Introduction to Power Analysis for Child Welfare Program Evaluations
Housekeeping

- Event is being recorded and the recording will be posted online afterward.
- The slides and speaker biographies are available online.
- All participants are muted.
- Type your questions or comments into the Q&A box at any time.
INTRODUCTIONS
Learning Goals

- Participants will be able to calculate and interpret an expected effect size.
- Participants will be able to calculate the minimum number of observations to demonstrate potential impacts.
Overview of Session

- Power and evaluation design 20 min
- Power’s four parameters 15 min
- Estimating power or estimating parameters? 10 min
- Interactive visualization 10 min
- Other design considerations 10 min
- Closing 10 min
Power and Evaluation Design
Design Determines a Statistical Test

- **Step 1**: What is your evaluation question?
- **Step 2**: What variables are you using?
- **Step 3**: How will you analyze?

Power is understood in the context of a statistical test.
A Statistical Test Is Used for Hypothesis Testing

- Hypothesis Testing
  - Null hypothesis $H_0$: parameter = value
  - Alternative hypothesis $H_1$: parameter $\neq$ value
- Statistical test = guidelines for assessing how much information we have about the null hypothesis

Power is embedded in hypothesis testing.
## Review of Hypothesis Testing

<table>
<thead>
<tr>
<th>Decision</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$ True</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Type I error $\alpha$</td>
</tr>
<tr>
<td>Accept $H_0$</td>
<td>Correct (1 - $\alpha$)</td>
</tr>
</tbody>
</table>

$H_0$ = Null hypothesis: no difference between groups  
$\alpha$ = Alpha, or % chance of deciding there is a difference when there isn’t one  
$\beta$ = Beta, or % chance of deciding there is NOT a difference when there is one
Power Defined

- **Technical**: the probability of rejecting a null hypothesis when it is false
- **Translation**: the chance of detecting an effect that actually exists
Your evaluation design influences whether you will be able to demonstrate that your intervention has an effect.

An underpowered design cannot answer your evaluation question.
EXAMPLE: Testing program impact with a continuous outcome

Statistically significant difference between treatment vs. control groups
Design Determines a Statistical Test: Example

- **Step 1: Evaluation Question**
  Did children in the intervention group experience greater well-being than children in the control group?

- **Step 2: Outcome Variable**
  Well-being (scale ranging from 0 to 50)

- **Step 3: Statistical test**
  Independent sample $t$-test
Hypothesis Testing: Example

- **Hypothesis Testing**
  - Null hypothesis $H_0$: diff $\bar{x} = 0$
  - Alternative hypothesis $H_1$: diff $\bar{x} \neq 0$

- **Statistical test**
  $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

  - $p < .05$

- **Impact Example**
  - No difference between groups
  - Difference between groups
Important Parameters

- $t = \text{test statistic}$
- Difference between groups (effect size or diff $\bar{x}$)
- Variation in outcome scores ($s$)
- # in each group ($n$)
- $p$ value threshold (alpha or $\alpha$)

$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$p < .05$
Null Hypothesis: Example

$H_0 = \text{Null hypothesis: No difference between groups}$

$\alpha = \text{Alpha, or } % \text{ chance of deciding there is a difference when there isn’t one}$

$\text{diff } \bar{x} = \text{difference between average scores of two groups}$

95%+ probability that there is no difference between treatment and control
Actual Parameters: Example

H₀ = Null hypothesis: No difference between groups
H₁ = Alternative hypothesis: Difference between groups
α = Alpha, or % chance of deciding there is a difference when there isn’t one
diff \( \bar{x} \) = difference between average scores of two groups

Treatment = 60
Control = 50
diff \( \bar{x} \) = 10
What Is Power?

$H_0$

$\beta$

$H_1$

\[
\text{Power} = 1 - \beta
\]
Back to Hypothesis Testing

<table>
<thead>
<tr>
<th>Finding</th>
<th>H₀ True</th>
<th>H₀ False</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Decision</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reject H₀</td>
<td>Type I error $\alpha$</td>
<td>Correct (1-$\beta$)</td>
</tr>
<tr>
<td>Accept H₀</td>
<td>Correct (1-$\alpha$)</td>
<td>Type II error $\beta$</td>
</tr>
</tbody>
</table>

H₀ = Null hypothesis: no difference between groups
$\alpha$ = Alpha, or % chance of deciding there is a difference when there isn’t one
$\beta$ = Beta, or % chance of deciding there is NOT a difference when there is one
KEY POINTS

1. Power is understood in the context of a statistical test.

2. Power is embedded in hypothesis testing.

3. Power is the chance of correctly rejecting the null hypothesis (finding a true difference).
Power depends on four parameters.
Power’s Four Parameters (Again, Better Graphic)

- Effect size
- Variance in Outcome
- p value
- Sample size
Ways to Increase Power

\[ t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 + s_2^2 / n_1 + n_2}} \]

- Increase diff \( \bar{x} \)
- Decrease \( s \)
- Increase \( n \)
- Increase \( \alpha \) value

\( p < .05 \)
# Four Parameters: Effect Size Example

<table>
<thead>
<tr>
<th></th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effect size</strong></td>
<td>5</td>
<td>8</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td><strong>Sample size</strong></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td><strong>p value</strong></td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>.42</td>
<td>.80</td>
<td>.99</td>
<td>1.00</td>
</tr>
</tbody>
</table>

INCREASING EFFECT SIZE
# Four Parameters: Variance Example

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect size</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Variance</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Sample size</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>p value</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>Power</td>
<td>1.00</td>
<td>.94</td>
<td>.65</td>
</tr>
</tbody>
</table>

**DECREASING VARIANCE**
# Four Parameters: Sample Size Example

<table>
<thead>
<tr>
<th></th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect size</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Variance</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Sample size</td>
<td>60</td>
<td>100</td>
<td>150</td>
<td>250</td>
</tr>
<tr>
<td>p value</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>Power</td>
<td>.44</td>
<td>.65</td>
<td>.82</td>
<td>.96</td>
</tr>
</tbody>
</table>

**Increasing Sample Size**
KEY POINT

4. Power has four primary parameters: effect size, variation, sample size, $p$ value.
Estimating power or estimating parameters?
Power = These Four Parameters:

- Effect size
- Variance of outcome
- \( p \) value
- Sample size
You May Already Have Some Ingredients:

- Fixed or maximum sample size
- Desired effect size (pilot studies, standardized measurement, expert opinion)
- Variance (pilot studies, standardized measurement)
- Power $\geq .80$ (common threshold)
- $p$ value $<.05$ (common threshold)
Solving for Effect Size: Sensitivity Power Analysis

- Smallest effect size that is meaningful to the field
- Minimum detectable effect (MDE)—minimum effect you are able to distinguish from a null effect
- You must already know (or be able to estimate):
  - variance of outcome
  - sample size
  - $p$-value
  - desired power threshold
Types of Effect Sizes

- Mean
- Mean difference/Cohen’s $d$
- Correlation coefficient/$r$
- Regression coefficient/$\beta$
- Proportion
- Proportion difference
- Odds ratio
Resources to Help Estimate Variance/Standard Deviation and Effect Size

- Pilot or historical data
- Published literature
- Standardized measures
- Expert opinion
- Develop a range of possibilities to see how different values affect power (scenarios)
A priori power analysis is done during planning phase

What sample size \( (n) \) is necessary to produce smallest effect size of interest, or MDE (minimum detectable effect)?

You must already know (or be able to estimate):

- variance of outcome
- effect size
- \( p \)-value
- desired power threshold
Example of Power Plot (G*Power)

Scenarios help determine a range of conditions under which your evaluation design will allow for a minimum detectable effect.
KEY POINTS

5. It’s typical to solve for sample size or minimum detectable effect.

6. It’s good practice to construct scenarios.
Interactive Activity
Interactive Visualization

Cohen’s $d = \frac{\bar{x}_1 - \bar{x}_2}{s_P}$

Cohen’s cutoffs

0.2 = small
0.5 = medium
0.8 = large

Interactive Visualization of Power
Other Design Considerations

The more complex your design, research questions, and statistical tests, the harder it is to get good power estimates.
Other Design Considerations: Increase Power

- Using covariates
- Including repeated measures, correlated outcomes (dependent variables) over time

Both reduce random error (or variance) in the outcome variable.
Other Design Considerations: Less Power

- Violating assumptions of the statistical test (independence, normality, heterogeneity)
- Number of statistical tests
- Interactions, stratification
- Missing data (e.g., data quality, attrition, service engagement)
- Cluster correlated data
- Complex sampling/weighted designs
CLOSING REFLECTION

What new insights do you have about power as it relates to your evaluation work?
Introduction to Power Analysis for Child Welfare Program Evaluations