L10: Effect Sizes and Power Analysis

When does “no” mean “no”?

How big an effect?

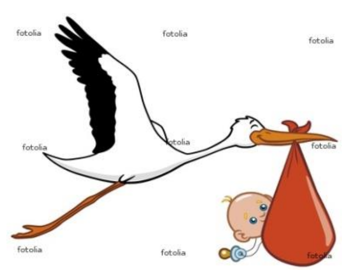
- For a comparison between two groups, effect size (ES) can be defined as the difference between means divided by a (pooled) SD:

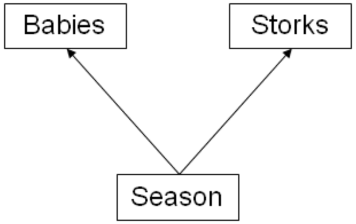
$$\text{Cohen's} \quad d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

(or for Hedge's Δ: s = control-group SD)

  – So, if undergrads have a mean IQ of 100 and postgrads have a mean IQ of 120, both with SD of 25, the ES is (120-100)/25 = 0.8

- For correlation, ES can be simply the *r* value ("proportion of shared variance").
- For linear models, ES can be the r^2 (bivariate) or R^2 (multiple predictors) value





Proportion of Explained Variance

- r^2 and R^2 are each known as a “proportion of explained variance”.
- R^2 can be corrected for number of variables: adjusted R^2

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{err}}}{\mathrm{SS}_{\mathrm{Tt}}} \qquad R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{MS}_{\mathrm{err}}}{\mathrm{MS}_{\mathrm{Tt}}}$$

- These don’t generalise to GLMs, multilevel models, etc...
- Binomial regression: try McFadden’s R^2

$$R_{\mathrm{L}}{}^2 = 1 - \frac{\log(L_{\mathrm{M}}) - k_{\mathrm{M}}}{\log(L_0) - k_0}$$

where k = nr. of parameters
Menard, S. (2000)
Am. Statistician, 54, 17–24

- Nagelkerke’s generalised R^2 (not penalised for k)

$$R^2 = 1 - \left(\frac{L_0}{L_{\mathrm{M}}}\right)^{2/N}$$

Nagelkerke (1991)
Biometrika 78, 691–692

Example

	Maze complexity			Means
	Simple	Medium	Complex	
Control	12±3.1, n=10	14±3.1, n=10	15±3.1, n=10	13.6, n=30
Manipulation	10±3.1, n=10	16±3.1, n=10	17±3.1, n=10	14.3, n=30
Means	11.0, n=20	15.0, n=20	16.0, n=20	14.0, n=60

Analysis of Variance for data

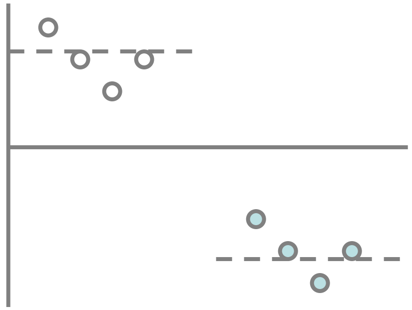
Source	DF	Seq SS	Adj SS	Adj MS	F	P
manip	1	6.667	6.667	[illegible]	0.68	0.413
complex	2	280.000	280.000		14.27	0.000
manip*complex	2	53.333	53.333		2.72	0.075
Error	54	529.823	529.823	[illegible]		
Total	59	869.823				

R^2=0.01 (small ES)

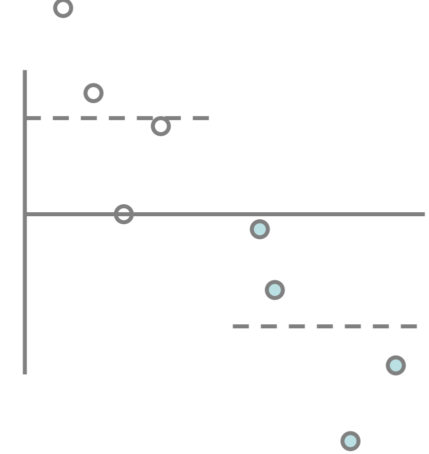
R^2=0.5 (large ES)

R^2=0.1 (medium ES)

Power analysis



Statistical Power



- Statistical noise (error) *blurs* statistical signals. If noise is large relative to an effect, the effect will be obscured.
- Your statistical method should be *sensitive* enough to detect an interesting size of effect.
- *It doesn't need to be sensitive enough to detect an uninteresting (tiny) effect!*

Power measures the probability of detecting an effect of a given size, if one exists: the *sensitivity* of a test.

Sensitivity depends on...

a number of factors, all of which can (potentially) be manipulated to increase the power:

1. **Effect size** – smaller effects are more difficult to find.
   - *Solution: Increase effect size - give bigger doses, leave longer for treatments to have an effect, etc.*

2. **Sampling error** (within-group variability) – the more noisy the data, the more difficult it is to detect an effect.
   - *Solution: Control extraneous variables, use blocking or measure covariates to allow statistical control*

3. **<u>Procedural error</u>** – the more error introduced, the more difficult it is to detect any effect.
   – *<u>Solution</u>: minimise measurement error – use same instruments, appropriate for the task (don't measure swallow's tails with a metre rule)*

4. **<u>Sample size</u>** – bigger samples reduce standard errors of parameters, so make it easier to partition out the signal from the noise.
   – *<u>Solution</u>: increase sample size – and justify it!*

5. **<u>Choice of analysis</u>** – some methods are more powerful than others.
   – *<u>Solution</u>: use optimal analysis controlling for variation due to other factors (i.e. include blocking, covariates etc). If variables are left out, their effects will be lumped in with the error term.*

Effect Size

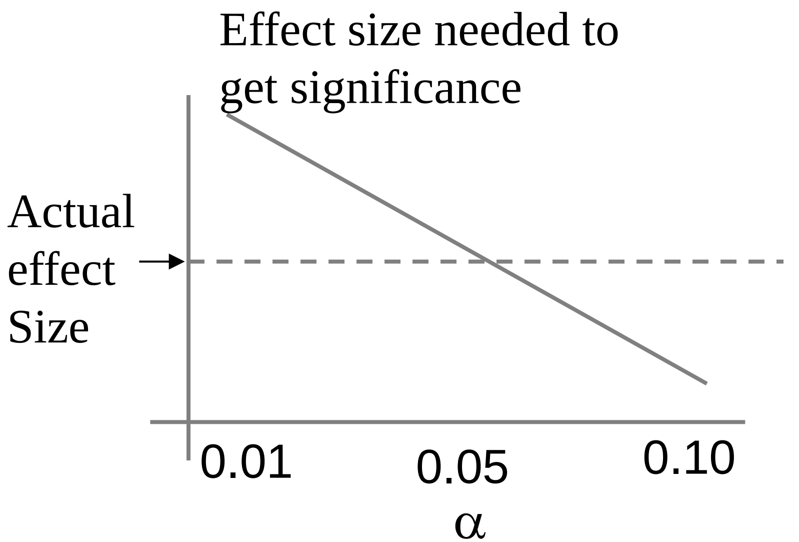
The power literature labels ES's of 0.2 as small, 0.5 as medium and 0.8+ as "large".

Effect size	2-sample: *d* = how many SDs apart	Binomial: Proportions of C and T groups "successful"	Correlation: *r*	Other: r^2 or R^2 = proportion of explained variance
small	0.2	0.45 vs 0.55	0.1 or -0.1	1%
medium	0.5	0.38 vs 0.62	0.3 or -0.3	9%
large	0.8	0.30 vs 0.70	0.5 or -0.5	25%

Alpha α

α = Prob {observed effect is consistent with H_0}. The greater the level of significance required, the more difficult it will be to obtain it.

- The more likely the guilty person is to go free *(civil vs criminal cases; "proof beyond reasonable doubt"…)*
- In other words, as α decreases, so does the power.



Types of error

- **Remember:** α is conventionally set at 5% - but it can be changed...
- Theories are built, and decisions made, on positive results, so we want to avoid concluding something is happening when it might not be. This requires a low α.
- But sometimes deciding that "no" means "no" is important – which requires maximising power.

LAW		Verdict:			
		Guilty		**Not Guilty**	
		Result	Prob	Result	Prob
Person is:	**Guilty**	**Good**	**$1-\beta$**	**Bad** **(OJ)**	β
	Not Guilty	**Bad** Derek Bentley	α	**Good**	**$1-\alpha$**

STATS		Significance:					
		Found			**Not found**		
		Result	Prob	Name	Result	Prob	Name
Effect	**Exists**	**Good**	**$1-\beta$**		**Error**	β	**Type II**
	Doesn't	**Error**	α	**Type I**	**Good**	**$1-\alpha$**	

- Both types of errors potentially bad: false convictions (Type I) and false releases (Type II).
- Both legal and statistical systems think Type I error to be more important (innocent till proven guilty, beyond reasonable doubt)
- Power (1- β) desirable to be at least 0.8 (80%) – i.e. β = 0.2.
- α at 0.05 and β at 0.2 suggests Type I errors are 4x more serious

Power of published studies

- Meta-analyses (some 10000 studies) indicated that:
  - 23% of studies find ES < 0.2 SDs
    - 77% have effects ≥ 0.2 SDs
  - 41% find ES between 0.2 and 0.499 SDs
    - 36% have effects ≥ 0.5 SDs
  - 24% find ES between 0.5 and 0.799 SDs
    - 12% have effects ≥ 0.8 SDs

Power of published studies

- Therefore most reported experiments are finding quite minor effect sizes.
- Often, they can't conclude that such effects are real. If they are, the sample size may be too small to get $P < 0.05$.
- Failing to look at power means unstated presumptions about what effect size we're interested in…
  - biologists' intuition?
  - or waste of resources?
  - **statistical vs. theoretical "significance"**

TABLE 1.1 Reviews of Statistical Power Levels in Various Research Domains

	Average statistical power reported for detecting:		
Research domain	*"Small" effects*	*"Medium" effects*	*"Large" effects*
Evaluation research	.28	.63	.81
Applied psychology	.25	.67	.86
Social psychology	.18	.48	.83
Sociology	.55	.84	.94
Education (a)	.13	.47	.73
Mathematics education	.24	.62	.83
Mass communication	.34	.76	.91
Management research	.31	.77	.91
Marketing research (b)	.24	.69	.87
Communication	.18	.52	.79
Speech pathology	.16	.44	.73
Occupational therapy (c)	.37	65	.93
Gerontology	.37	88	.96
Medicine	.14	39	.61

.27±.11 .63±.15 .84±.10

NOTE: (a) Included only F- and t-tests yielding statistical significance
(b) Experimental studies only
(c) Power in each category only for studies with computed effect sizes in the indicated range

SOURCES (respectively): Lipsey et al., 1985; Chase and Chase, 1976; Cohen, 1962; Spreitzer, 1974 (cited in Chase and Tucker, 1976); Brewer, 1972; Clark, 1974 (cited in Reed and Slaichert, 1981); Chase and Baran, 1976; Mazen, Graf, Kellogg, and Hemmasi, 1987; Sawyer and Ball, 1981; Chase and Tucker, 1975; Kroll and Chase, 1975; Ottenbacher, 1982; Levenson, 1980; Reed and Slaichert, 1981

- "average" study has power <50%... so no conclusion from "n.s." results
- *May as well toss a coin?*

Calculating Power

- Power is a function of ES, α and N. So if you know three of these quantities it should be possible to obtain the fourth.
- If you have an estimate of a theoretically-significant effect size, and specify α=0.05, you can work out the *sample size* needed to give you reasonable power (e.g. 80%).
- Alternatively, if N, ES and α are fixed, you can estimate the *power*.
- Estimate power when designing an experiment
  - Post-hoc power analysis using observed effect sizes is self-fulfilling: non-significant results always come with low power!

TABLE 6.5 Approximate Sample Size per Experimental Group Needed to Attain Various Criterion Levels of Power for a Range of Effect Sizes at Alpha = .05

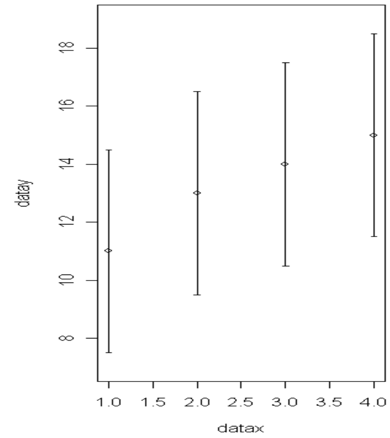
	Power Criterion		
Effect size	*.80*	*.90*	*.95*
.10	1570	2100	2600
.20	395	525	650
.30	175	235	290
.40	100	130	165
.50	65	85	105
.60	45	60	75
.70	35	45	55
.80	25	35	45
.90	20	30	35
1.00	20	25	30

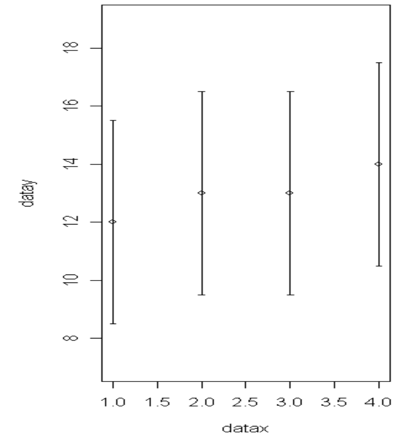
From Cohen (1988)

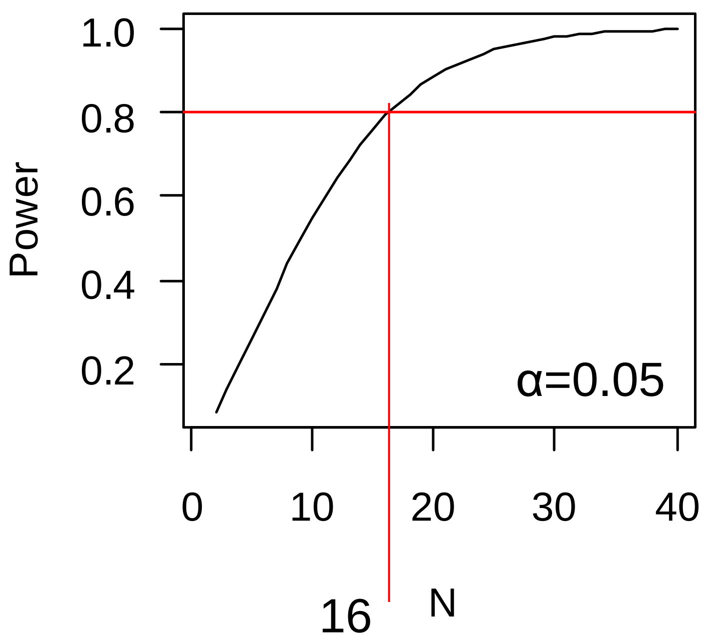
Power for ANOVA

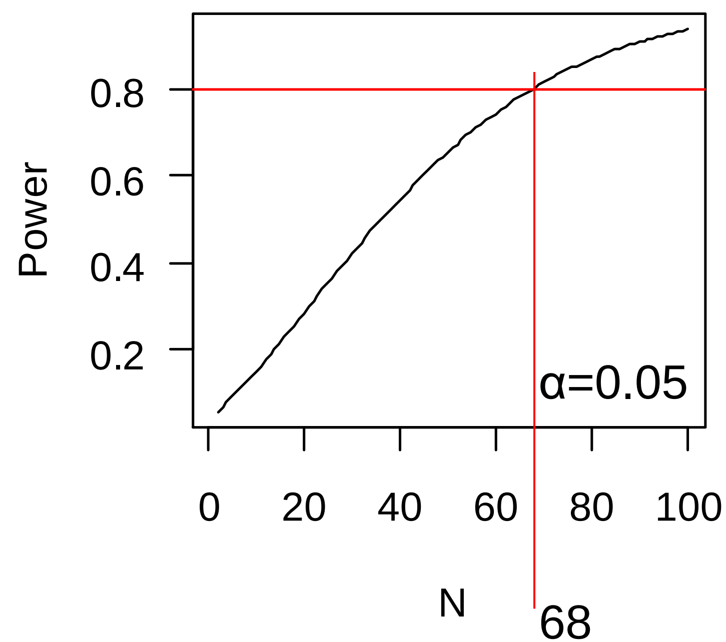
Suppose we know how much within-group variation to expect and how much between-group variation we want to detect:

```
gp.x1 <- c(11,13,14,15)
gp.x2 <- c(12,13,13,14)
error.sd <- 3.5
```









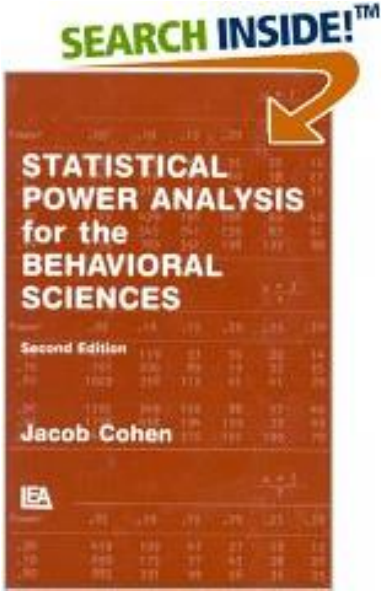
How to do it (*before* the experiment!)

☺

```
power.anova.test(groups=4, n=5, sig.level=0.05,
   between.var=1, within.var=3)
# Power = 0.3535594
power.anova.test(groups=4, sig.level=0.05,
   between.var=1, within.var=3, power=.80)
# n = 11.92613
```

or:

- By bootstrapping
- By the book…
- Using web sites:
  - http://www.dssresearch.com/toolkit/spcalc/power.asp
  - http://www.stat.uiowa.edu/~rlenth/Power/index.html
  - http://statpages.org/#Power



In library “stats” (base distribution):

`power.anova.test()`

Power Calculations for Balanced One-Way Analysis of Variance Tests

`power.prop.test()`

Power Calculations for Two-Sample Test for Proportions

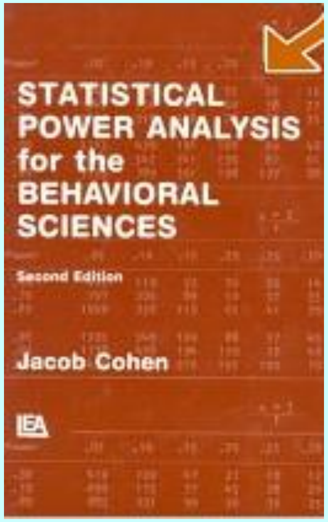
`power.t.test()`

Power Calculations for One and Two Sample t Tests

Lots more in library “pwr”…

References

- MW **Lipsey** (1990) Design Sensitivity: statistical power for experimental research. SAGE Publications, Newbury Park, Calif.
- Jacob **Cohen** (1969; 2nd edn 1988) Statistical power analysis for the behavioural sciences. Academic Press.



Summary of modelling

More complex structures

Gen. least squares
gls (nlme)

Gen. Additive
gam (mgcv)

Multilevel
lme (nlme)

Multilevel GLMs
lmer (lme4)

Permutations
sample

“Non-linear” LMs
(least squares) **nls**

quasi-GLMs
glm

Linear models
(least squares) **lm**

GLMs
glm

Non-Normal errors