basic Mechanics of Power Analysis

## Dr. Carrie Markovitz

***Principal Research Scientist***

***NORC***