Power, Sample Size, Effect Size: Considerations for Research

Carol B. Thompson

JH Biostatistics Center

SON Brown Bag – November 20, 2012

Research Approaches

- Comparisons – statistical hypotheses
- Estimates – precision (confidence intervals)

Population vs Research Views

What conclusion is reached by the researcher?	What is true in the real world? →	There is no effect (null = true)	There is an effect (null = false)
	No effect (ES = 0)	Correct conclusion ($p = 1 - \alpha$)	Type II error (p = b)
	There is an effect (ES ≠ 0)	Type I error (p = a)	Correct conclusion (p = 1 − b)

Figure 3.2 Four outcomes of a statistical test

(ELLIS)

Type I and Type II Errors (Which is Worse Risk?)

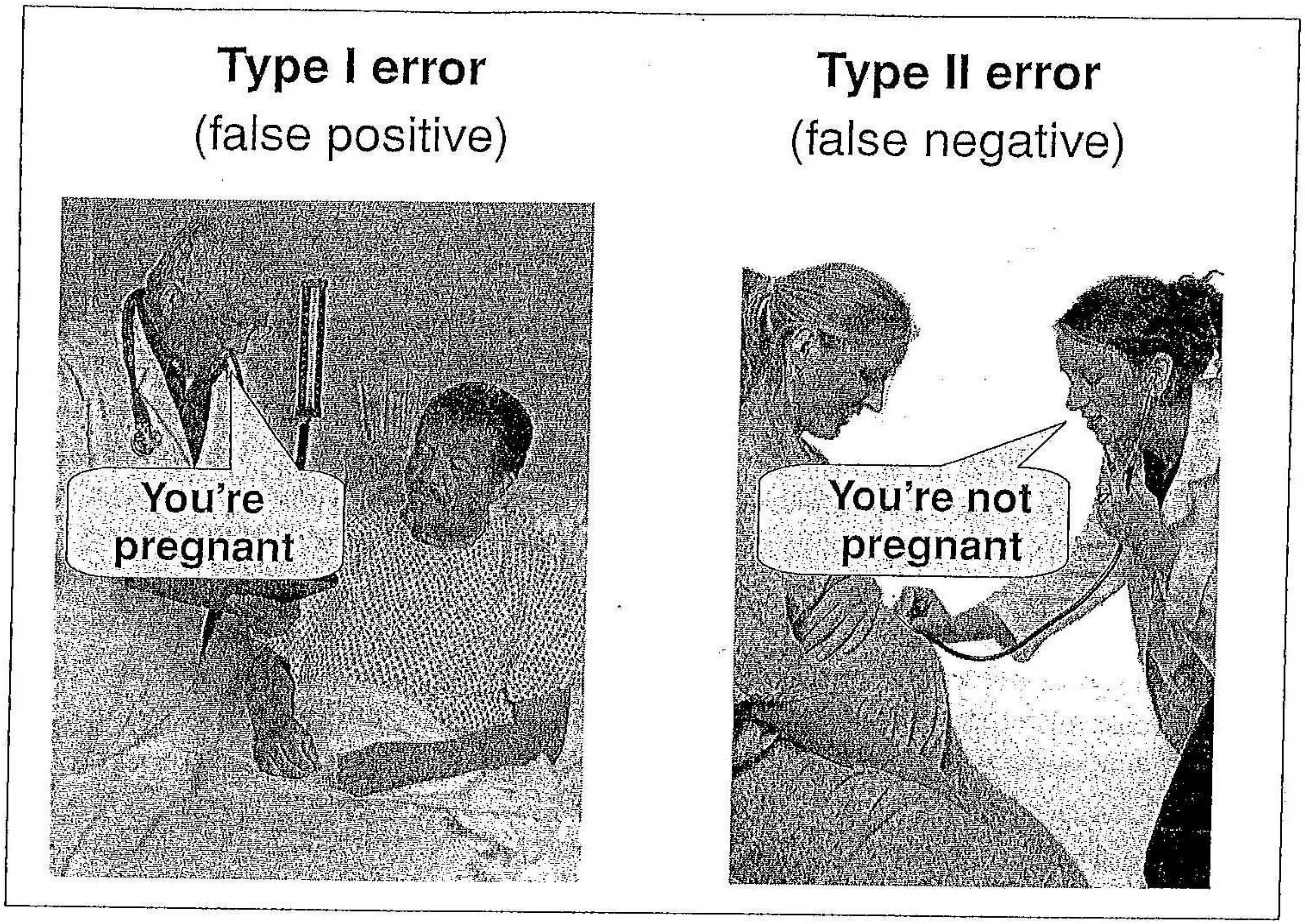

Figure 3.1 Type I and Type II errors (ELLIS)

Related Parameters for Prospective Analysis

- Effect Size
- Sample Size
- α
- Power ($1-\beta$)

Parameters for α and β

Thus, for any fixed $\Delta = \mu_1 - \mu_0$, two types of errors can occur, a false positive Type I error with probability α and a false negative Type II error with probability β, as presented by the following:

	H_0	H_1: $\mu_1 - \mu_0 = \Delta$
Reject: +	α	$1 - \beta(\Delta, N, \alpha)$
Fail to Reject: −	$1 - \alpha$	$\beta(\Delta, N, \alpha)$
	1.0	1.0

(3.9)

(LACHIN)

α vs β

- α doesn't rely on any of the other parameters
- β or power relies on 3 parameters (N, α, ES)
  - Which relate to a specific H_A
- For same sample size and ES, lower α → higher β

Comparing Two Means

The formula for the sample size required to compare two population means, μ_0 and μ_1, with common variance, σ^2, is (VAN BELLE)

$$n = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2}. \quad (2.5)$$

This equation is derived from equation (2.1). For $\alpha = 0.05$ and $\beta = 0.20$ the values of $z_{1-\alpha/2}$ and $z_{1-\beta}$ are 1.96 and 0.84, respectively; and $2(z_{1-\alpha/2} + z_{1-\beta})^2 = 15.68$,

Choosing Power Level - 1

- Underpowered study
  - Waste resources; can't reject H_0
  - Can misdirect future studies if results are NS
  - Unethical if subjecting individual to inferior treatment
- Overpowered study
  - Waste resources?
    - Pick up essentially trivial results – meaningless?
    - Costs of collecting data > benefits

Choosing Power Level - 2

- Balance between risks
- Power of 0.8 due to Jacob Cohen
- Generally Type I error is considered worse
- If can tolerate 5% α, can tolerate 20% β
- Meant as a guideline in considering competing risks, but taken as more absolute these days.

Effect Size

- Practical vs statistical significance of results
- Based on:
  - Carefully chosen samples in comparable popns
  - General/dimensionless value
    - Jargon-free language
    - Allows comparison of disparate research results
- Less reliance on just p-values; more information

Effect Size Types

- 70+ varieties
- d family – difference between groups
- r family – association between measures
- Can convert between r and d ES, if needed

d Effect Sizes - 1

- Dichotomous outcomes
  - Difference in probabilities
  - Risk ratio or relative risk
  - Odds ratio

d Effect Sizes - 2

- Continuous Outcomes (e.g. 2 groups)
  - Difference between 2 means in SD units
  - SD options
    - Cohen's D – If SDs are roughly the same, use pooled SD.
    - Glass' Δ - If SDs are not homogenous, use control's SD (not affected by treatment).
    - Hedges' g – If SDs are not homogenous and different N's, use weighted SD relative to Ns.

r Effect Size

- Pearson's r, Spearman's ρ, Kendall's τ
- Proportion of variance: r^2, R^2, adjusted R^2
- Eta^2 → % of variance based on group diffs
- Cohen's f or f^2 → incremental effect of adding β to basic model

Relative Effect Size Examples - 1

Table 2.1 *Cohen's effect size benchmarks*

Test	Relevant effect size	Effect size classes		
		Small	Medium	Large
Comparison of independent means	d, Δ, Hedges' g	.20	.50	.80
Comparison of two correlations	q	.10	.30	.50
Difference between proportions	Cohen's g	.05	.15	.25
Correlation	r	.10	.30	.50
	r^2	.01	.09	.25
Crosstabulation	w, φ, V, C	.10	.30	.50
ANOVA	f	.10	.25	.40
	η^2	.01	.06	.14
Multiple regression	R^2	.02	.13	.26
	f^2	.02	.15	.35

Notes: The rationale for most of these benchmarks can be found in Cohen (1988) at the following pages: Cohen's d (p. 40), q (p. 115), Cohen's g (pp. 147–149), r and r^2 (pp. 79–80), Cohen's w (pp. 224–227), f and η^2 (pp. 285–287), R^2 and f^2 (pp. 413–414). (ELLIS)

Relative Effect Size Examples - 2

Table 1.2 Measures of Effect Size, Their Use, and a Rough Guide to Interpretation

Effect Size	Common Use/Presentation	Small	Medium	Large
Φ (also known as V or w)	Omnibus effect for χ^2	0.10	0.30	0.50
h	Comparing proportions	0.20	0.50	0.80
d	Comparing two means	0.20	0.50	0.80
r	Correlation	0.10	0.30	0.50
q	Comparing two correlations	0.10	0.30	0.50
f	Omnibus effect for ANOVA/ regression	0.10	0.25	0.40
η^2	Omnibus effect for ANOVA	0.01	0.06	0.14
f^2	Omnibus effect for ANOVA/ regression	0.02	0.15	0.35
R^2	Omnibus effect for regression	0.02	0.13	0.26

(ABERSON)

Choosing Effect Size

- Are effects meaningful ?
  - convert to actual units
- What are raw differences you wish to detect?
- Previous studies may overrepresent larger effects because of publication bias
  - Consider lowest ES as conservative
- Pilot study

Relationships Between 4 Parameters

- For same N and α, ES ↑→ power ↑
- For same ES and α, N ↑→ power ↑
- For same N and ES, α ↓→ power ↓
- For same N and power, ES ↑ → α ↓

Sample Size/Power by Effect Size

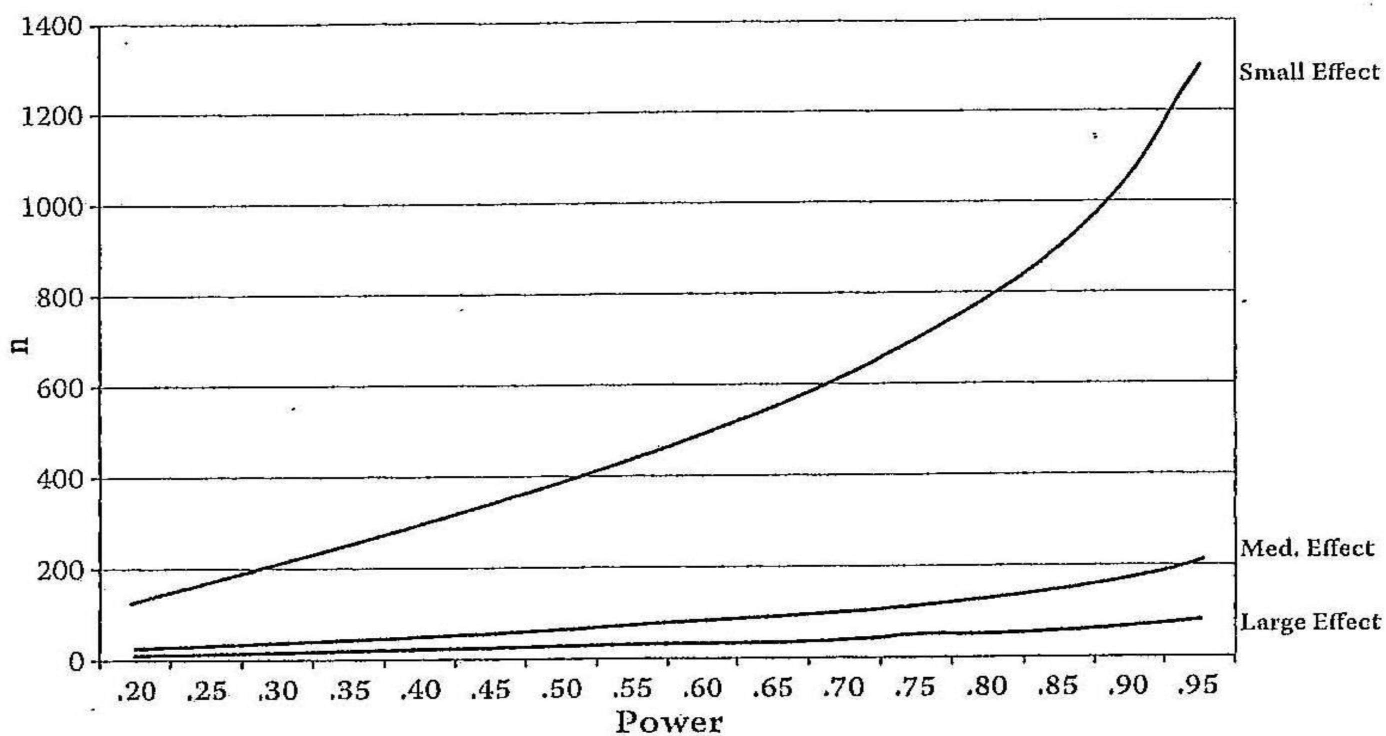

Figure 1.7 Sample size and power for small, medium, and large effects. (ABERSON)

Sample Size for r and d Effect Sizes (Ellis)
α = 0.05, power = 0.8

Table 3.2 *Smallest detectable effects for given sample sizes*

	r		*d*	
Sample size	One-tailed	Two-tailed	One-tailed	Two-tailed
10	.705	.761	1.725	2.024
20	.526	.579	1.156	1.325
30	.437	.485	.931	1.060
40	.382	.426	.801	.909
50	.344	.384	.713	.809
60	.315	.352	.650	.736
70	.292	.327	.600	.679
80	.274	.307	.561	.634
90	.259	.290	.528	.597
100	.246	.276	.501	.566
110	.235	.263	.477	.539
120	.225	.252	.457	.516
130	.216	.243	.438	.495
140	.208	.234	.422	.477
150	.201	.226	.408	.460
160	.195	.219	.395	.446
170	.189	.213	.383	.432
180	.184	.207	.372	.420
190	.179	.202	.362	.409
200	.175	.197	.353	.398

Impacts on Power

- Measurement error – decreases ES
- Subgroup analyses – estimate smallest subgroup size
- Multiple subgroup analyses – adjust α
- Multiple regression – multiple effects
- Correlated measurements/clustered observations – adjust ES

Power for Multiple Effects

Table 3.3 *Power levels in a multiple regression analysis with five predictors*

	Sample size		
Power to detect. . .	100	200	400
At least one effect	.84	.99	>.99
Any single specified effect	.26	.48	.78
All effects	<.01	.01	.22

Note: Every predictor has a medium correlation ($r = .3$) with the outcome variable. $\alpha = .05$.
Source: Adapted from Maxwell (2004, Table 3). (ELLIS)

Boosting Power

- Larger ES – reasonable to expect?
- Increase sample size – tradeoff with cost
- Reliable measures
- Type of statistical test
  - Parametric > non-parametric
  - 1-tailed > 2-tailed
  - Metric > nominal or ordinal
- Relax alpha

Influences on Effect Size

- Research design – sampling methods
- Variability within participants/clusters
- Time between administration of treatment and collection of data
- ES later study < ES early study – larger effect sizes required for earlier studies
- Regression to the mean

Post-hoc Power Analysis

- Can't separate low power from no effect if NS
- Better to quantify uncertainty with CI
- Can't be used to interpret current study
- Can be used to assess sensitivity of future studies – same ES
- Can be useful for pooling estimates from multiple studies

Power vs Precision

- Related questions:
  - How much power to detect certain ES?
  - How precise should my estimate be?
- ES impacts power, but no direct relation to accuracy/precision
- Decide on study aim: comparison, estimate or both

Power and Precision

- If seeking medium ES, then as bare minimum the desired CI should at least exclude the possibility of values suggesting small and large ES.
- For example, ES = 0.5 with CI = (0.15, 0.85) → small (0.2) and large (0.8) ES are in the possible range. Thus CI is not precise enough to detect ES of interest vs others.

Precision of Estimates - CIs

- Point estimate of parameter $\pm$ margin of error
  - Sampling error and variability in population
  - Based on sampling distribution of parameter (SE)
- Provides plausible region for popn parameter
- α - risk that CI will exclude true value
- $1-\alpha$ – not probability CI contains true value
- Gives more info about effects than p-value

References

- Aberson CL. Applied Power analysis for the Behavioral Sciences. 2010. Routledge/Taylor & Francis Group.
- Ellis PD. The Essential Guide to Effect Sizes. 2010. Cambridge University Press.
- Lachin JM. Biostatistical Methods: The Assessment of Relative Risks. 2011. John Wiley & Sons, Inc.
- Van Belle G. Statistical Rules of Thumb, 2^{nd} ed. 2008. John Wiley & Sons, Inc.