Effect Size and Statistical Power

Joseph Stevens, Ph.D., University of Oregon
(541) 346-2445, stevensj@uoregon.edu

An Introductory Problem or Two:

Which Study is Stronger?

Answer: Study B

Study A: $t\,(398) = 2.30$, $p = .022$

ω^2 for Study A = .01

Study B: $t\,(88) = 2.30$, $p = .024$

ω^2 for Study B = .05

Examples inspired by Rosenthal & Gaito (1963)

Study C Shows a Highly Significant Result

Study C: $F = 63.62$, $p < .0000001$

η^2 for Study C = .01, $N = 6{,}300$

Study D: $F = 5.40$, $p = .049$

η^2 for Study D = .40, $N = 10$

Correct interpretation of statistical results requires consideration of statistical significance, effect size, and statistical power

Three Fundamental Questions Asked in Science

Is there a relationship?

Answered by Null Hypothesis Significance Tests (NHST; e.g., *t* tests, *F* tests, $\chi 2$, *p*-values, etc.)

What kind of relationship?

Answered by testing if relationship is linear, curvilinear, etc.

How strong is the relationship?

Answered by effect size measures, <u>not</u> NHST's (e.g., R^2, r^2, η^2, ω^2, Cohen's *d*)

The Logic of Inferential Statistics

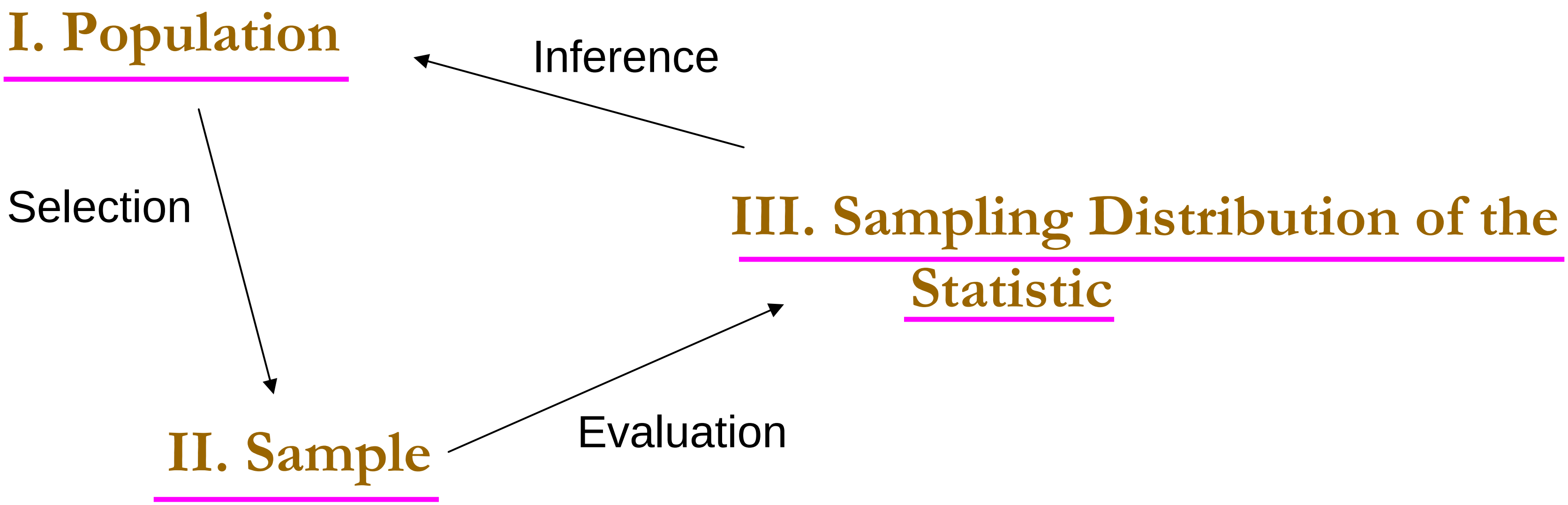
Three Distributions Used in Inferential Statistics:

- Population: the entire universe of individuals we are interested in studying (μ, σ, ∞)

- Sample: the selected subgroup that is actually observed and measured ($\overline{X}$, $\hat{s}$, N)

- Sampling Distribution of the Statistic: A theoretical distribution that describes how a statistic behaves across a large number of samples ($\mu_{\overline{X}}$, $\hat{s}_{\overline{X}}$, ∞)

The Three Distributions Used in Inferential Statistics



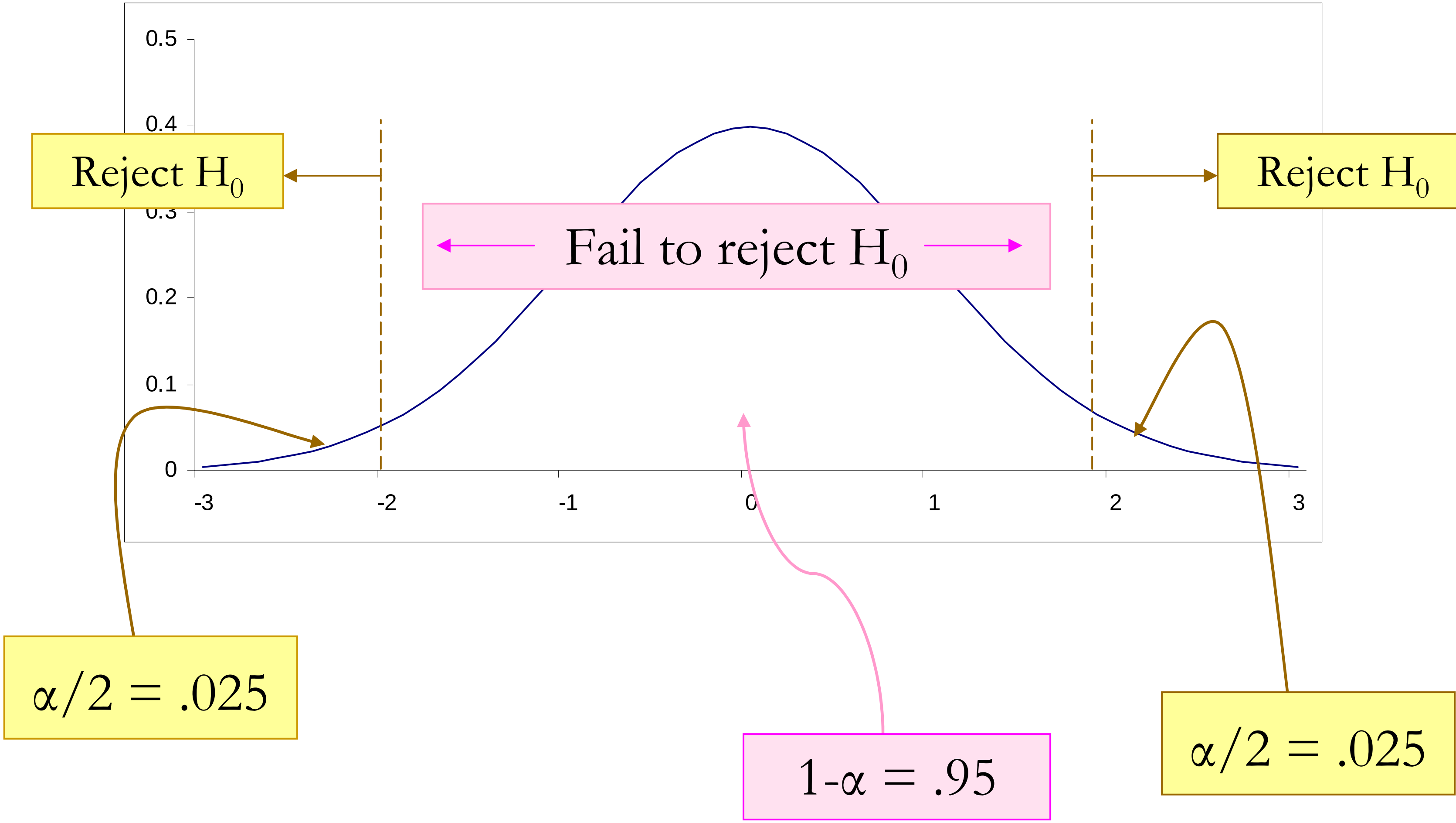
The NHST Decision Model (based on the sampling distribution of the statistic)

True State

Statistical Decision

	H_0 True	H_0 False
Fail to Reject H_0	**Correct Decision, $(1 - \alpha)$**	**Type II Error (β), False Negative**
Reject H_0	**Type I Error (α), False Positive**	**Correct Decision $(1 - \beta)$, Statistical Power**

H_0 True



Note: Sampling distributions are called Central Distributions when H_0 is true

True State

The value of α is set by convention which also determines $1 - \alpha$

Statistical Decision	H_0 True	H_0 False
Reject H_0	$(1 - \alpha) = .95$	**Type II Error** $\beta =$ **?**
Reject H_0	**Type I Error** $\alpha = .05$	**Statistical Power** $(1 - \beta) =$ **?**

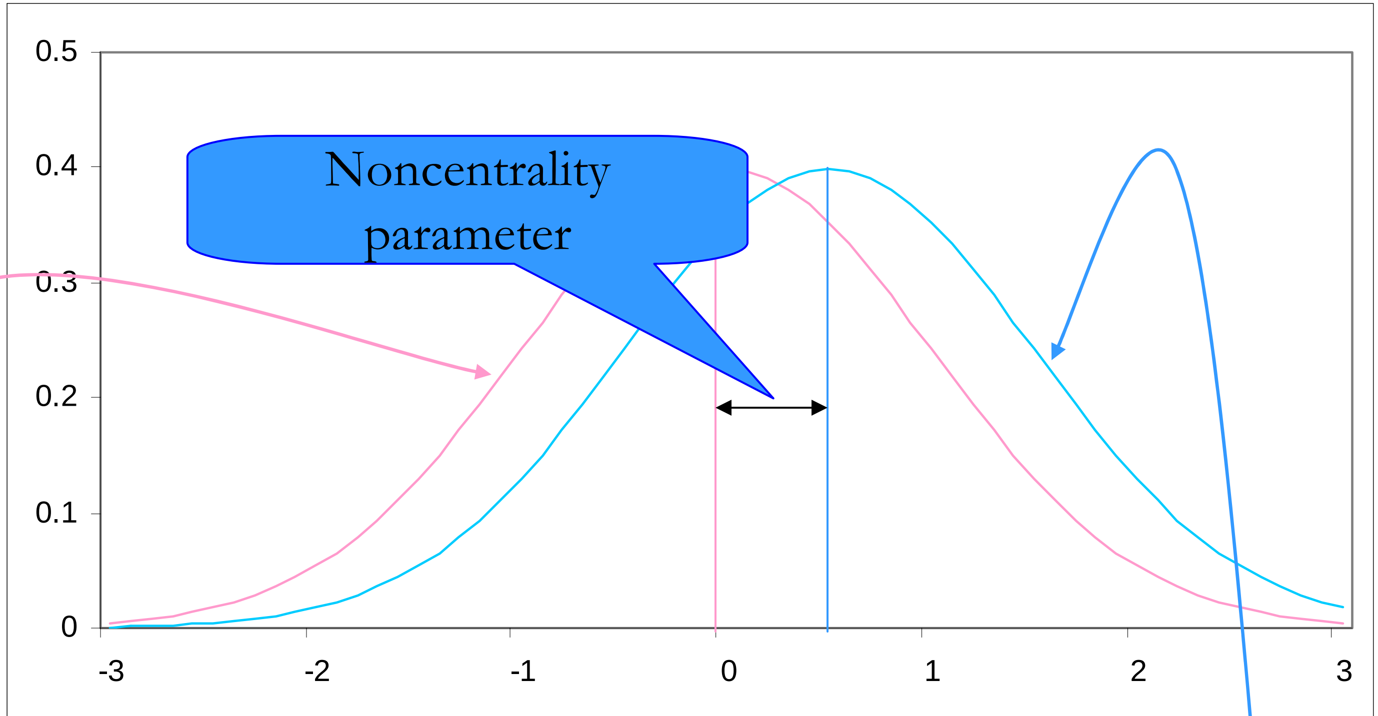
And, if H_0 is really true, then $\beta = 0$

But if H_0 is false, what are the values of β and $(1-\beta)$?

What if H_0 is False?

- If the null hypothesis is false, the sampling distribution and model just considered is incorrect
- In that case, a different sampling distribution describes the true state of affairs, the noncentral distribution
  - In fact there is a family of sampling distributions when the null is false that depend on just how large an effect is present
  - The size of the difference between the central and noncentral distributions is described by a noncentrality parameter

Central and Noncentral Distributions

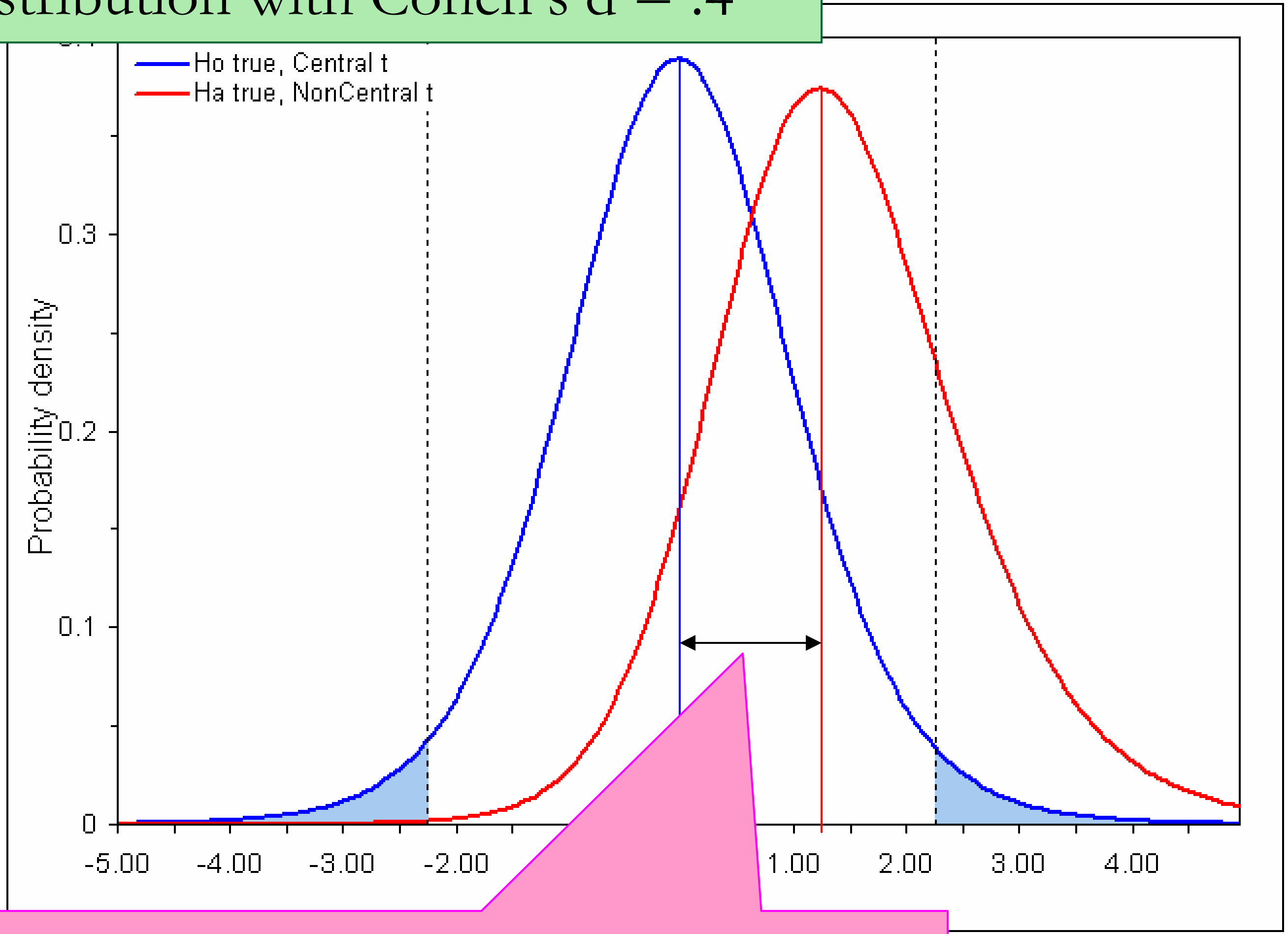


The noncentrality parameter represents the lack of overlap or displacement of the two distributions that results from a true difference between groups or nonzero relationship between variables

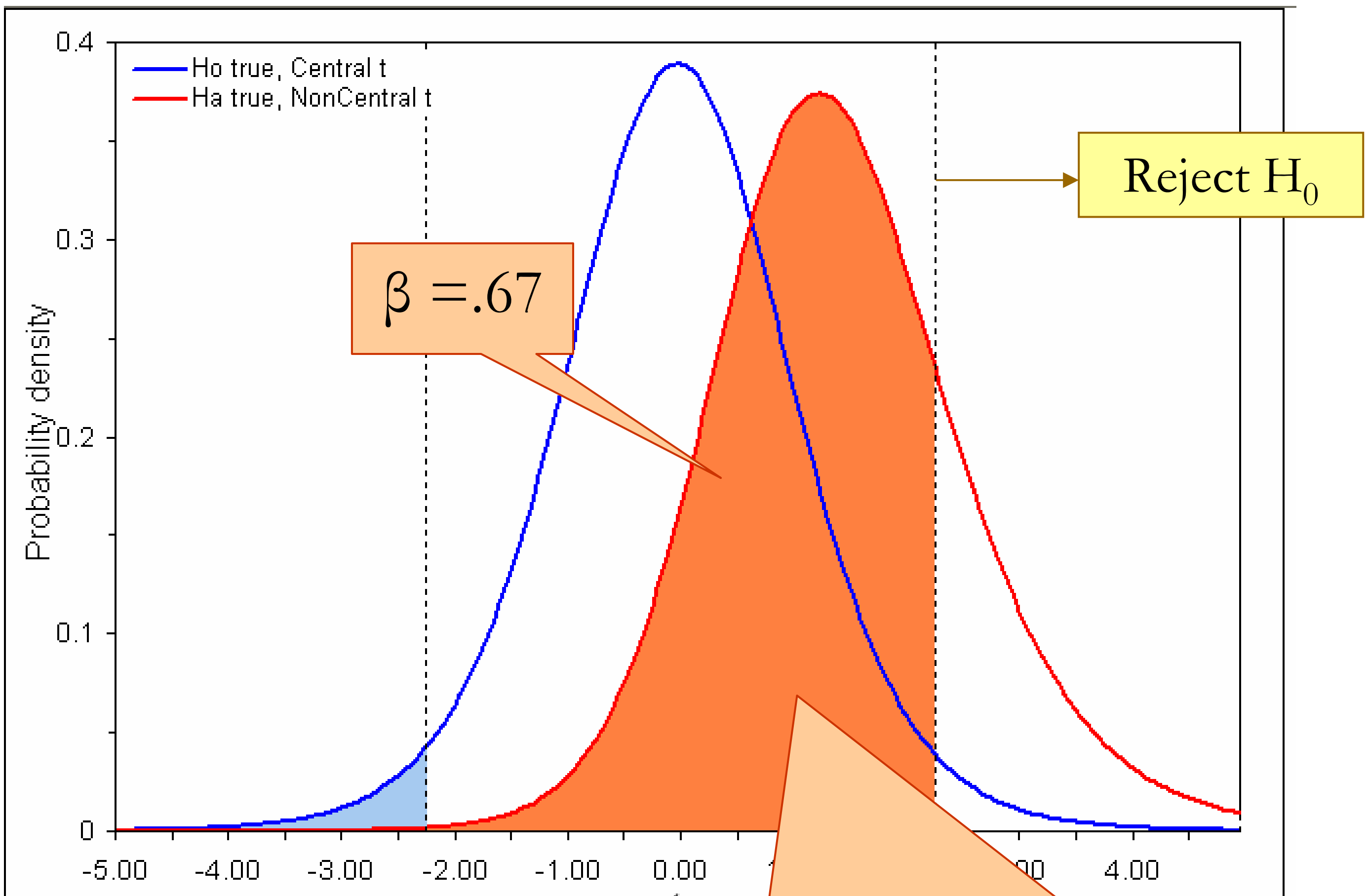
Distribution, H_0 False

H_0 True

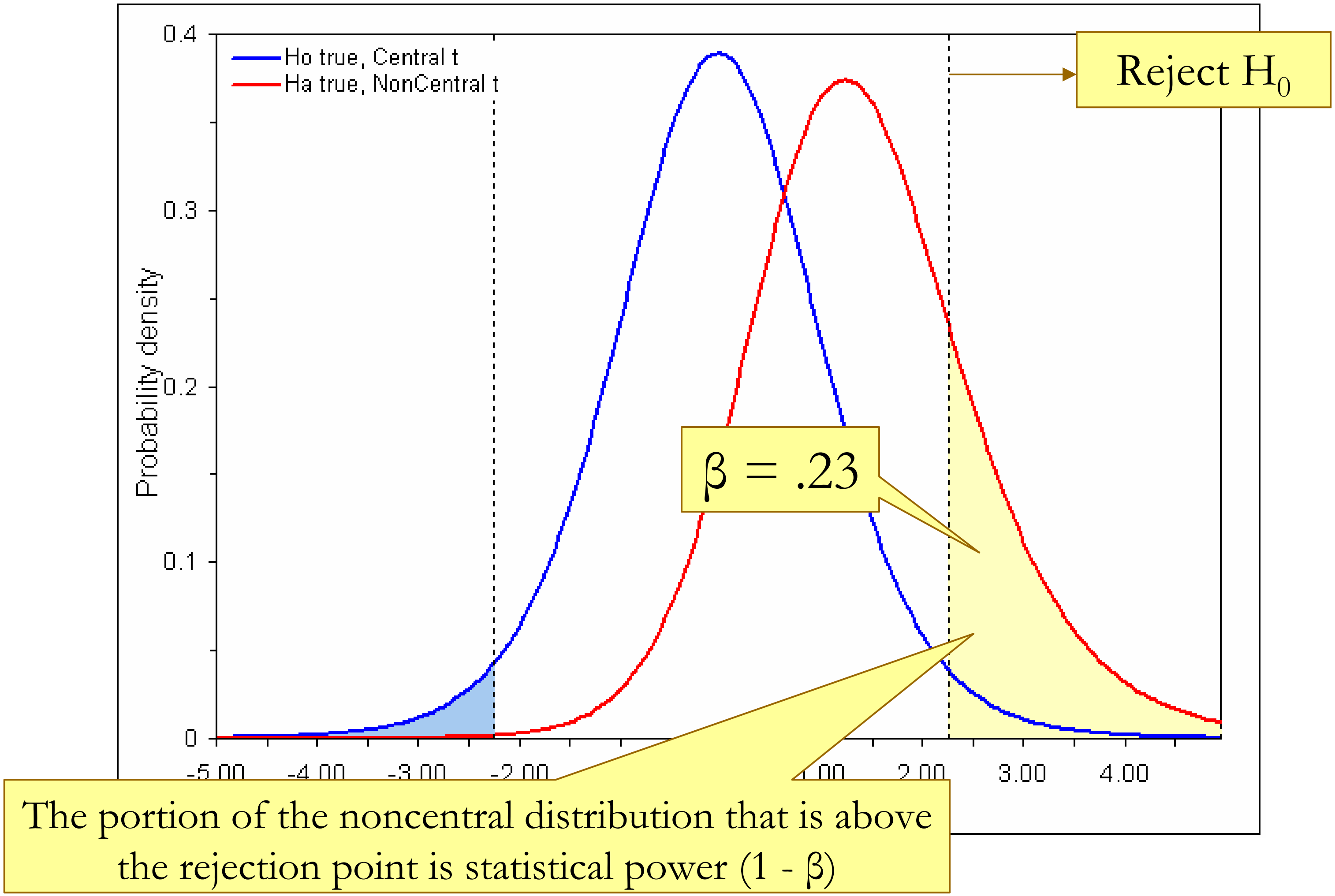
Assume an example using the t distribution with Cohen's d = .4



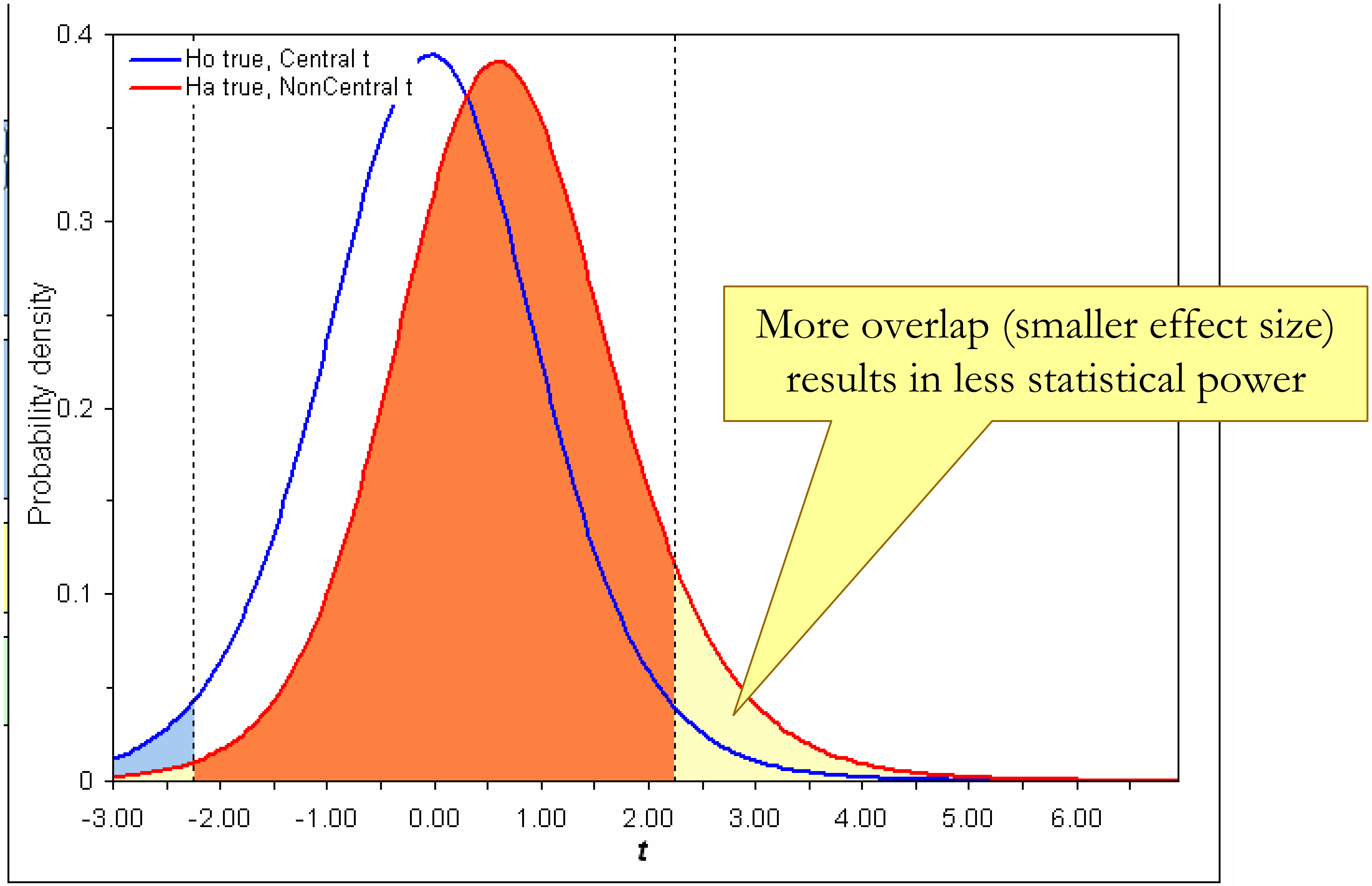
Note the disparity between the central and noncentral sampling distributions

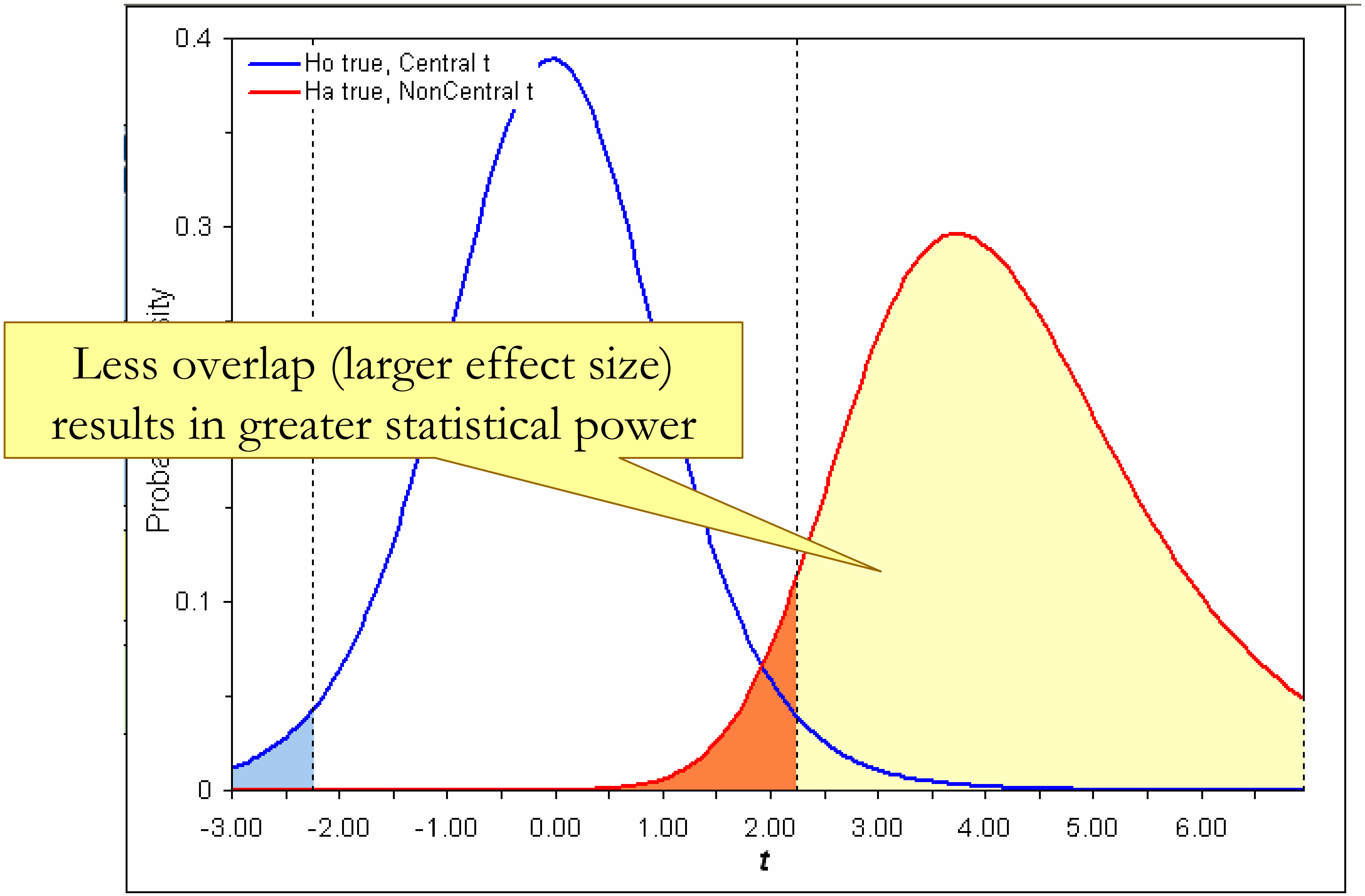


The portion of the noncentral distribution that is below the rejection point represents the probability of a Type II error (β)



Ho true, Central t
Ha true, NonCentral t
Probability density
0.4
0.3
0.2
0.1
0
-5.00
-4.00
-3.00
-2.00
2.00
3.00
4.00
Reject H_0
$\beta = .23$
The portion of the noncentral distribution that is above the rejection point is statistical power (1 - β)


0.4
0.3
0.2
0.1
0
Probability density
Ho true, Central t
Ha true, NonCentral t
-3.00
-2.00
-1.00
0.00
1.00
2.00
3.00
4.00
5.00
6.00
t
More overlap (smaller effect size) results in less statistical power



[ESCI Software]

The Relationship Between Effect Size and Statistical Significance

- It should be apparent that statistical significance depends on the size of the effect (e.g., the noncentrality parameter)
- And, statistical significance also depends on the size of the study (N)
- Statistical significance is the product of these two components

Significance Test Results = Effect Size X Size of Study

$$\boldsymbol{t} = \frac{r}{\sqrt{1-r^2}} \quad \text{X} \quad \sqrt{df}$$

$$\boldsymbol{t} = \frac{(\bar{X}_1 - \bar{X}_2)}{\hat{s}} \quad \text{X} \quad \frac{1}{\sqrt{\frac{1}{n_1}} + \sqrt{\frac{1}{n_2}}}$$

Significance Test Results = Effect Size X Size of Study

$$\boldsymbol{F} = \frac{r^2}{1-r^2} \quad \mathrm{X} \quad \mathrm{df}$$

$$\boldsymbol{F} = \frac{eta^2}{1-eta^2} \quad \mathrm{X} \quad \frac{df_{error}}{df_{means}}$$

Significance
Test Results = Effect Size X Size of Study

- To make correct interpretations, additional information beyond statistical significance is needed

- When results are statistically significant, it is very important to estimate effect size to determine the magnitude of results

Two Kinds of Metric for Measuring the Magnitude of Effects

- Standardized Difference Measures – Express the size of group difference in standard deviation units (e.g., Cohen's *d*)

- Strength of Association Measures – Express magnitude of effect as a proportion or percentage (e.g., r^2, η^2, ω^2)

Strength of Association Measures

- Pearson's *r*
- Multiple *R*
- Multivariate
  - Canonical *r*
  - Wilk's Lambda $(1 - \Lambda)$
- Effect size can be interpreted in units of r (see BESD below) or after squaring and multiplying by 100 as Percent Shared Variance (PSV)

$$\boxed{\text{PSV} = r^2 \text{ X } 100}$$

Strength of Association Measures

Correlation ratio

- Omega squared (ω^2)
- Eta squared (η^2)
- Partial eta squared ($\eta^2{}_p$)

Strength of Association Measures

- Cohen also uses f^2 as a metric of effect size
- This is easily expressed as R^2 or η^2

$$f^2 = \frac{R^2}{(1-R^2)} \qquad f^2 = \frac{\eta^2}{(1-\eta^2)}$$

Strength of Association Measures: ω^2

<u>Omega Squared for an independent *t*-test</u>:

$$\omega^2 = (t^2 - 1) \; / \; (t^2 + N_1 + N_2 - 1)$$

Example:

	Group 1	Group 2
Mean	65.50	69.00
Variance	20.69	28.96
N	30	30

$$t = 65.5 - 69 \, / \, 1.29 = -2.71$$

$$\omega^2 = (2.71)^2 - 1 \; / \; [(2.71)^2 + 30 + 30 - 1]$$

$$= 0.096, \text{ about } 10\% \text{ shared variance}$$

Strength of Association Measures: ω^2

Omega Squared for a one-factor ANOVA:

$$\omega^2 = \frac{[SS_{Between} - (a-1)(MS_{Residual})]}{(SS_{Total} + MS_{Residual})}$$

Strength of Association Measures: ω^2

<u>Omega Squared for a two-factor ANOVA</u>:

$$\omega^2 = [SS_A - (a-1)(MS_{Residual})] \;/\; (SS_{Total} + MS_{Residual})$$

$$\omega^2 = [SS_B - (b-1)(MS_{Residual})] \;/\; (SS_{Total} + MS_{Residual})$$

$$\omega^2 = [SS_{AB} - (a-1)(b-1)(MS_{Residual})] \;/\; (SS_{Total} + MS_{Residual})$$

Strength of Association Measures: ω^2

Example:

Source	SS	df	MS	*F*	*p*
A	3.61	1	3.61	2.76	.101
B	13.94	3	4.65	3.55	.019
AB	12.34	3	4.11	3.14	.030
Residual	94.30	72	1.31		
Total	757.00	80			

Strength of Association Measures: ω^2

$$\omega^2 = [SS_A - (a-1)(MS_{Residual})] / (SS_{Total} + MS_{Residual})$$

$$= [3.61 - (1)(1.31)] / (757 + 1.31) = .003$$

$$\omega^2 = [SS_B - (b-1)(MS_{Residual})] / (SS_{Total} + MS_{Residual})$$

$$= [13.94 - (3)(1.31)] / (757 + 1.31) = .013$$

$$\omega^2 = [SS_{AB} - (a-1)(b-1)(MS_{Residual})] / (SS_{Total} + MS_{Residual})$$

$$= [12.34 - (3)(1)(1.31)] / (757 + 1.31) = .011$$

Strength of Association Measures: η^2

$$\eta^2 = SS_{Effect} / SS_{Total}$$

An alternative measure is partial eta squared:

$$\eta^2_p = SS_{Effect} / (SS_{Effect} + SS_{Residual})$$

Note. Partial eta may sum to more than 100% in multifactor designs

Strength of Association Measures: η^2_p

An alternative formula using only *F* and df:

$$\eta_p^2 = \frac{[(F)(df_{effect})]}{[(F)(df_{effect}) + df_{residual}]}$$

Example using the interaction effect from above:

$$\eta_p^2 = \frac{[(F)(df_{effect})]}{[(F)(df_{effect}) + df_{residual}]} = \frac{(3.14)(3)}{[(3.14)(3) + 72]} = .116$$

Comparing Strength of Association Measures

Note the problems with partials:

- Different denominator for each effect
- Partials may sum to more than 100% in multifactor designs

			η^2_p
			.037
			.129
AB	.011	.016	.116

$$\eta^2_p = SS_{Effect} / (SS_{Effect} + SS_{Residual})$$

Note that: $\omega^2 \leq \eta^2 \leq \eta^2_p$

Group Difference Indices

- There are a variety of indices that measure the extent of the difference between groups
- Cohen's *d* is the most widely used index (two groups only)
- Generalization of Cohen's to multiple groups is sometimes called δ, but there is great variation in notation
- Hedges' *g* (uses pooled <u>sample</u> standard deviations)
- For multivariate, Mahalanobis' D^2

The Standardized Mean Difference: Cohen's d

$$d = \frac{(\bar{X}_1 - \bar{X}_2)}{\hat{s}_{pooled}}$$

$$\hat{s}_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

The Standardized Mean Difference: Cohen's d

Example:

	Group 1	Group 2
Mean	65.50	69.00
Variance	20.69	28.96
N	30	30

$$\hat{s}_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} = \sqrt{\frac{20.69(29) + 28.96(29)}{30 + 30 - 2}} = 4.98$$

$$d = \frac{(\bar{X}_1 - \bar{X}_2)}{\hat{s}_{pooled}} = \frac{(65.5 - 69.0)}{4.98} = -0.70$$

Interpreting Effect Size Results (How big is big?

- There is no simple answer to “How large should an effect size be?”
- The question begs another: “For what purpose?”
- The answer does not depend directly on statistical considerations but on the utility, impact, and costs and benefits of the results

Interpreting Effect Size Results

Cohen's "Rules-of-Thumb"

- standardized mean difference effect size (Cohen's *d*)
  - small = 0.20
  - medium = 0.50

"If people interpreted effect sizes (using fixed benchmarks) with the same rigidity that $p = .05$ has been used in statistical testing, we would merely be being stupid in another metric"
(Thompson, 2001; pp. 82–83).

  - large = 0.50

Interpreting Effect Size Results: Rosenthal & Rubin's BESD

Are Small Effects Unimportant?

The Binomial Effect Size Display (BESD) Corresponding to Various Values of r^2 and r

[illegible]	[illegible]	[illegible]	[illegible]	Success Rate Difference
[illegible]	[illegible]	[illegible]	[illegible]	.02
.00	.04	.48	.52	.04
.00	.06	.47	.53	.06
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
.01	.10	.45	.55	.10
.01	.12	.44	.56	.12
.03	.16	.42	.58	.16
.04	.20	.40	.60	.20
.06	.24	.38	.62	.24
.09	.30	.35	.65	.30
.16	.40	.30	.70	.40
.25	.50	.25	.75	.50
.36	.60	.20	.80	.60
.49	.70	.15	.85	.70
.64	.80	.10	.90	.80
.81	.90	.05	.95	.90
1.00	1.00	.00	1.00	1.00

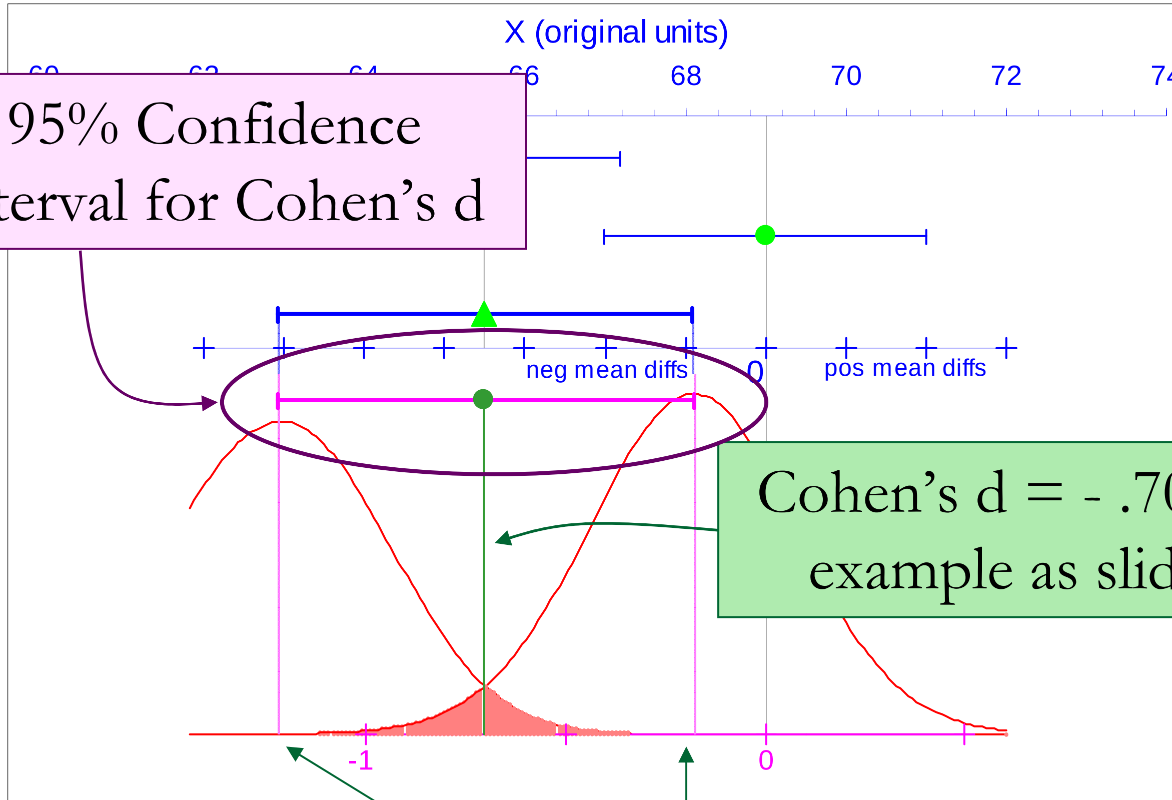
“Small” effects may be associated with important differences in outcomes

Success Rate Increase Associated with an r^2 of .10			
Condition	Alive	Dead	Total
Treatment	66	34	100
Control	34	66	100
Total	100	100	200

Note. Both tables from Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Also see Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Confidence Intervals for Effect Size



95% Confidence interval for Cohen's d

Cohen's d = - .70 (same example as slide 35)

95% CI = -1.2 to -0.2

[ESCI Software]()

Note. See Cumming & Finch (2001) or http://www.latrobe.edu.au/psy/esci/

Intermission

Statistical Power

- Statistical power, the probability of detecting a result when it is present
- Often the concern is "How many participants do I need?"
- While estimating *N* is important, a more productive focus may be on effect size and design planning
- How can I strengthen the research?

Factors Affecting Statistical Power

- Sample Size
- Effect Size
- Alpha level
- Unexplained Variance
- Design Effects

Effect of Sample Size on Statistical Power

All things equal, sample size increases statistical power at a geometric rate (in simple designs)

- This is accomplished primarily through reduction of the standard error of the sampling distribution
- With large samples, inferential statistics are very powerful at detecting very small relationships or very small differences between groups (even trivial ones)
- With small samples, larger relationships or differences are needed to be detectable

Effect of Sample Size on Statistical Power

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$$

$$\hat{s}_{\overline{X}} = \frac{\hat{s}}{\sqrt{N}}$$

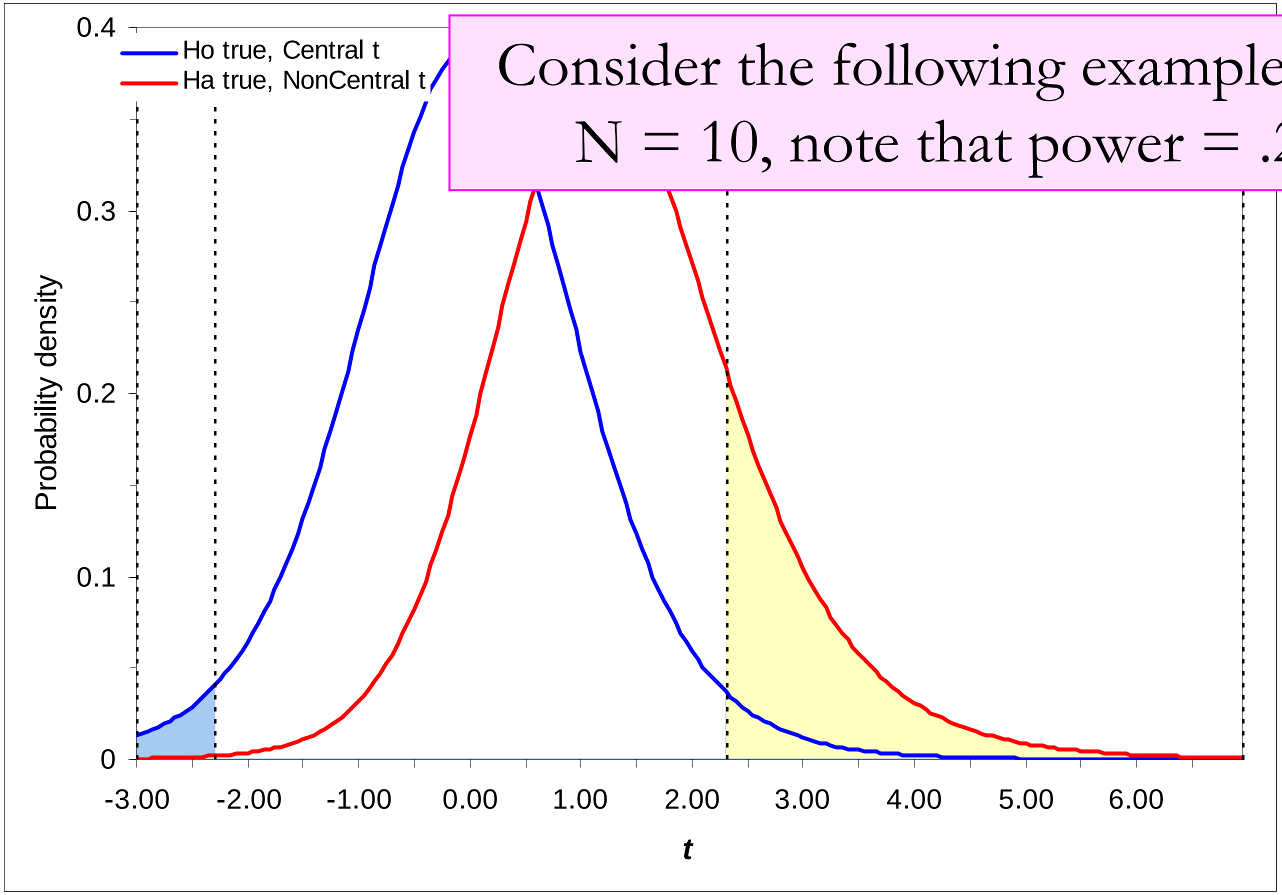
As an example, if the estimated population standard deviation was 10 and sample size was 4 then:

$$\hat{s}_{\overline{X}} = \frac{10}{\sqrt{4}} = 5$$

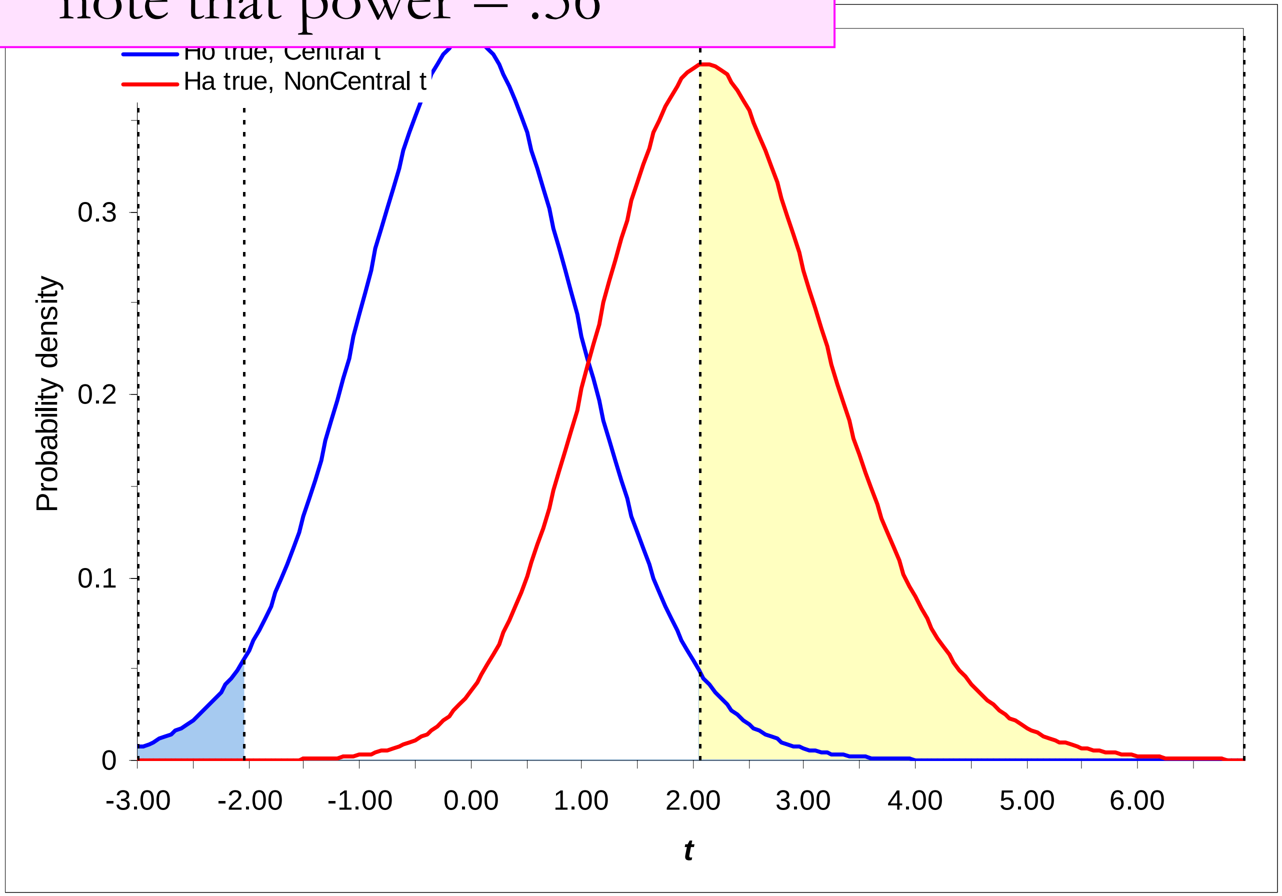
But if sample was 16 (4 times larger) then the standard error is 2.5 (smaller by half):

$$\hat{s}_{\overline{X}} = \frac{10}{\sqrt{16}} = 2.5$$

Consider the following example with N = 10, note that power = .21

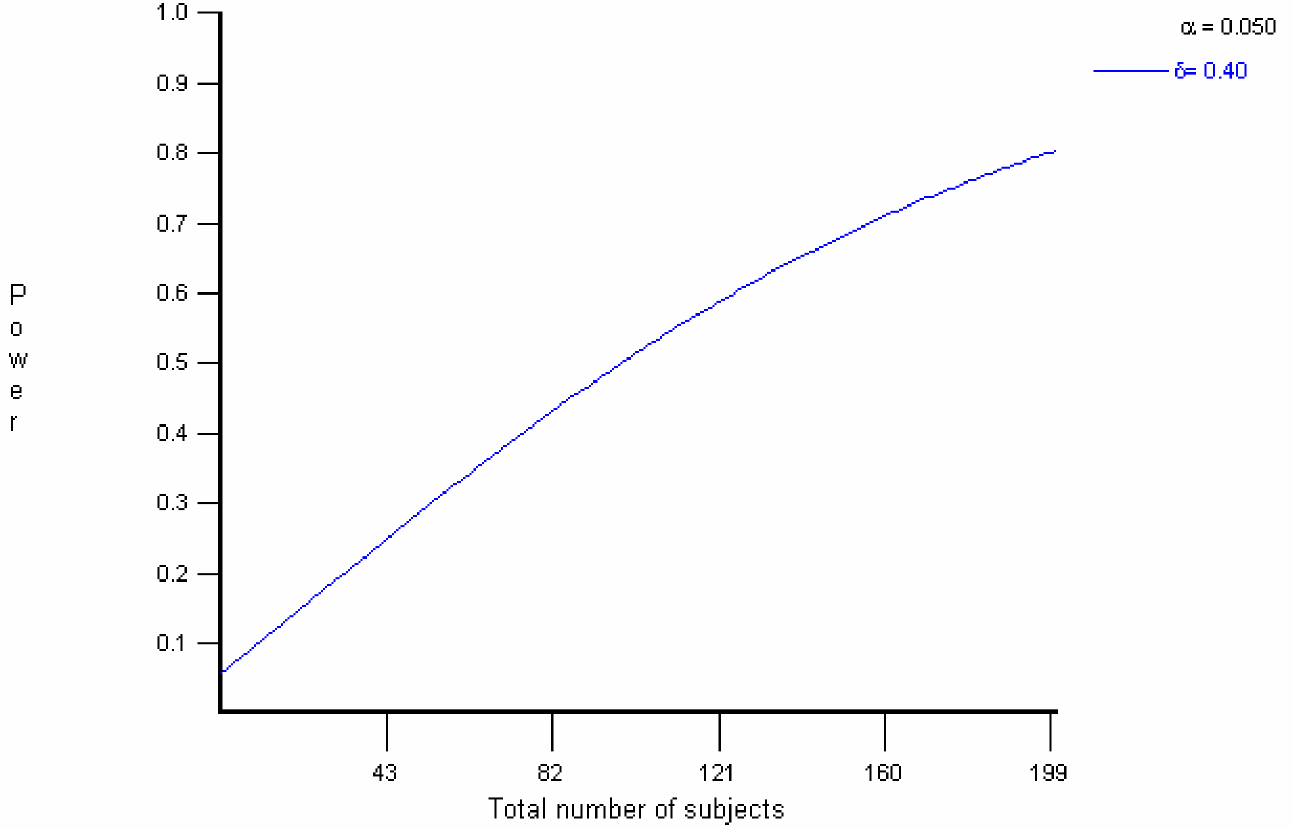


Versus a second example with N = 30, note that power = .56

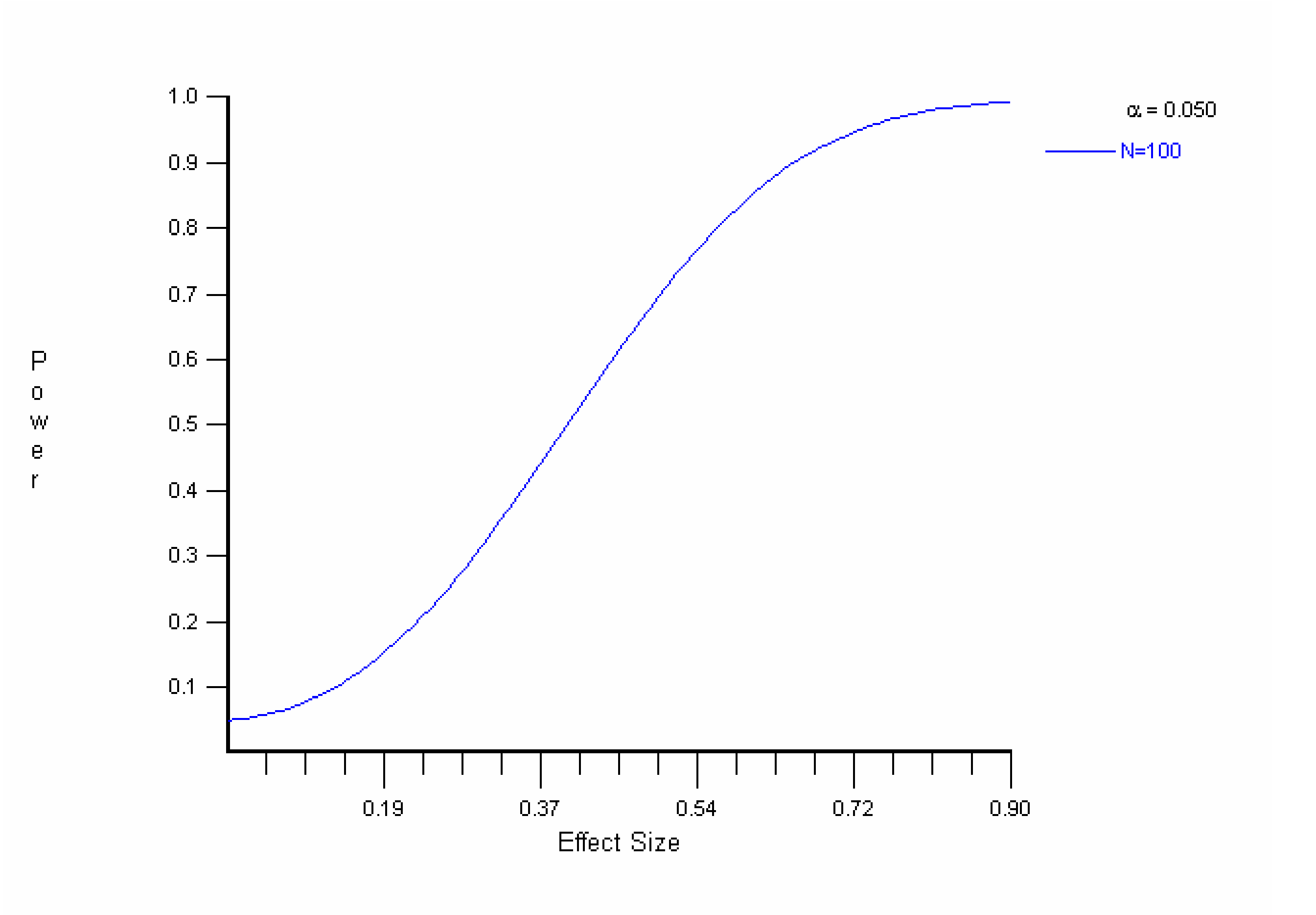


ESCI Software

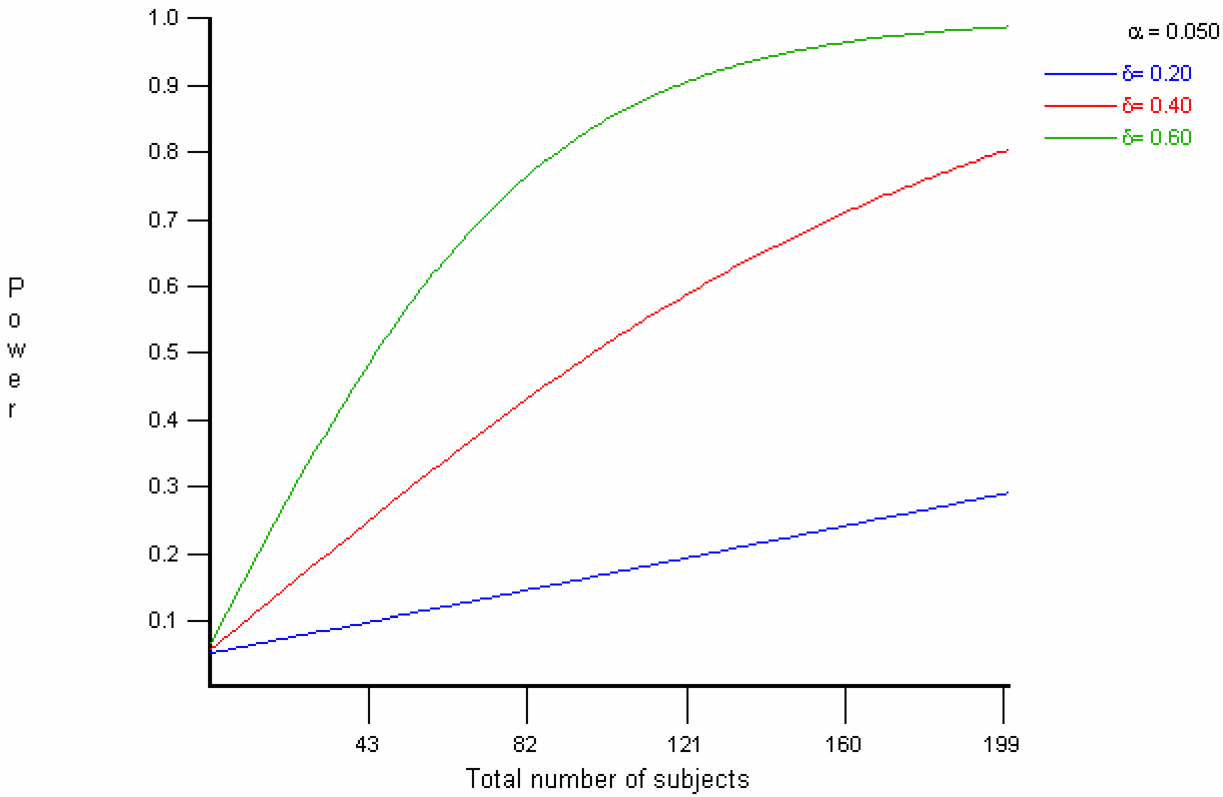
Impact of Sample Size on Statistical Power



Impact of Effect Size on Statistical Power



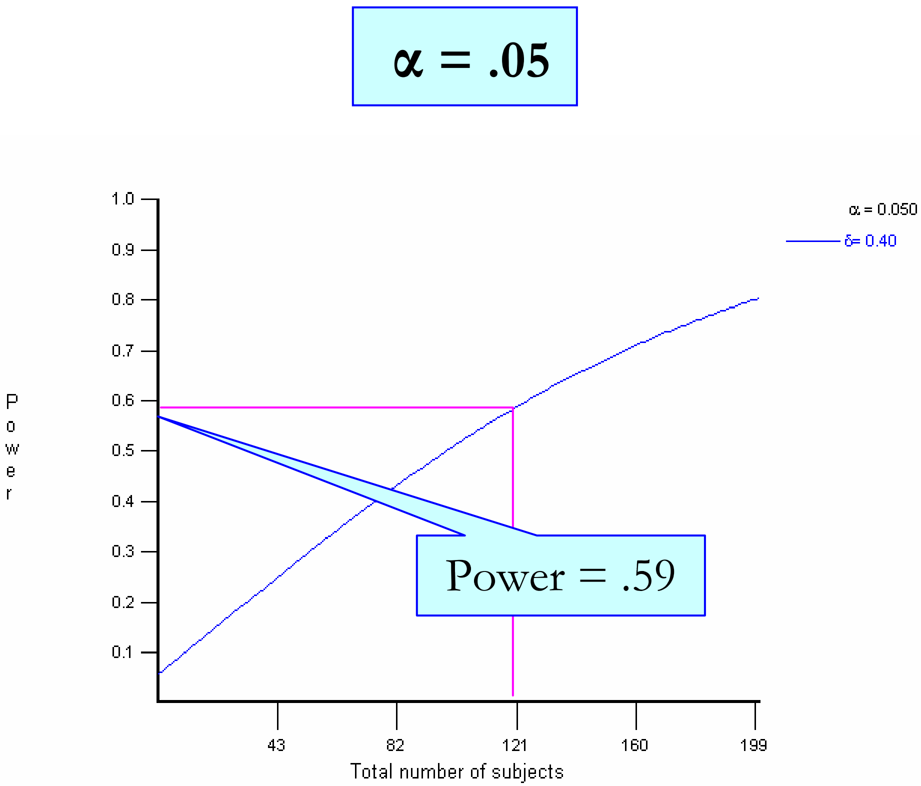
Impact of Sample and Effect Size on Statistical Power



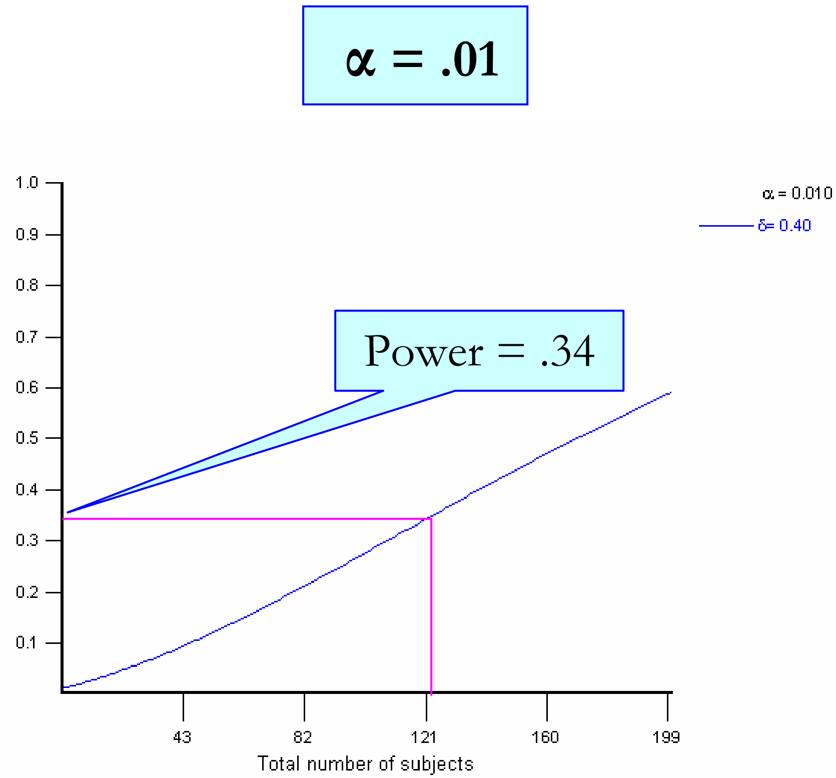
Effect of Alpha Level on Statistical Power

- One-tailed tests are more powerful than two-tailed tests
  - Require clear a priori rationale
  - Requires willingness to ignore results in the wrong direction
  - Only possible with certain statistical tests (e.g., t but not F)

- Larger alpha values more powerful (e.g., $p < .10$)
  - May be difficult to convince reviewers
  - Can be justified well in many program evaluation contexts (when only one direction of outcome is relevant)
  - Justifiable with small sample size, small cluster size, or if, a priori, effect size is known to be small

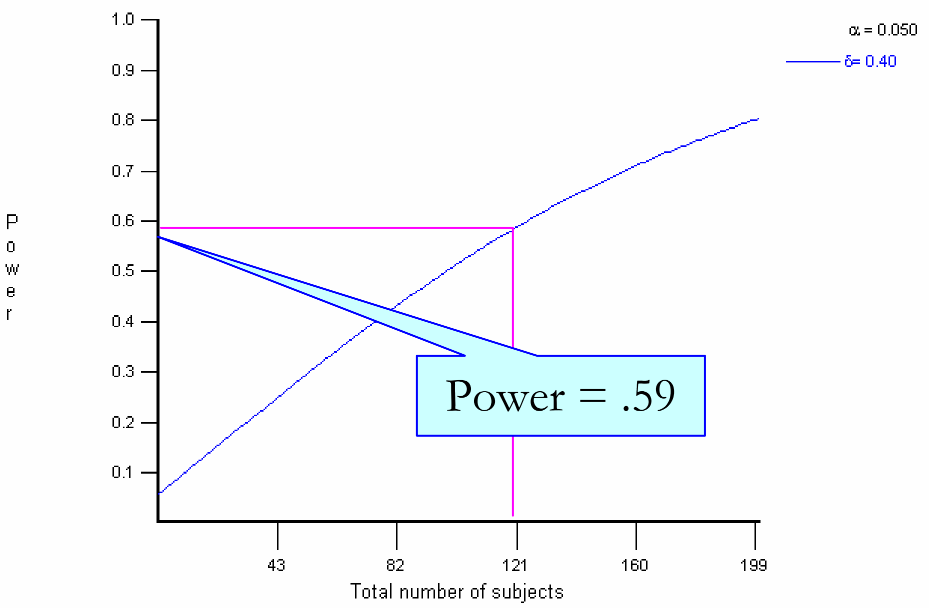


α = .05
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
P
o
w
e
r
α = 0.050
δ= 0.40
Power = .59
43
82
121
160
199
Total number of subjects



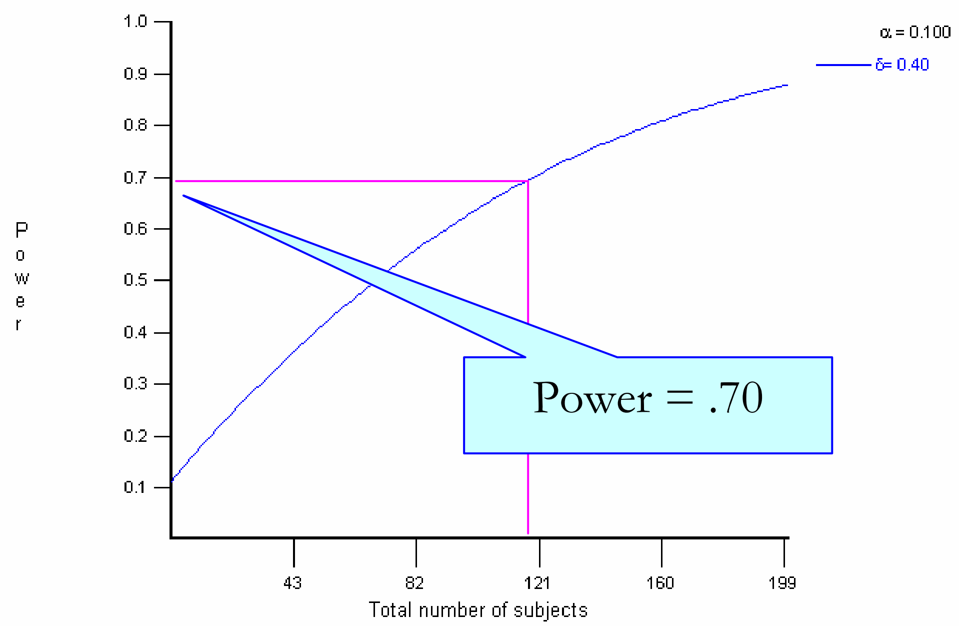
α = .01
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
α = 0.010
δ= 0.40
Power = .34
43
82
121
160
199
Total number of subjects

α = .05



α = 0.050
δ= 0.40
Power = .59
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
P o w e r
43
82
121
160
199
Total number of subjects

α = .10



α = 0.100
δ= 0.40
Power = .70
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
P o w e r
43
82
121
160
199
Total number of subjects

Effect of Unexplained Variance on Statistical Power

- Terminology: "error" versus unexplained or residual
- Residual variance reduces power
  - Anything that decreases residual variance, increases power (e.g., more homogeneous participants, additional explanatory variables, etc.)
- Unreliability of measurement contributes to residual variance
- Treatment infidelity contributes to residual variance

Effect of Design Features on Statistical Power

- Stronger treatments!
- Blocking and matching
- Repeated measures
- Focused tests ($df = 1$)
- Intraclass correlation
- Statistical control, use of covariates
- Restriction of range (IV and DV)
- Measurement validity (IV and DV)

Effect of Design Features on Statistical Power

- Multicollinearity (and restriction of range)

$$s_{b_{y1.2}} = \sqrt{\frac{s^2_{y12}}{\Sigma x_1^2 (1 - r_{12}^2)}}$$

- Statistical model misspecification
  - Linearity, curvilinearity,…
  - Omission of relevant variables
  - Inclusion of irrelevant variables

Options for Estimating Statistical Power

- Cohen's tables
- Statistical Software like SAS and SPSS using syntax files
- Web calculators
- Specialized software like G*Power, Optimal Design, ESCI, nQuery

Estimating Statistical Power

- Base parameters on best information available
- Don't overestimate effect size or underestimate residual variance or ICC
- Consider alternative scenarios
  - What kind of parameter values might occur in the research?
  - Estimate for a variety of selected parameter combinations
  - Consider worst cases (easier to plan than recover)

Recommendations for Study Planning

- Greater attention to study design features
- Explore the implications of research design features on power

- Base power estimation on:
  - Prior research
  - Pilot studies
  - Plausible assumptions
  - Thought experiments
  - Cost/benefit analysis

Power in Multisite and Cluster Randomized Studies

- More complex designs involving data that are arranged in inherent hierarchies or levels
- Much educational and social science data is organized in a multilevel or nested structure
  - Students within schools
  - Children within families
  - Patients within physicians
  - Treatments within sites
  - Measurement occasions within individuals

Power in Multisite and Cluster Randomized Studies

Factors affecting statistical power

- Intraclass Correlation (ICC)
- Number of participants per cluster (N)
- Number of clusters (J)
- Between vs. within cluster variance
- Treatment variability across clusters
- Other factors as discussed above

Intraclass Correlation Coefficient (ρ)

$$\text{Total } \sigma^2_Y = \tau^2 + \sigma^2$$

$$\text{ICC} = \frac{\text{population variance between units}}{\text{total variance}}$$

$$= \quad \tau^2 / (\tau^2 + \sigma^2)$$

As ICC approaches 0, multilevel modeling is not needed and power is the same as a non-nested design, but even small values of ICC can impact power

Intraclass Correlation (ρ)

- The Intraclass Correlation Coefficient (ICC) measures the correlation between a grouping factor and an outcome measure
- In common notation there are 1 to J groups
- If participants do not differ from one group to another, then the ICC = 0
- As participants' outcome scores differ due to membership in a particular group, the ICC grows large

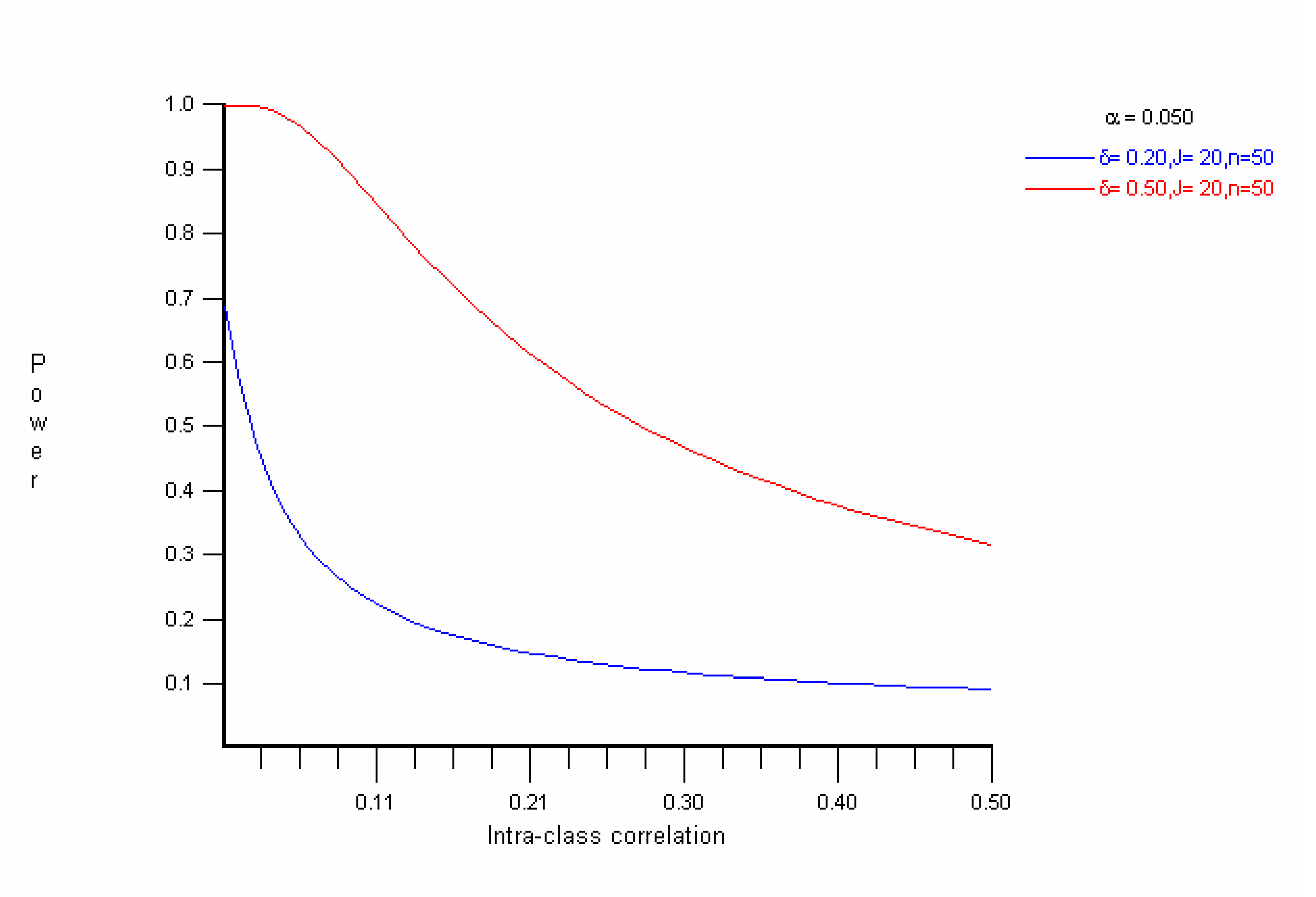
Intraclass Correlation (ρ)

- ICC becomes important in research design when:
  - Random assignment is accomplished at the group level
  - Multistage sampling designs are used
  - Group level predictors or covariates are used
- If there is little difference from one group to another (ICC nears zero), power is similar to the total sample size ignoring the clustering of groups
- The more groups differ (ICC is nonzero), effective sample size for power approaches the number of groups rather than the total number of participants

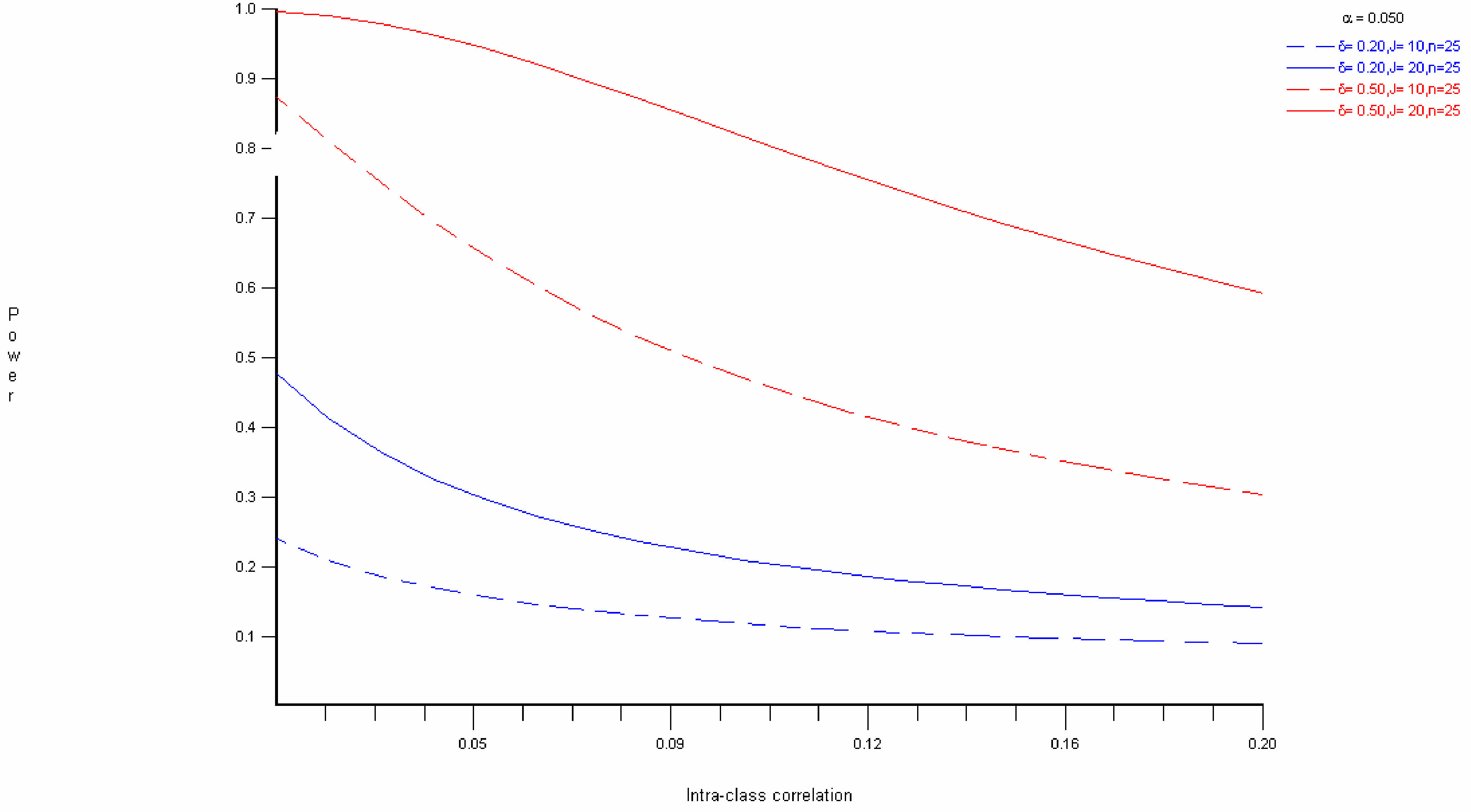
Intraclass Correlation (ρ_I)

- ICC varies with outcome and with type of group and participants
- Small groups that may be more homogenous (e.g., classrooms) are likely to have larger ICCs than large groups with more heterogeneity (e.g., schools or districts)
- What size of ICCs are common?
  - Concentrated between 0.01 and 0.05 for much social science research (Bloom, 2006)
  - Between 0.05 and 0.15 for school achievement (Spybrook et al., 2006)
- The guideline of 0.05 to 0.15 is more consistent with the values of covariate adjusted intraclass correlations; unconditional ICCs may be larger (roughly 0.15 to 0.25; Hedges & Hedberg, in press)
- "It is unusual for a GRT to have adequate power with fewer than 8 to 10 groups per condition" (Murray et al., 2004)

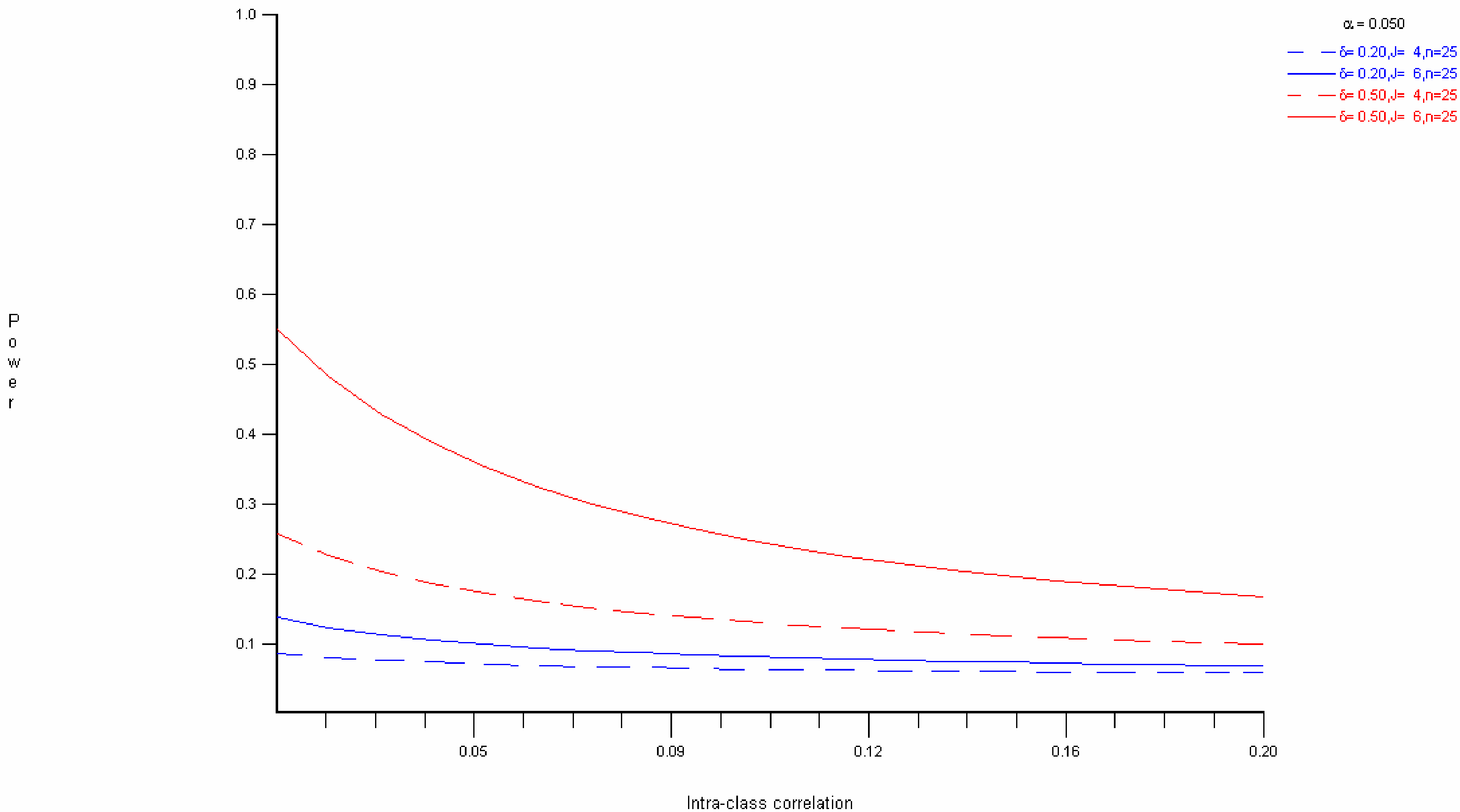
Relationship of ICC and power



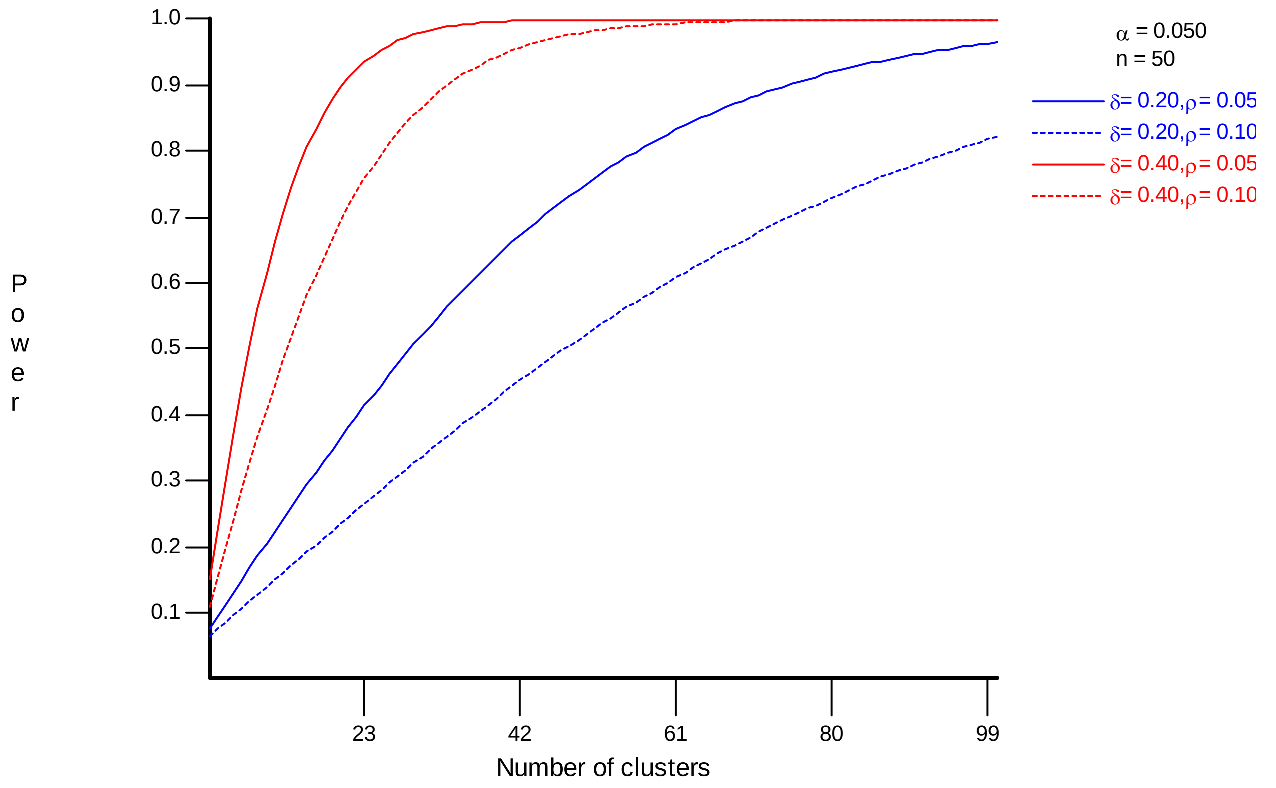
Relationship of ICC, Effect Size, Number of Clusters and Power



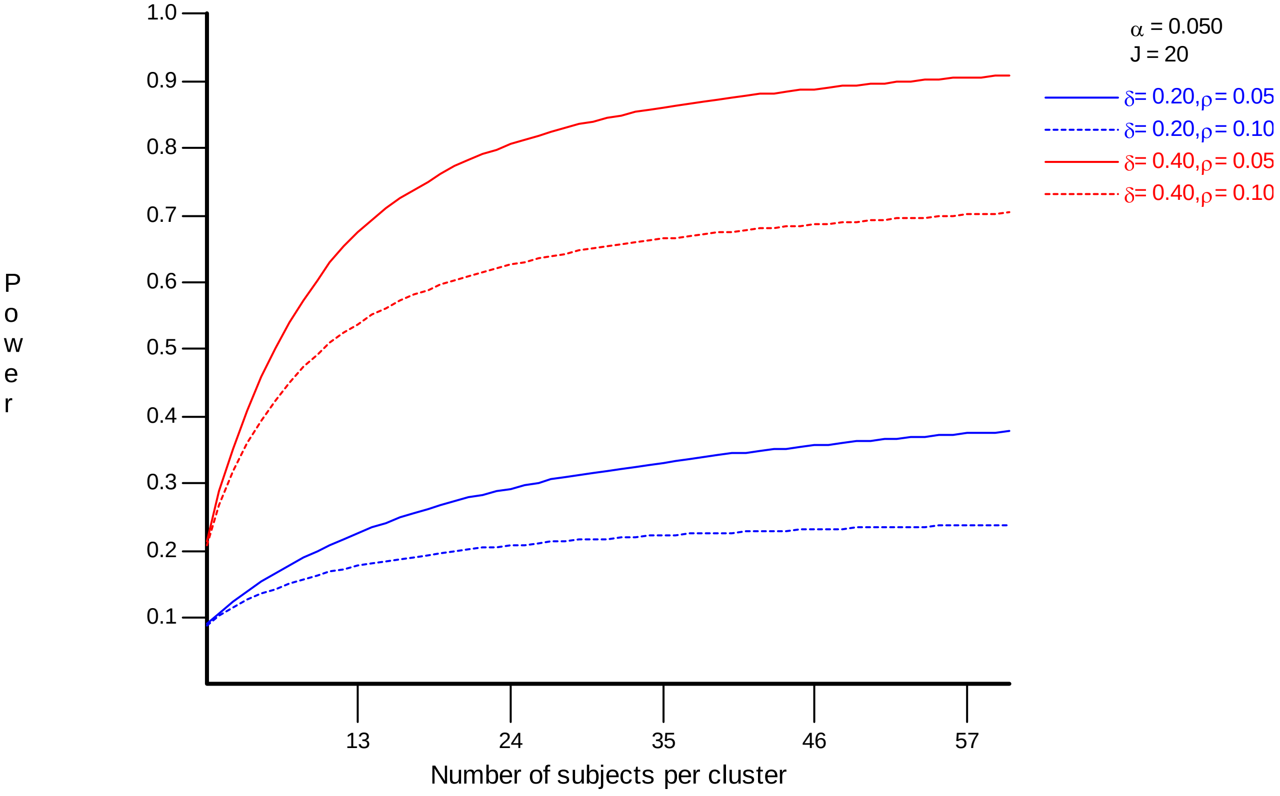
Relationship of ICC, Effect Size, Number of Clusters and Power When J is Small



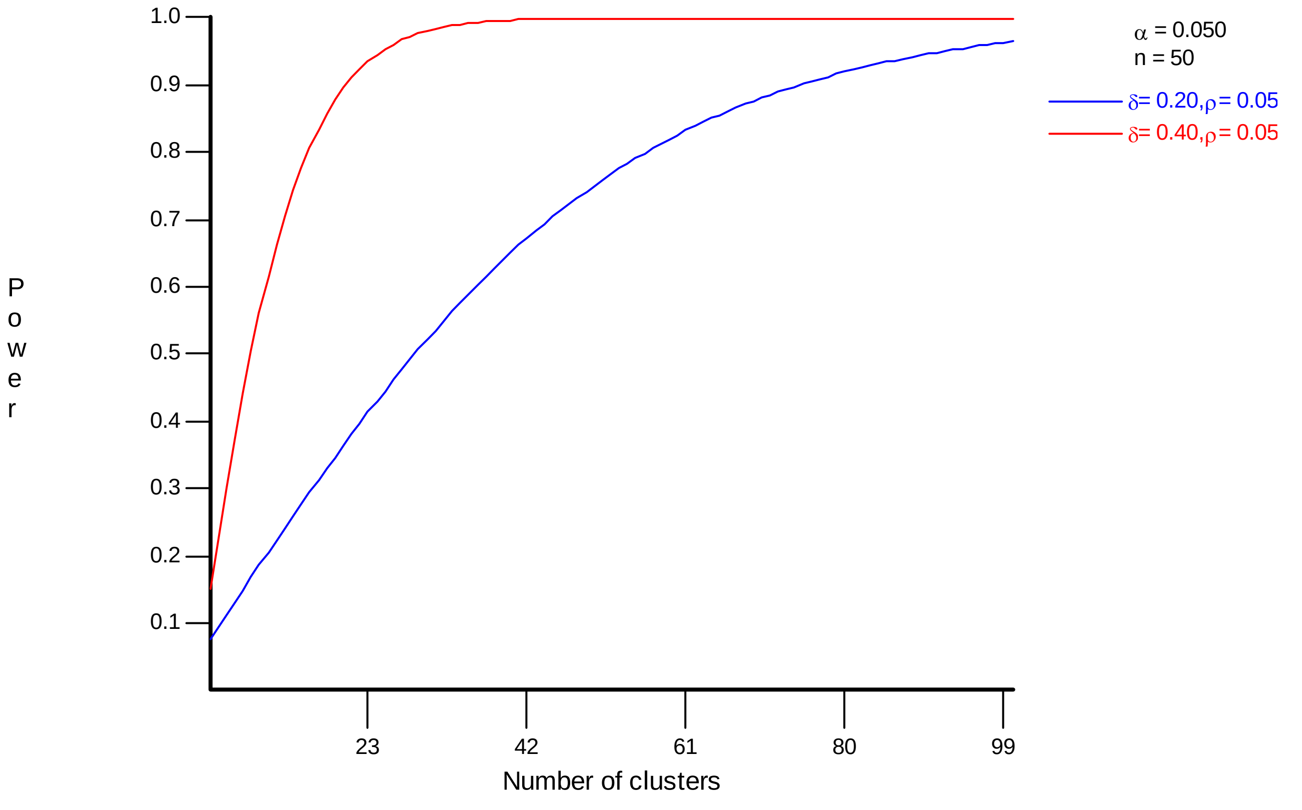
Relationship of ICC, effect size, number of clusters and power



Effect of Cluster Size (n)



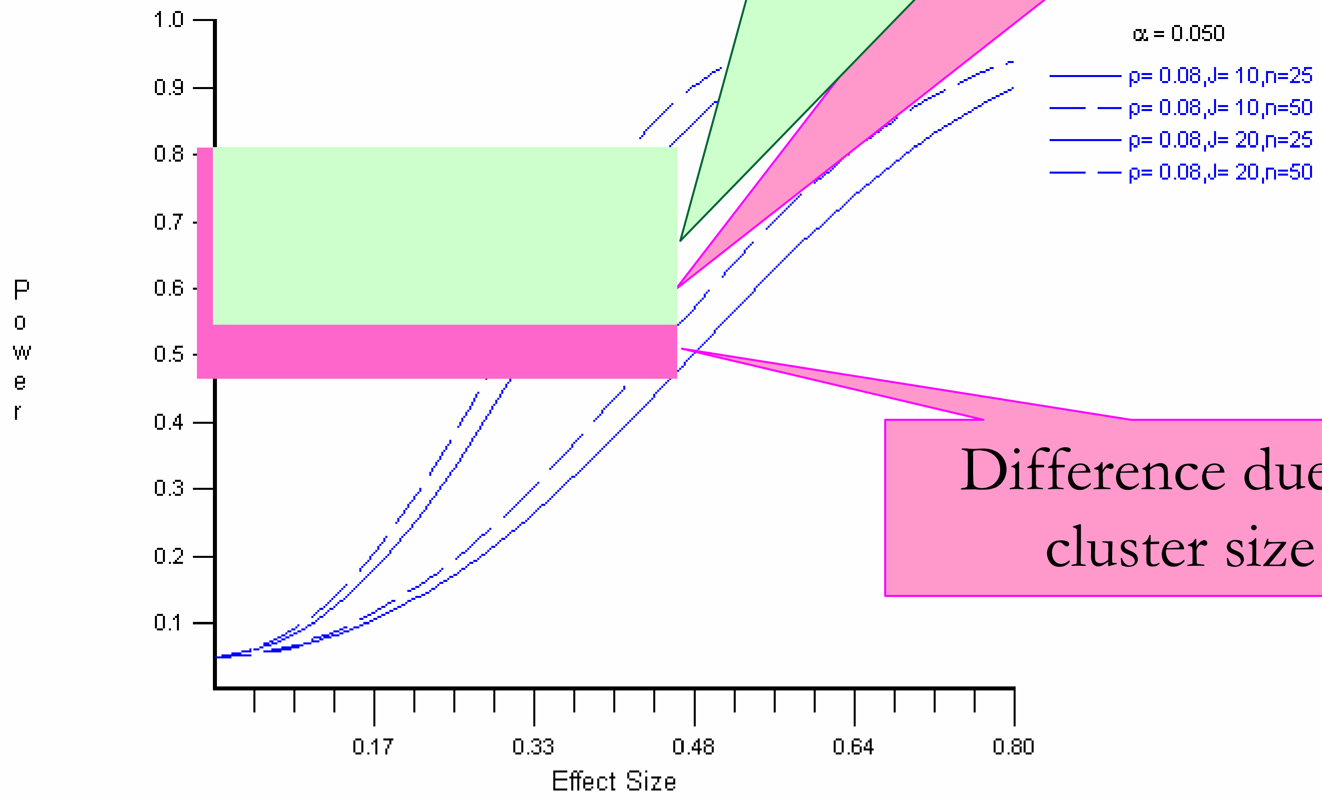
Effect of Number of Clusters (J)



The number of cluster
power than the cluster

Note the difference in power for nj = 500 arranged as 50 per 10 vs. nj = 500 arranged as 25 per 20 clusters

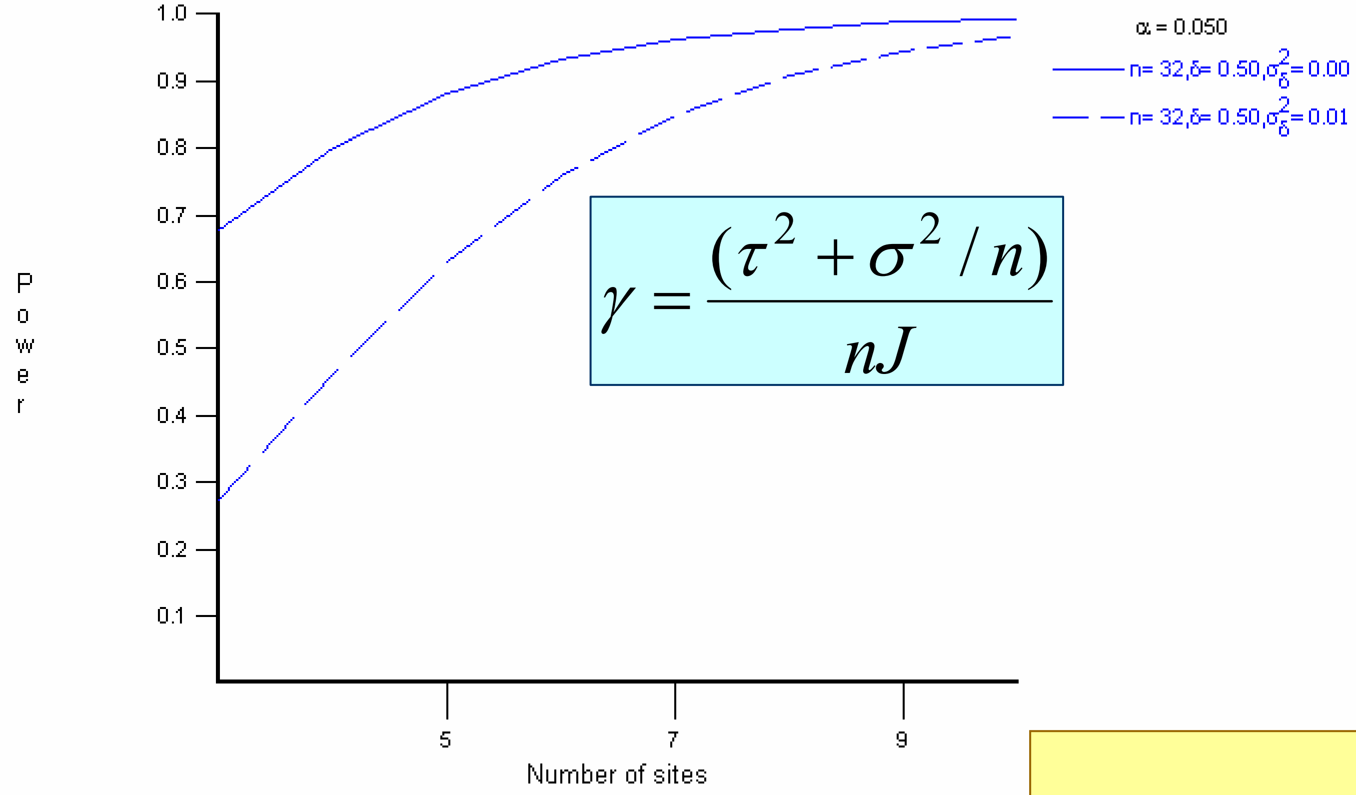
mber of clusters



Difference due to cluster size

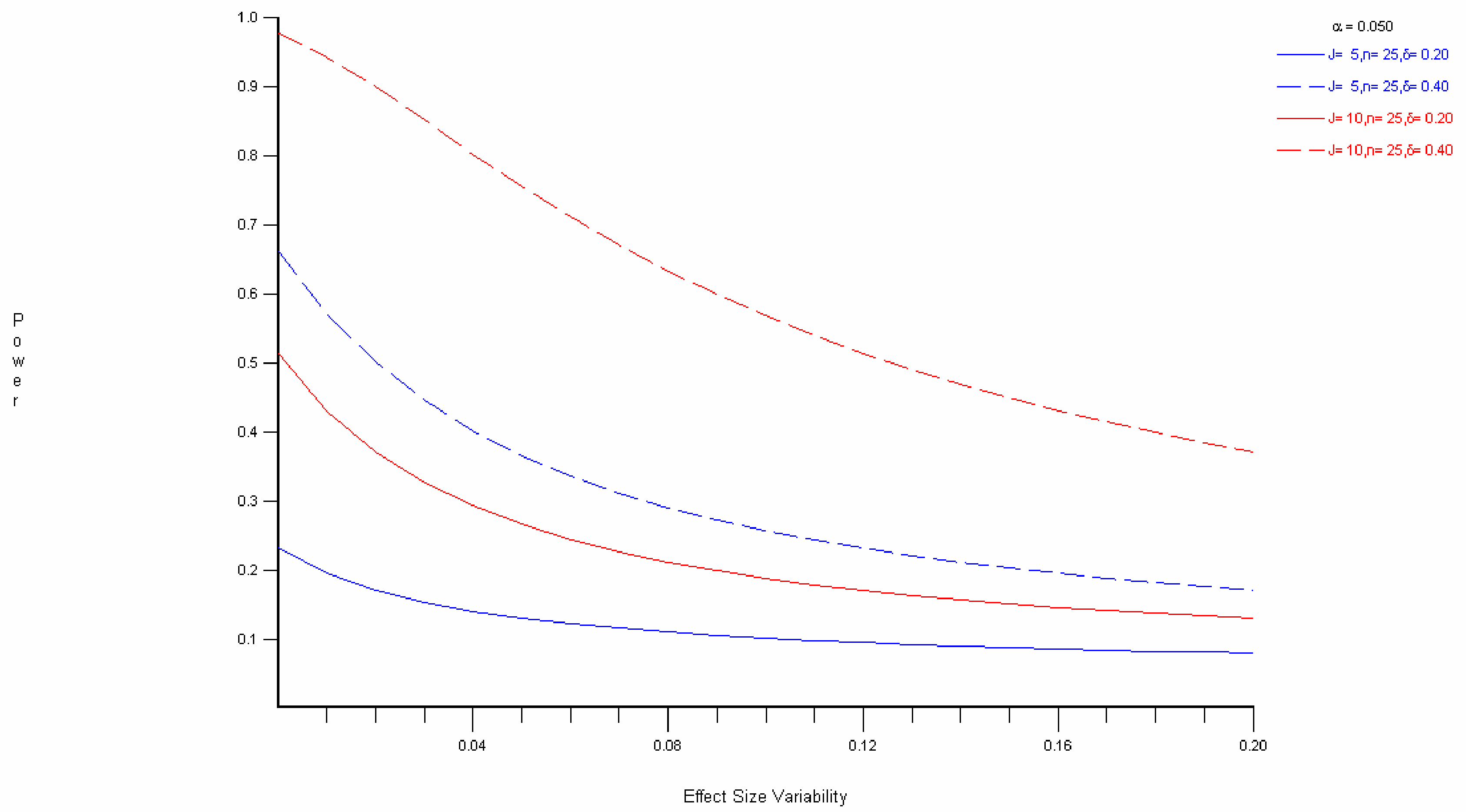
Ignoring Hierarchical Structure vs. Multilevel Modeling

Variance of the treatment effect across clusters



$$\gamma = \frac{(\tau^2 + \sigma^2 / n)}{nJ}$$

Effect of Effect Size Variability (σ_δ^2)



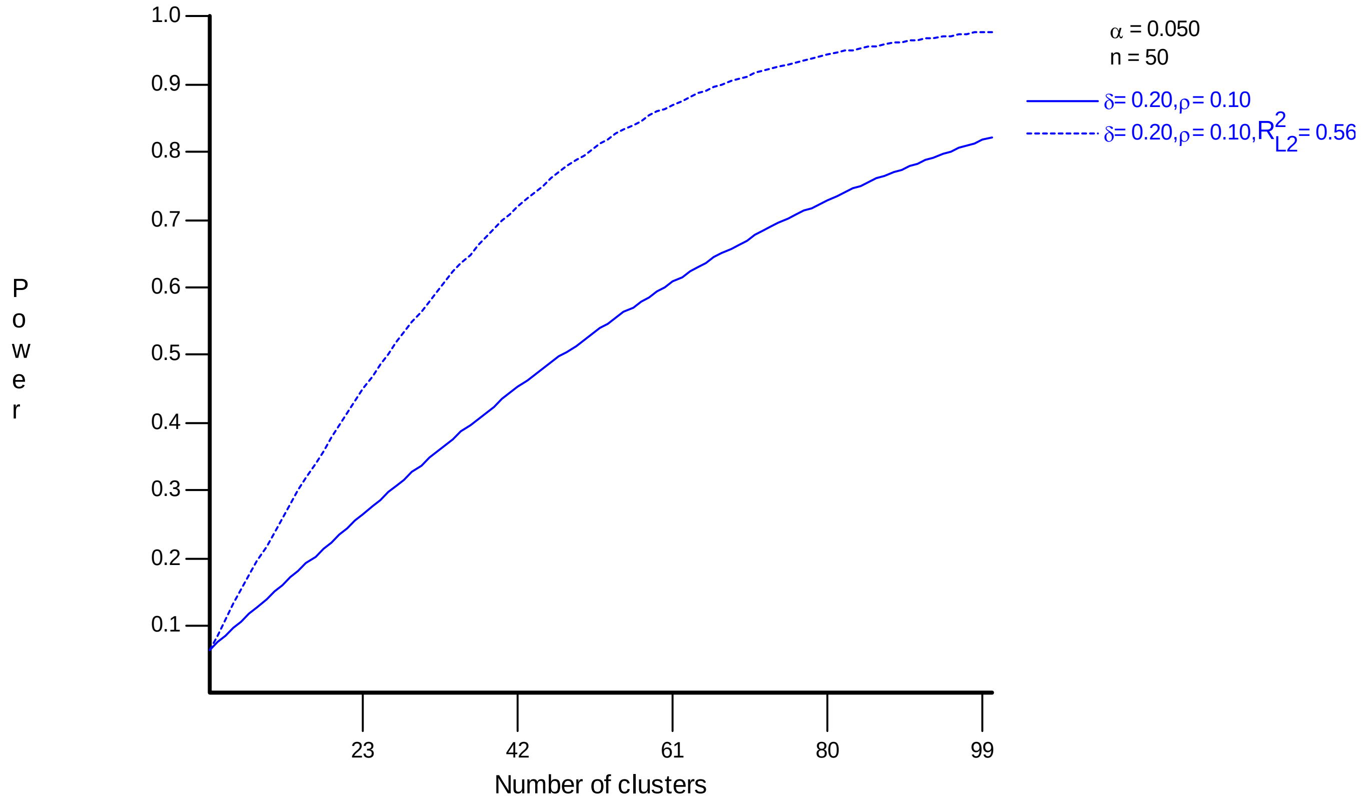
The number of clusters has a stronger influence on power than the cluster size as ICC departs from 0

- The standard error of the main effect of treatment is:

$$SE(\hat{\gamma}_{01}) = \sqrt{\frac{4(\rho + (1-\rho)/n)}{J}}$$

- As ρ increases, the effect of n decreases
- If clusters are variable (ρ is large), more power is gained by increasing the number of clusters sampled than by increasing n

Effect of a Covariate on Power



The Group Effect Multiplier

ICC (ρ)	Randomized group size (n)					
	10	20	50	100	200	500
0.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	1.04	1.09	1.22	1.41	1.73	2.48
0.02	1.09	1.17	1.41	1.73	2.23	3.31
0.03	1.13	1.25	1.57	1.99	2.64	4.00
0.04	1.17	1.33	1.72	2.23	2.99	4.58
0.05	1.20	1.40	1.86	2.44	3.31	5.09
0.06	1.24	1.46	1.98	2.63	3.60	5.56
0.07	1.28	1.53	2.10	2.82	3.86	5.99
0.08	1.31	1.59	2.22	2.99	4.11	6.40
0.09	1.35	1.65	2.33	3.15	4.35	6.78
0.10	1.38	1.70	2.43	3.30	4.57	7.13
0.20	1.67	2.19	3.29	4.56	6.39	10.04

Note: The group effect multiplier equals $\sqrt{1+(n-1)\rho}$; table from Bloom (2006).

The Minimum Detectable Effect Expressed as a Multiple of the Standard Error

Number of groups (*J*)	Two-tailed test	One-tailed test
4	5.36	3.98
6	3.72	3.07
8	3.35	2.85
10	3.20	2.75
12	3.11	2.69
14	3.05	2.66
16	3.01	2.63
18	2.99	2.61
20	2.96	2.60
30	2.90	2.56
40	2.87	2.54
60	2.85	2.52
120	2.83	2.50
infinite	2.80	2.49

Note: The group effect multipliers shown here are for the difference between the mean program group outcome and the mean control group outcome, assuming equal variances for the groups, a significance level of .05, and a power level of .80; table from Bloom (2006).

The Minimum Detectable Effect Size

Intraclass correlation (ρ_I) = 0.01

	Randomized group size (n)					
Number of groups (J)	10	20	50	100	200	500
4	1.77	1.31	0.93	0.76	0.66	0.59
6	1.00	0.74	0.52	0.43	0.37	0.33
8	0.78	0.58	0.41	0.33	0.29	0.26
10	0.67	0.49	0.35	0.29	0.25	0.22
20	0.44	0.32	0.23	0.19	0.16	0.15
30	0.35	0.26	0.18	0.15	0.13	0.12
40	0.30	0.22	0.16	0.13	0.11	0.10
60	0.24	0.18	0.13	0.10	0.09	0.08
120	0.17	0.13	0.09	0.07	0.06	0.06

Note: The minimum detectable effect sizes shown here are for a two-tailed hypothesis test, assuming a significance level of .05, a power level of .80, and randomization of half the groups to the program; table from Bloom (2006).

The Minimum Detectable Effect Size

Intraclass correlation (ρ_I) = 0.05

	Randomized group size (n)					
Number of groups (J)	10	20	50	100	200	500
4	2.04	1.67	1.41	1.31	1.26	1.22
6	1.16	0.95	0.80	0.74	0.71	0.69
8	0.90	0.74	0.62	0.58	0.55	0.54
10	0.77	0.63	0.53	0.49	0.47	0.46
20	0.50	0.41	0.35	0.32	0.31	0.30
30	0.40	0.33	0.28	0.26	0.25	0.24
40	0.35	0.28	0.24	0.22	0.21	0.21
60	0.28	0.23	0.19	0.18	0.17	0.17
120	0.20	0.16	0.14	0.13	0.12	0.12

Note: The minimum detectable effect sizes shown here are for a two-tailed hypothesis test, assuming a significance level of .05, a power level of .80, and randomization of half the groups to the program; table from Bloom (2006).

The Minimum Detectable Effect Size

Intraclass correlation (ρ_I) = 0.10

	Randomized group size (n)					
Number of groups (J)	10	20	50	100	200	500
4	2.34	2.04	1.84	1.77	1.73	1.71
6	1.32	1.16	1.04	1.00	0.98	0.97
8	1.03	0.90	0.81	0.78	0.77	0.76
10	0.88	0.77	0.69	0.67	0.65	0.64
20	0.58	0.50	0.46	0.44	0.43	0.42
30	0.46	0.40	0.36	0.35	0.34	0.34
40	0.40	0.35	0.31	0.30	0.29	0.29
60	0.32	0.28	0.25	0.24	0.24	0.23
120	0.22	0.20	0.18	0.17	0.17	0.16

Note: The minimum detectable effect sizes shown here are for a two-tailed hypothesis test, assuming a significance level of .05, a power level of .80, and randomization of half the groups to the program; table from Bloom (2006).

Using G*Power

- Free software for power estimation available at:

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register

- Estimates power for a variety of situations including *t*-tests, *F*-tests, and χ^2

- G*Power

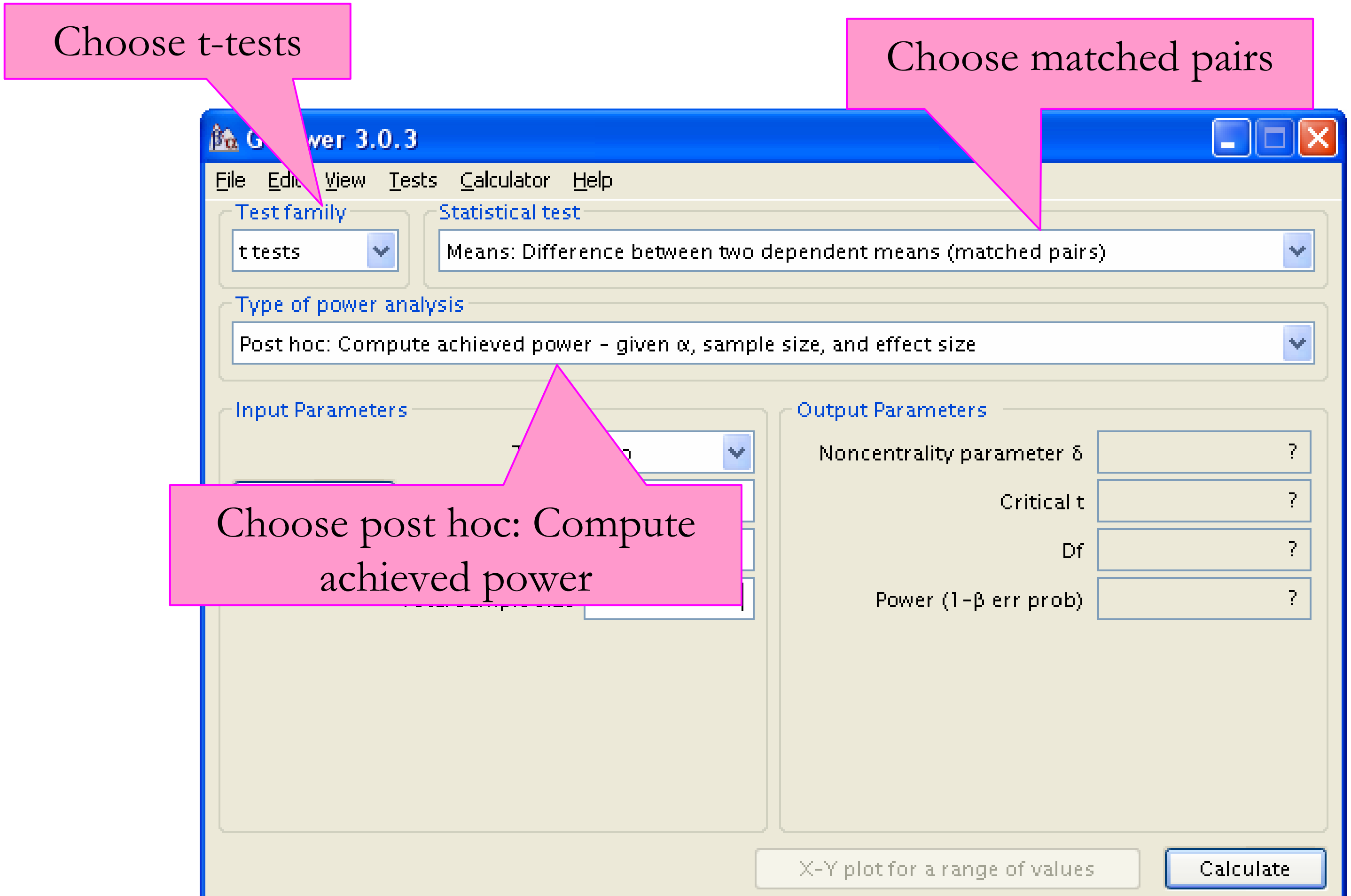
Examples using G*Power

Luft & Vidoni (2002) examined preservice teachers' knowledge about school to career transitions before and after a teacher internship. Some of the obtained results were:

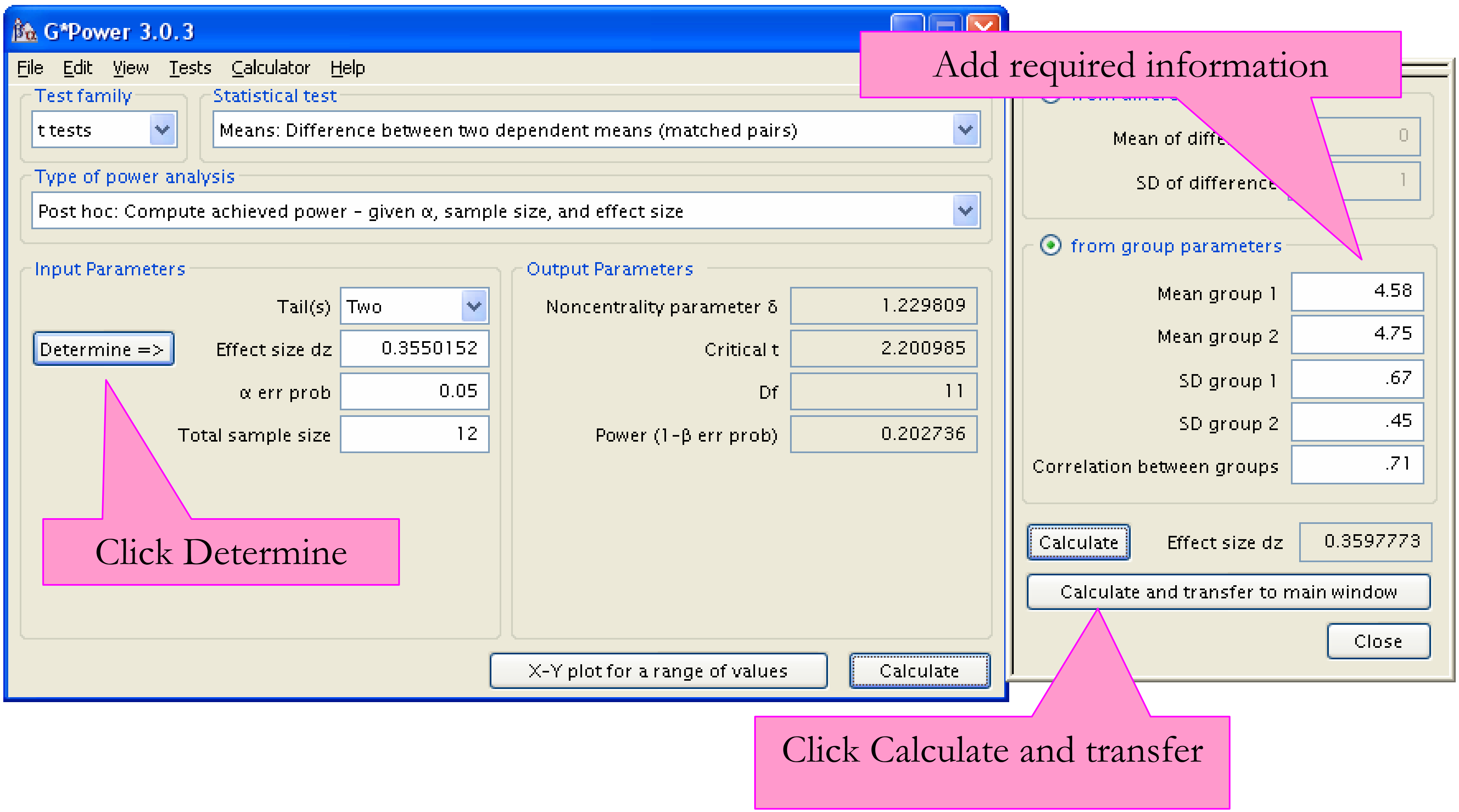
	Before		After				
Knowledge about:	$\overline{X}$	*sd*	$\overline{X}$	*sd*	*t*	*p*	*r*
Writing	2.92	1.44	3.92	.79	-2.25	.05	.59
Use of Hands-on activities	4.58	.67	4.75	.45	-1.00	.34	.71
Class assignments	3.67	.49	4.08	.79	-1.82	.10	.56

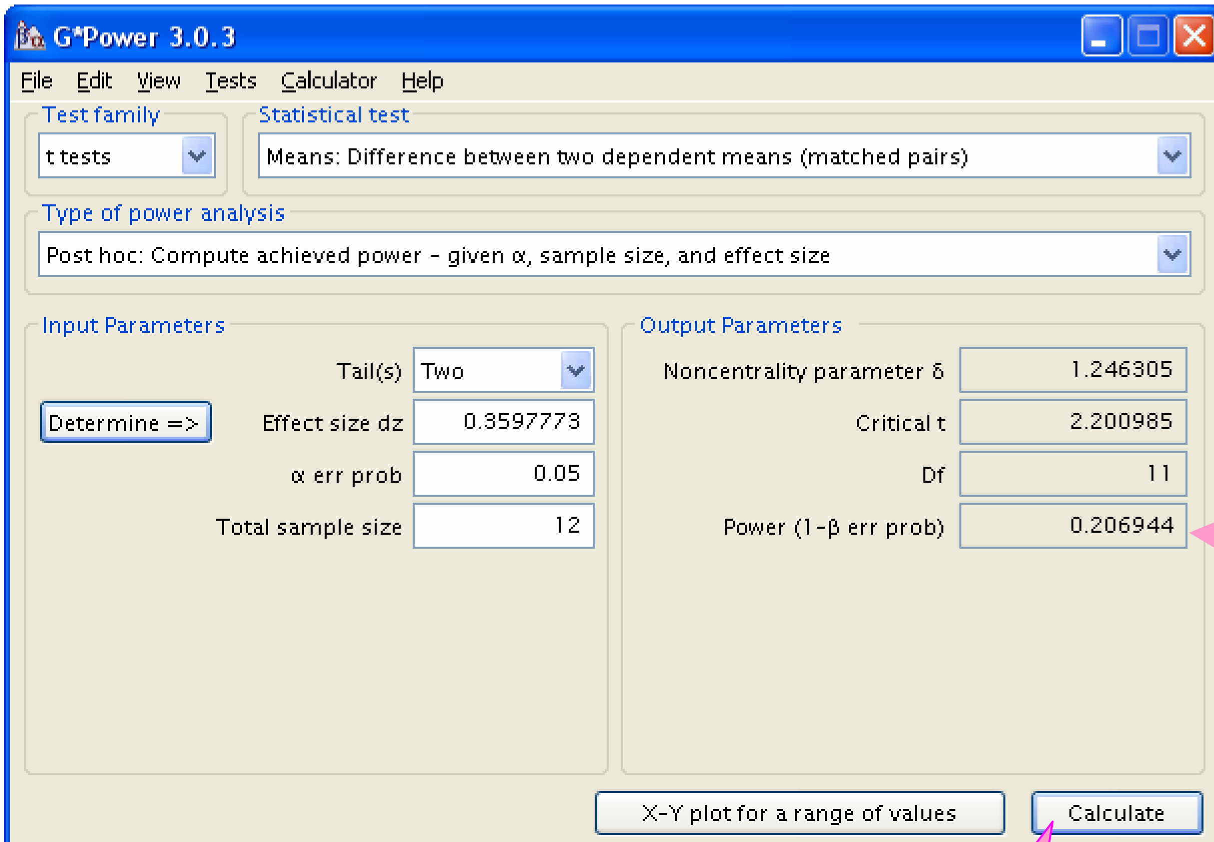
Twelve students participated in the study and completed the pre and post testing.

Example 1. Using G*Power, estimate the power of the repeated measures *t*-test for knowledge of hands-on activities. Use the supplied information in the table.



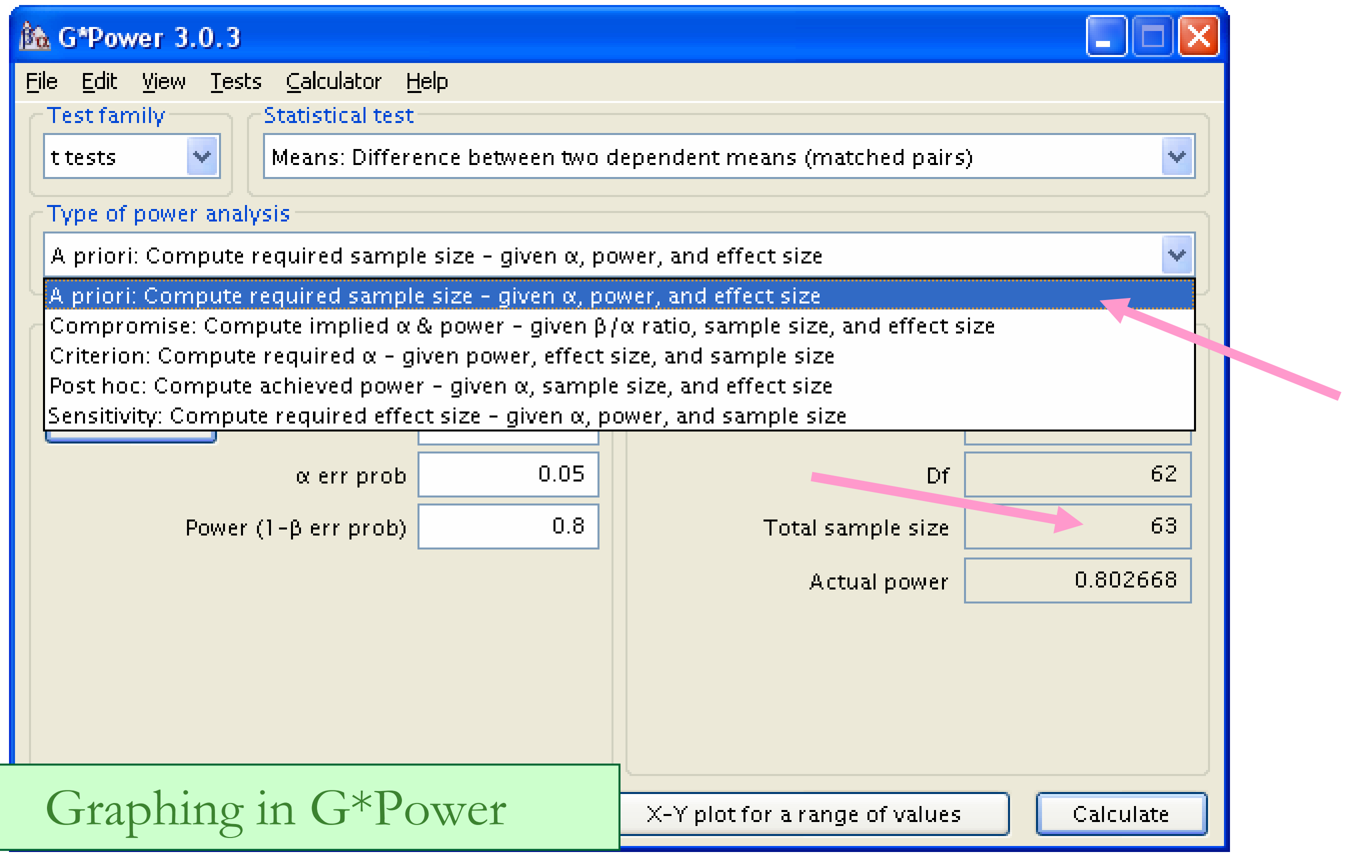
Next calculate an effect size based on the supplied table information:





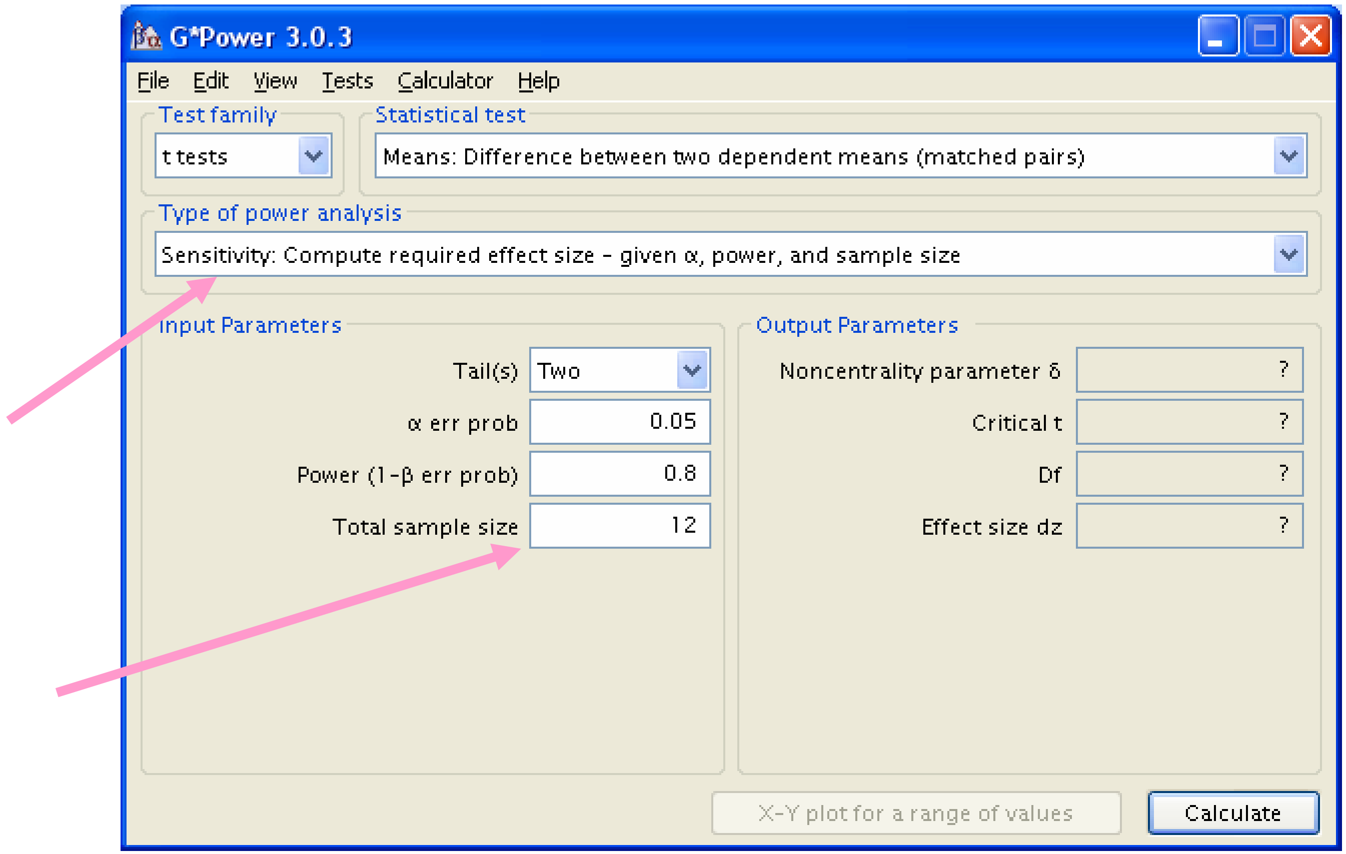
Click calculate

Example 2. Using the same information as example 1, determine the necessary sample size to achieve a power of .80



Graphing in G*Power

Example 3. Continue with the same information and determine the minimum detectable effect size if power is .80



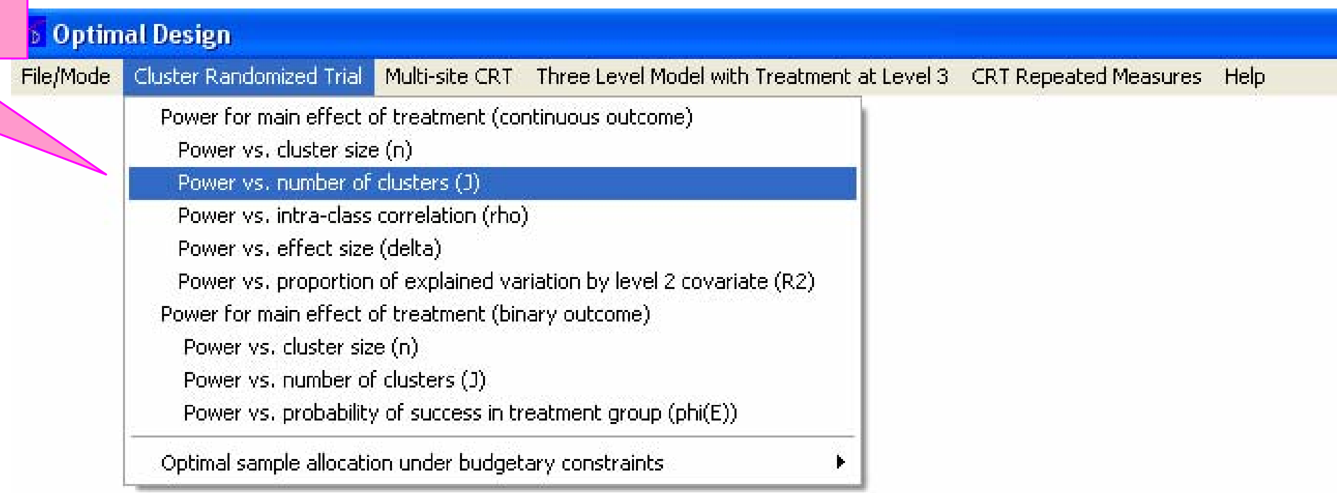
Using the Optimal Design Software

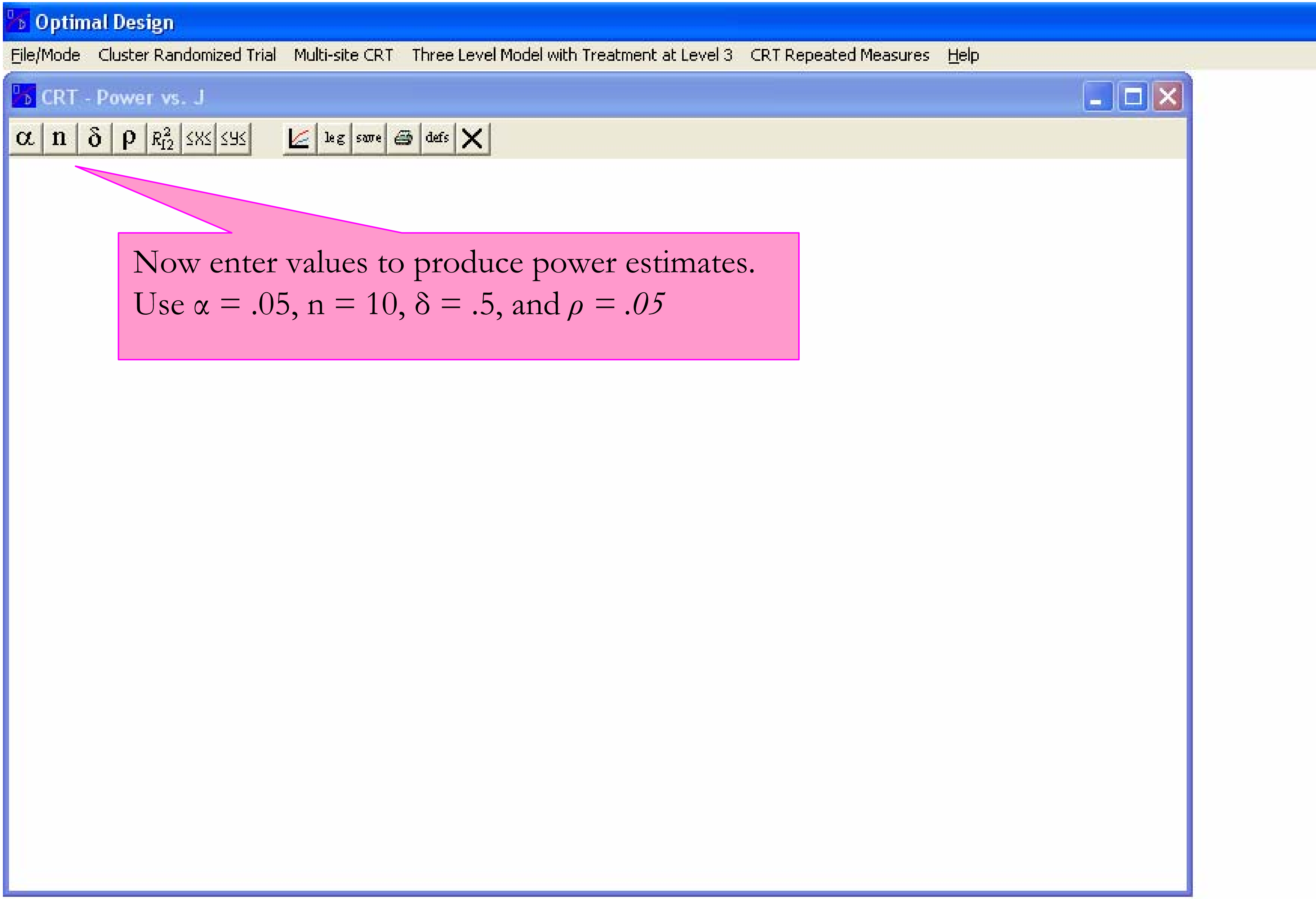
- The Optimal Design Software can also be used to estimate power in a variety of situations
- The particular strength of this software is its application to multilevel situations involving cluster randomization or multisite designs
- Available at:

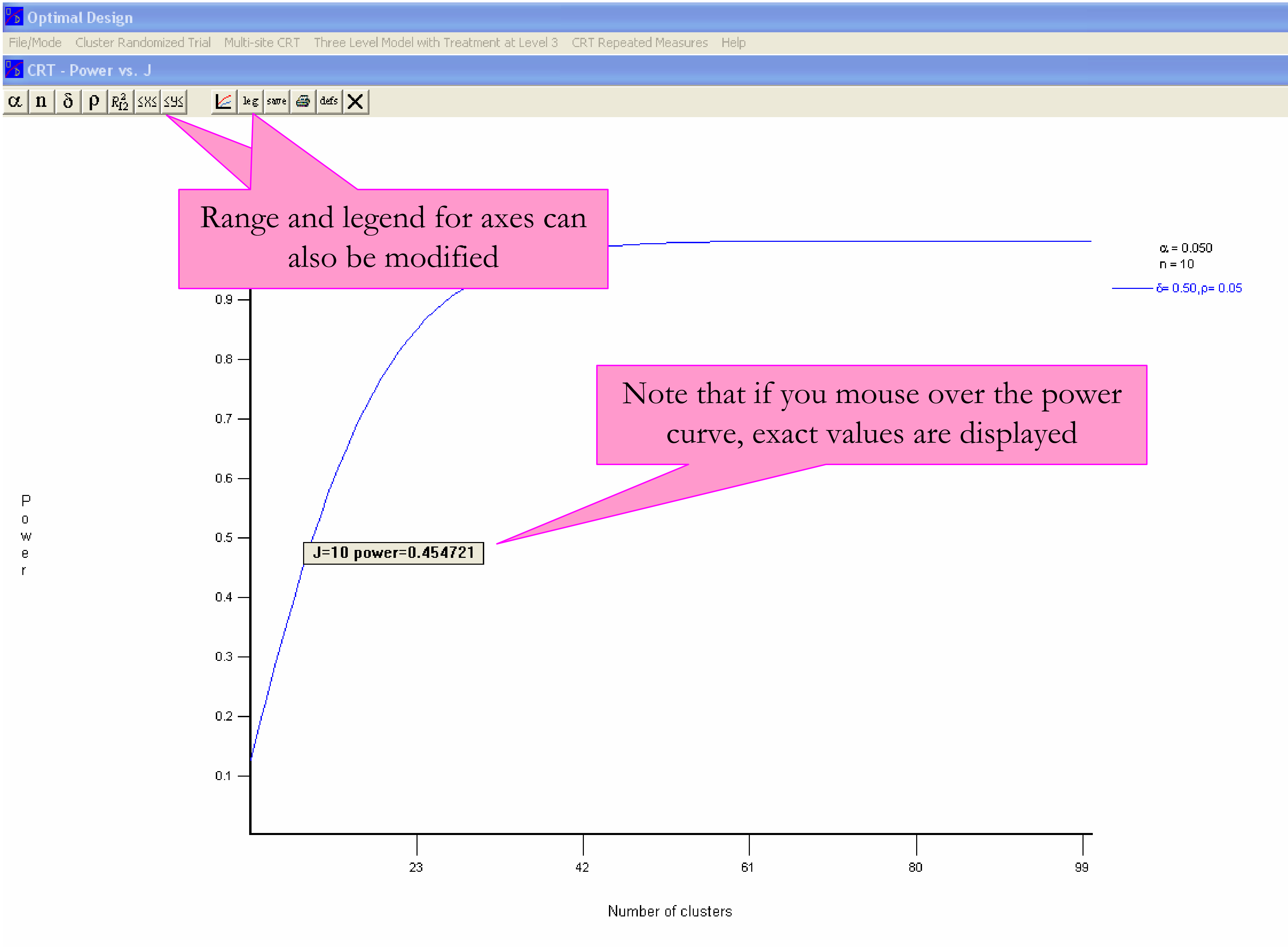
  http://sitemaker.umich.edu/group-based/optimal_design_software
- Optimal Design

Using Optimal Design (OD), estimate the power for a group randomized study under several conditions. Start by choosing "File/Mode" on the toolbar and then "Optimal Design for Group Randomized Trials"

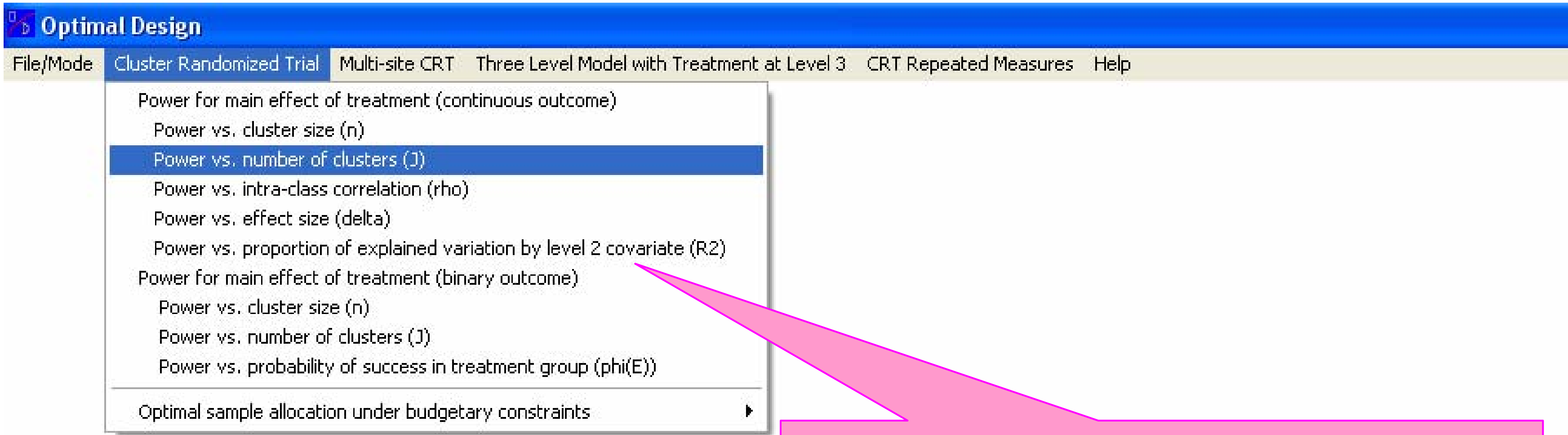
Next choose Power vs. number of clusters




Optimal Design
File/Mode
Cluster Randomized Trial
Multi-site CRT
Three Level Model with Treatment at Level 3
CRT Repeated Measures
Help
CRT - Power vs. J
leg
save
defs
Now enter values to produce power estimates.
Use α = .05, n = 10, δ = .5, and ρ = .05

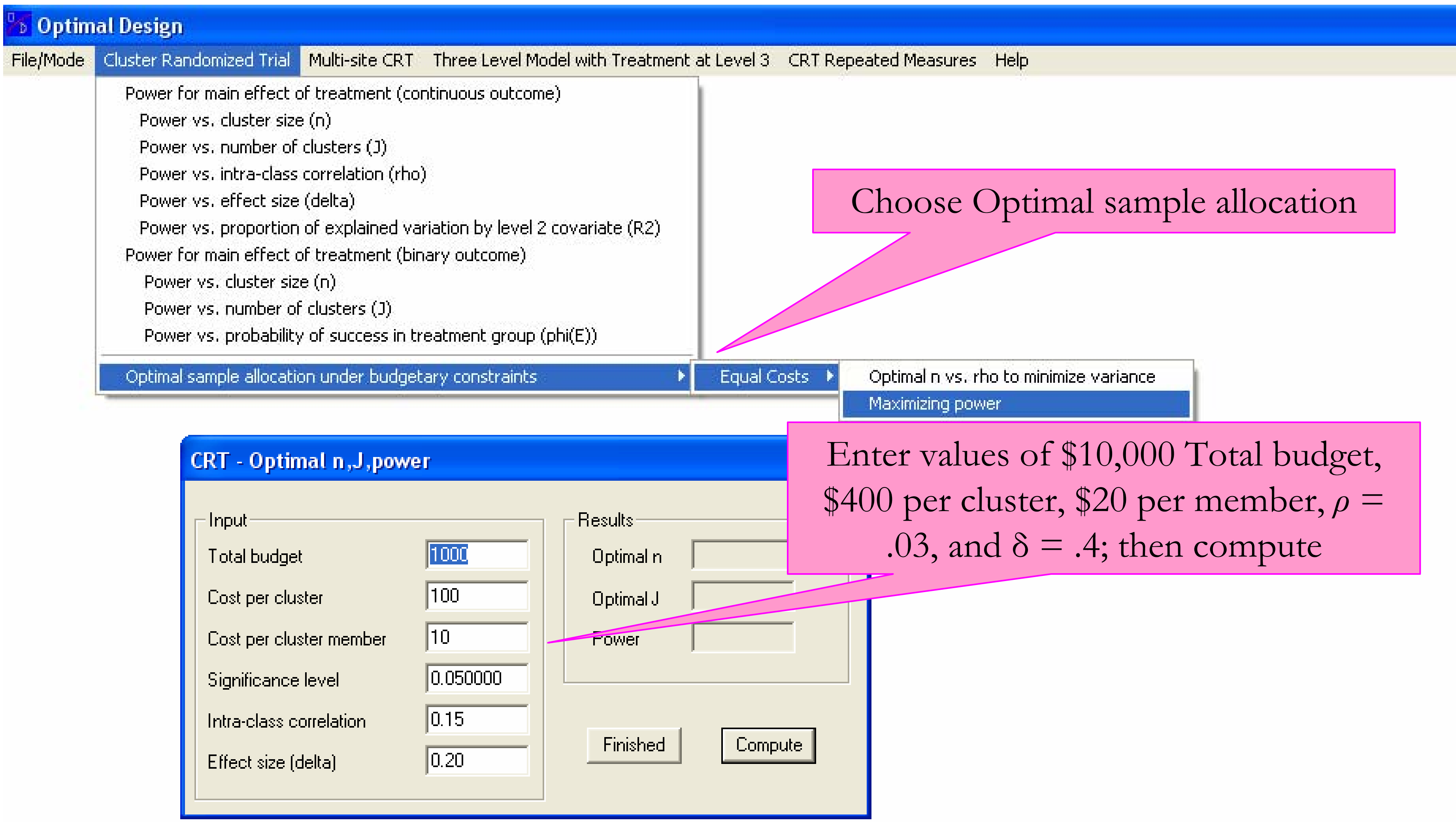


Optimal Design
File/Mode
Cluster Randomized Trial
Multi-site CRT
Three Level Model with Treatment at Level 3
CRT Repeated Measures
Help
CRT - Power vs. J
Range and legend for axes can also be modified
Note that if you mouse over the power curve, exact values are displayed
J=10 power=0.454721
α = 0.050
n = 10
δ= 0.50,ρ= 0.05
Power
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
23
42
61
80
99
Number of clusters



Now explore the use of OD for examining power as a function of n, ρ, δ, and R^2

The OD software can also be used to determine the best combination of design features under cost constraints



Optimal Design

What if the ICC was lower, .01?

What if the ICC was higher, .08?

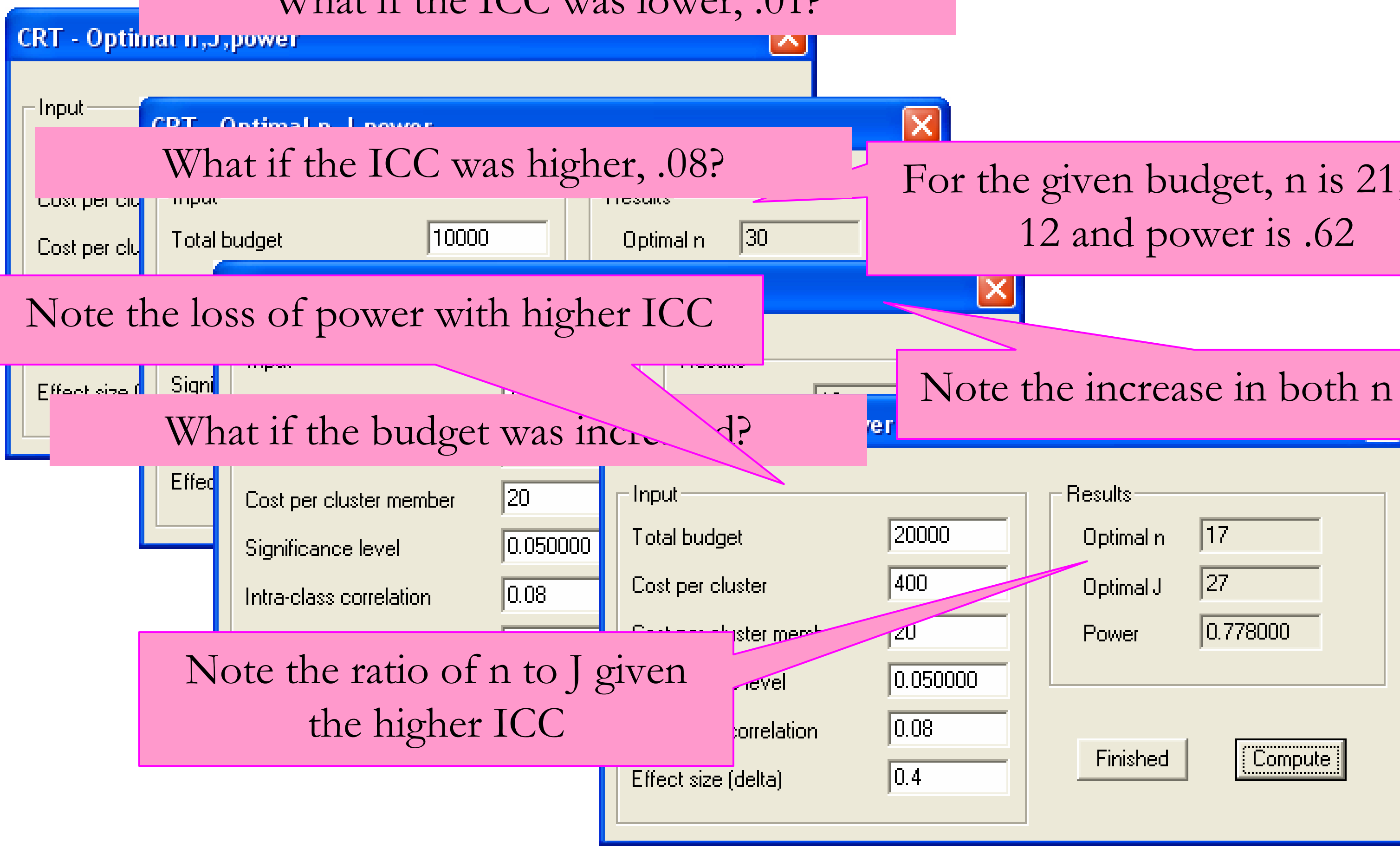
For the given budget, n is 21, J is 12 and power is .62

Note the loss of power with higher ICC

Note the increase in both n and power

What if the budget was increased?

Note the ratio of n to J given the higher ICC



One Last Example: Multisite CRT

- The primary rationale in this approach is to extend the idea of blocking to the multilevel situation
- Clusters are assigned to blocks with other similar clusters and then randomly assigned to treatment
- Blocking creates greater homogeneity and less residual variance, thereby increasing power
- For example, schools are collected into blocks based on whether school composition is low, medium, or high SES
- Schools are within each block are randomly assigned to treatment
- Between school SES variability is controlled by the blocking

Multisite CRT

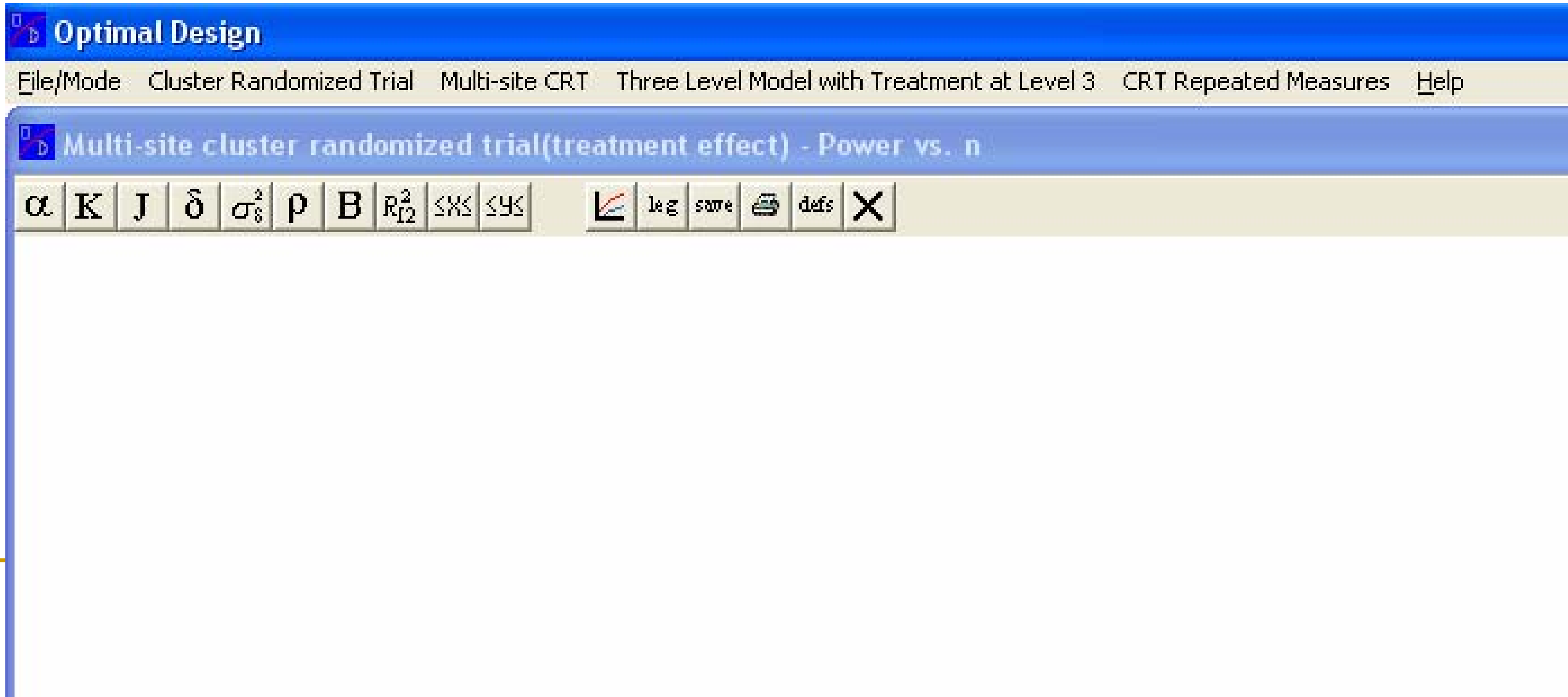
Two additional parameters are used in estimation:

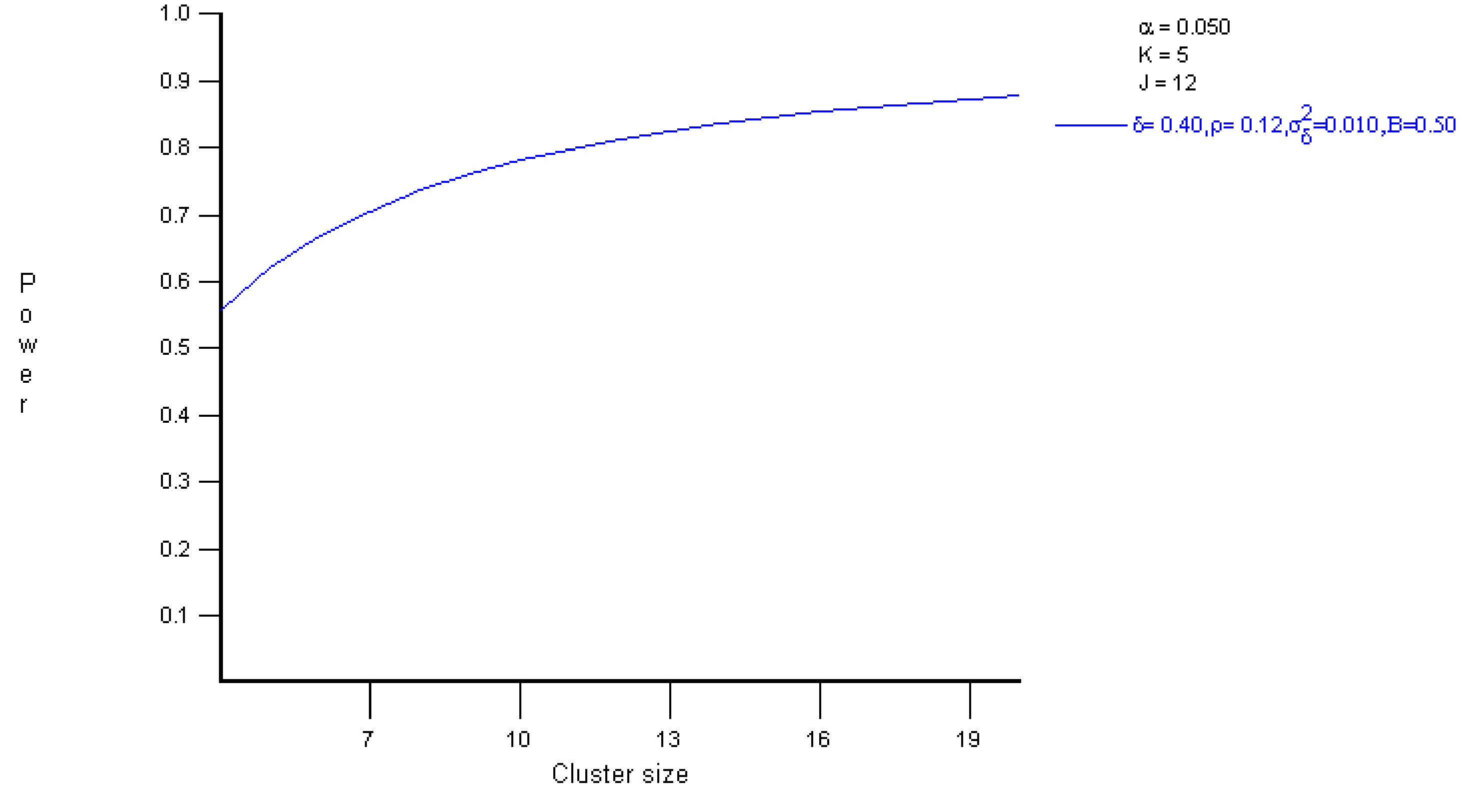
- Number of sites or blocks, K
- The effect size variability, σ_δ^2
- σ_δ^2 represents the variability of effect size from one cluster to another within a site
- This variability represents within site replications of the study

Multisite CRT

Example:

- 5 cities, 12 schools per city, d = .4, ICC =.12, σ_δ^2 = .01, blocking accounts for 50% of the variation in the outcome





1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
P
o
w
e
r
7
10
13
16
19
Cluster size
α = 0.050
K = 5
J = 12
δ= 0.40,ρ= 0.12,σ$^2_δ$=0.010,B=0.50

Applications

- For the remainder of the workshop you may
  - complete exercises on power estimation
  - calculate power estimates for your own research
- Exercises can be downloaded from:

  http://www.uoregon.edu/~stevensj/workshops/exercises.pdf
- When you finish the exercises, you can obtain answers at:

  http://www.uoregon.edu/~stevensj/workshops/answers.pdf
- Discussion as time permits

Bibliography

Bloom, H. S. (2006). *Learning More from Social Experiments: Evolving Analytic Approaches*. New York, NY: Russell Sage Foundation Publications.

Boling, N. C., & Robinson, D. H. (1999). Individual study, interactive multimedia, or cooperative learning: Which activity best supplements lecture-based distance education? Journal of Educational Psychology, 91, 169-174.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A Power Primer, Psychological Bulletin, 112, 155-159.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997– 1003.

Cooper, H., & Hedges, L. (1994). The Handbook of Research Synthesis. New York, NY: Russel Sage Foundation.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–575.

Elashoff, J. D. (2002). *NQuery Advisor Version 5.0 User's Guide*. Los Angeles, CA: Statistical Solutions Limited.

Elmore, P., & Rotou, O. (2001, April). *A primer on basic effect size concepts*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Hallahan & Rosenthal (1996). Statistical Power: Concepts, Procedures and Applications, *Behavior Research and Therapy,* 34, 489-99.

Harlow, L. L. Mulaik, S. A. , & Steiger, J. H. (1997). *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

Hays, W. L. (1963). *Statistics for psychologists*. New York: Holt, Rinehart & Winston.

Hedges, L. V., & Hedburg, E. C. (in press). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis.*

Huberty, C. (2002). A History of Effect Size Indices, *Educational and Psychological Measurement, 62*, 227-240.

Luft, V. D., & Vidoni, K. (2002). Results of a school-to-careers preservice teacher internship program, *Education, 122*, 706-714.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments, *American Journal of Public Health, 94*, 423-432.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials, *Psychological Methods, 2*(2), 173-185.

Rosenthal & Gaito (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55*, 33-38.

Rosenthal, R. & Rosnow, R. L. (1991). *Essentials of behavioral research* (2nd Ed.). New York: McGraw-Hill, Inc.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166-169.

Spybrook, J., Raudenbush, S., & Liu, X.-f. (2006). *Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design Software*. New York: William T. Grant Foundation.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25* (2), 26– 30.

Thompson (2002). What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes, *Educational Researcher, 31*, 25-32.

Wilkinson, L. & Task Force on Statistical Inference (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations, *American Psychologist, 54* (8), 594–604. [Retrieved from: http://www.apa.org/journals/amp/amp548594.html#c1].