Power, Effect Sizes, Confidence Intervals, & Scientific Integrity

Lecture 10

Survey Research & Design in Psychology

James Neill, 2012

Overview

1. Significance testing
2. Inferential decision making
3. Power
4. Effect size
5. Confidence intervals
6. Publication bias
7. Scientific integrity

2

Readings

1. Ch 6.9 Effect sizes and Ch8 Power (Howell Statistical Methods). Note that these concepts rely upon:
   - Ch3 The Normal Distribution
   - Ch4 Sampling Distributions and Hypothesis Testing
   - Ch7 Hypothesis Tests Applied to Means
2. Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

3

Significance Testing

4

Significance Testing: Overview

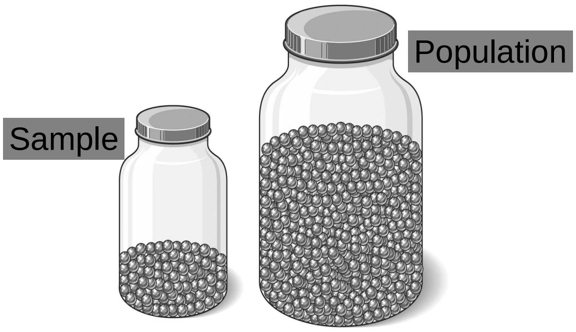
- Logic
- History
- Criticisms
- Decisions
- Inferential decision making table
  - –Correct decisions
  - –Errors (Type I & II errors)

5

Logic of significance testing

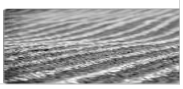
How many heads in a row would I need to throw before you'd protest that something “wasn't right”?

Logic of significance testing



History of significance testing

- Developed by Ronald Fisher (1920's-1930's)
- To help determine what agricultural methods (IVs) yielded greater output (plant growth) (DVs).
- Method used to test whether the variation in produce per acre for agriculture crop was due to chance or not
  .



History of significance testing

- Agricultural research designs couldn't be fully experimental, therefore it was needed to determine whether variations in the DV were due to chance or the IV(s).

Logic of significance testing (ST)

- Null hypothesis (H_0) reflects expected effect in the population (or no effect)
- Obtain *p*-value from sample data to determine the likelihood of H_0 being true
- Researcher tolerates some false positives (critical α) to make a decision about H_0

History of significance testing

- ST spread to other fields, including social sciences
- Spread aided by the development of computers and training.
- In the latter decades of the 20th century, widespread use of ST attracted critique for its over-use and mis-use.

Criticisms of significance testing

- Critiqued as early as 1930
- Cohen's (1980's-1990's) critique
- During the late 1990's a critical mass of awareness developed
- During the 2000's there has been change in publication criteria and (more slowly) teaching about over-reliance on ST and alternative and adjunct techniques.

Criticisms of significance testing

- The null hypothesis is rarely true
- ST only provides a binary decision (yes or no) and the direction of the effect
- But mostly we are interested in the size of the effect – i.e., *how much* of an effect?
- Statistical vs. practical significance
- Sig. is a function of ES, *N* and α

13

Statistical significance

- **Statistical significance** means that the observed mean differences are not likely to be due to sampling error
  - Can get statistical significance, even with very small population differences, if *N* and ES are large enough

14

Practical significance

- **Practical significance** is about whether the difference is large enough to be of value in a practical sense
  - Is it an effect worth being concerned about – are these noticeable or worthwhile effects?
  - e.g., a 5% increase in well-being probably has practical value

15

Criticisms of significance testing

- Whether a result is significant or not is a function of:
  - Effect size (ES)
  - *N*
  - Critical alpha (α) level
- Sig. can be manipulated by tweaking any of the three
  - as each of them increase, so does the likelihood of a significant result

Criticisms of significance testing

ears. For example, Frank Yates (1951), a contemporary of Fisher, observe
ıat the use of the null hypothesis significance test

> has caused scientific research workers to pay undue attention to the results of the tests of significance that they perform on their data and too little attention to the estimates of the magnitude of the effects they are investigating. . . . The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers often have regarded the execution of a test of significance on an experiment as the ultimate objective. (pp. 32-33)

Criticisms of significance testing

A more strongly worded criticism of null hypothesis significance testi
ɩas written by Paul Meehl (1978):

> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (p. 817)

GILL, CALIFORNIA POLYTECHNIC STATE UNIVERSITY

The current method of hypothesis testing in the social sciences is under intense criticism, yet most political scientists are unaware of the important issues being raised. Criticisms focus on the construction and interpretation of a procedure that has dominated the reporting of empirical results for over fifty years. There is evidence that null hypothesis significance testing as practiced in political science is deeply flawed and widely misunderstood. This is important since most empirical work argues the value of findings through the use of the null hypothesis significance test. In this article I review the history of the null hypothesis significance testing paradigm in the social sciences and discuss major problems, some of which are logical inconsistencies while others are more interpretive in nature. I suggest alternative techniques to convey effectively the importance of data-analytic findings. These recommendations are illustrated with examples using empirical political science publications.

APA Style Guide recommendations about effect sizes, CIs and power

- APA 5th edition (2001) recommended reporting of ESs, power, etc.
- APA 6th edition (2009) further strengthened the requirements to use NHST as a starting point and to also include ESs, CIs and power.

NHST and alternatives

"Historically, researchers in psychology have relied heavily on null hypothesis significance testing (NHST) as a starting point for many (but not all) of its analytic approaches. APA stresses that NHST is but a starting point and that additional reporting such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results... complete reporting of all tested hypotheses and estimates of appropriate ESs and CIs are the minimum expectations for all APA journals."
(APA Style Manual (6th ed., 2009, p. 33)

Recommendations

- Learn to use traditional Fisherian logic methodology (inferential testing)
- Learn to use alternative and complementary techniques (ESs and CIs)
- Look for practical significance
- Recognise merits and shortcomings of each approach

22

Significance testing: Summary

- **Logic**:
  - Examine sample data to determine *p* that it represents a population with no effect or some effect. It's a “bet”.
- **History**:
  - Developed by Fisher for agricultural experiments in early 20th C
  - During the 1980's and 1990's, ST was increasingly criticised for over-use and mis-application.

23

Significance testing: Summary

- **Criticisms**:
  - Binary, Doesn't directly indicate ES, Dependent on *N*, ES, and alpha, Need practical significance
- **Recommendations**:
  - Use complementary or alternative techniques, including power, effect size (ES) and CIs
  - Wherever you report a *p*-level, also report an ES

24

Inferential Decision Making

Hypotheses in inferential testing

Null Hypothesis (H_0):
No differences or effect

Alternative Hypothesis (H_1):
Differences or effect

Inferential decisions

When we test a hypothesis we draw a conclusion based on the sample data; either we

Do not reject H_0

p is not sig. (i.e. not below the critical α)

Reject H_0

p is sig. (i.e., below the critical α)

Inferential Decisions: Correct Decisions

We are hoping to make a correct inference from the sample; either:

- Do not reject H_0:
  - Correctly retain H_0 when there is no real difference/effect in the population
- Reject H_0 (Power):
  - Correctly reject H_0 when there is a real difference/effect in the population

28

Inferential Decisions: Type I & II Errors

However, when we fail to reject or reject H_0, we risk making errors:

- Type I error:
  - Incorrectly reject H_0 (i.e., there is no difference/effect in the population)
- Type II error:
  - Incorrectly fail to reject H_0 (i.e., there is a difference/effect in the population)

29

Inferential Decision Making Table

		Reality	
		H_0 False	H_0 True
Test	Reject H_0	Correct rejection H_0 = Power = 1 - β	Type I error = α
	Accept H_0	Type II error	Correct acceptance of H_0

Inferential decision making: Summary

- Correct acceptance of H_0
- Power (correct rejection of H_0) = 1- β
- Type I error (false rejection of H_0) = α
- Type II error (false acceptance of H_0) = β
- Traditional emphasis has been too much on Type I errors and not enough on Type II error – balance needed.

Statistical Power

Statistical power

Statistical power is the probability of

- correctly rejecting H_0
- rejecting a false H_0
- a sig. result when there is a real difference in the population

Statistical power

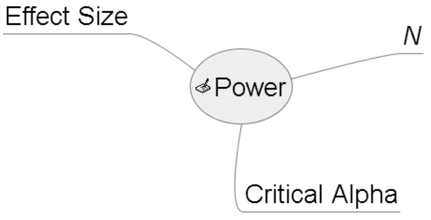
		Reality	
		H_0 False	H_0 True
Test	Reject H_0	POWER	Type I error = α
	Accept H_0	Type II error	Correct acceptance of H_0

Statistical power

- Desirable power > .80
- Typical power (in the social sciences) ~ .60
- Power depends on the:
  - Critical alpha (α)
  - Sample size (N)
  - Effect size (Δ)

Statistical Power

An inferential test is more ‘powerful’ (i.e. more likely to get a significant result) when any of these 3 increase:

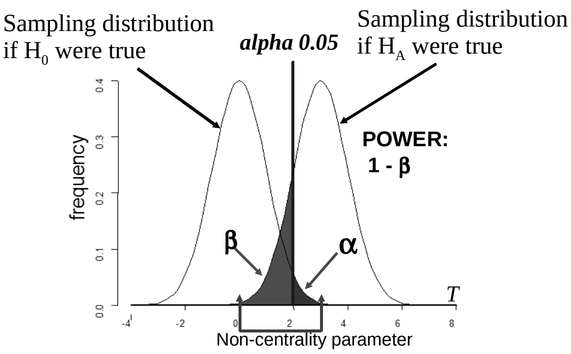


Power analysis

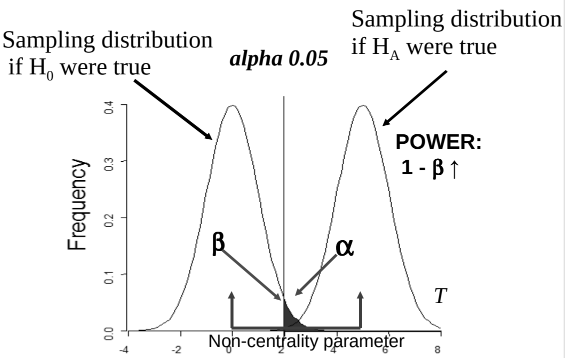
- If possible, calculate expected power before conducting a study, based on:
  - Estimated *N*,
  - Critical α,
  - Expected or minimum ES (e.g., from related research)
- Report actual power in the results.

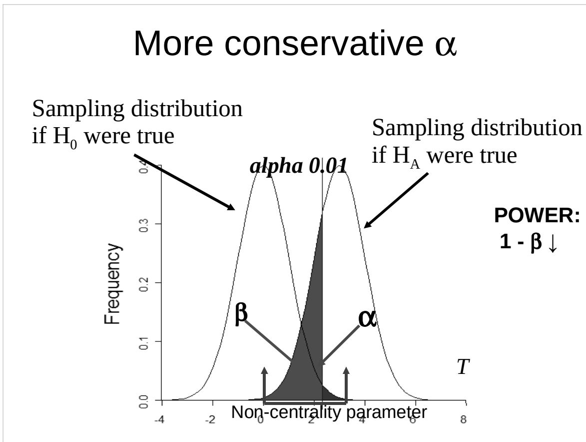
37

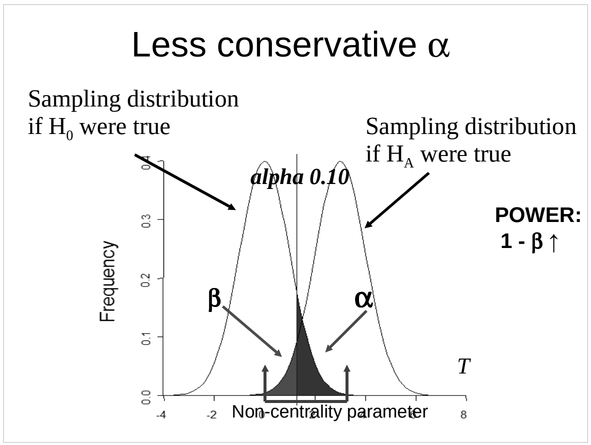
Typical scenario

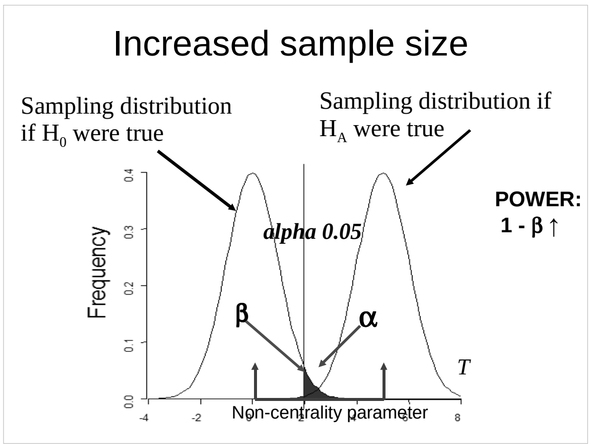


Increased effect size




More conservative α
Sampling distribution
if H0 were true
Sampling distribution
if HA were true
alpha 0.01
POWER:
1 - β ↓
Frequency
β
α
T
Non-centrality parameter


Less conservative α
Sampling distribution
if H0 were true
Sampling distribution
if HA were true
alpha 0.10
POWER:
1 - β ↑
Frequency
β
α
T
Non-centrality parameter


Increased sample size
Sampling distribution
if H0 were true
Sampling distribution if
HA were true
alpha 0.05
POWER:
1 - β ↑
Frequency
β
α
T
Non-centrality parameter

Statistical Power: Summary

- Power = likelihood of detecting an effect as statistically significant
- Power can be increased by:
  - ↑ N
  - ↑ critical α
  - ↑ ES
- Power over .8 “desirable”
- Power of ~.6 is more typical
- Can be calculated prospectively and retrospectively

43

Effect Sizes

What is an effect size?

A measure of the **strength** of a relationship or effect.

Where *p* is reported, also present an effect size.

45

Why use an effect size?

- An inferential test may be statistically significant (i.e., unlikely to have occurred by chance), but this doesn't necessarily indicate how large the effect is.
- There may be non-significant, notable effects esp. in low powered tests.
- Unlike significance, effect sizes are not influenced by *N*.

46

Commonly used effect sizes

Mean differences
- Cohen's *d*
- η^2, η_p^2

Correlational
- r, r^2
- R, R^2

47

Standardised mean difference

The difference between two means in standard deviation units.

-ve = negative difference/effect

0 = no difference/effect

+ve = positive difference/effect

48

Standardised mean difference

- A standardised measure of the difference between two *M*s
  - $d = M_2 - M_1 / \sigma$
  - $d = M_2 - M_1$ / pooled *SD*
- Often called Cohen's d, sometimes called Hedges' g
- Not readily available in SPSS; use a separate calculator e.g., Cohensd.xls

49

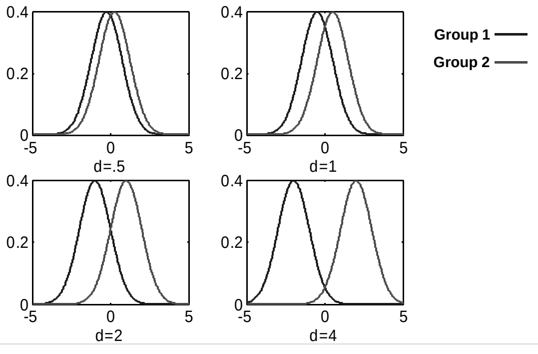
Standardised mean difference

$$ES = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_{pooled}} \qquad s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

- Represents a standardised group contrast on an *inherently continuous* measure
- Uses the pooled standard deviation (some situations use control group standard deviation)

50

Example effect sizes



Rules of thumb for interpreting standardised mean differences

- Cohen (1977): .2 = small
  .5 = moderate
  .8 = large

- Wolf (1986): .25 = educationally significant
  .50 = practically significant (therapeutic)

Standardised Mean ESs are proportional, e.g., .40 is twice as much change as .20

52

Interpreting effect size

- No agreed standards for how to interpret an ES
- Interpretation is ultimately subjective
- Best approach is to compare with other studies

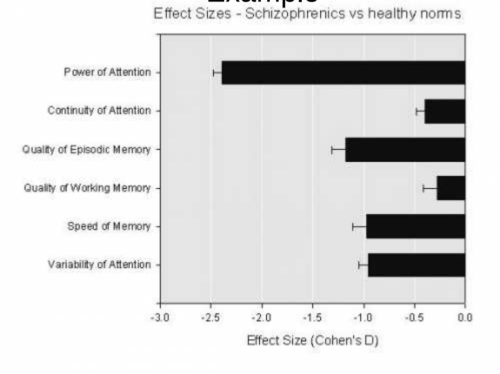
53

The meaning of an effect size depends on context

- A small ES can be impressive if, e.g., a variable is:
  - difficult to change (e.g. a personality construct) and/or
  - very valuable (e.g. an increase in life expectancy).
- A large ES doesn't necessarily mean that there is any practical value e.g., if
  - it isn't related to the aims of the investigation (e.g. religious orientation).

54

Graphing standardised mean effect size - Example



Standardised mean effect size table - Example

Table 3 Means, SE,95%CIs, P levels and effect sizes (Cohen's d) for outcome variables across no leaflet (n = 375) and leaflet (n = 394) groups

	Group	Mean	Std. error	95% Confidence interval Lower	Upper	*P* level	Effect size
Knowledge about oral cancer	no leaflet	26.11	0.19	25.73	26.48	0.001	1.29
	leaflet	30.87	0.18	30.51	31.24		
Attitudes about negative consequences	no leaflet	3.97	0.08	3.81	4.13	0.038	0.15
	leaflet	3.73	0.08	3.57	3.88		
Attitudes about lack of control	no leaflet	7.91	0.09	7.72	8.10	0.078	0.13
	leaflet	7.67	0.09	7.49	7.86		
Normative beliefs	no leaflet	13.34	0.25	12.84	13.83	0.019	0.17
	leaflet	12.51	0.24	12.03	12.99		
Anxiety about screening procedure	no leaflet	5.58	0.13	5.31	5.85	0.069	0.13
	leaflet	5.23	0.13	4.97	5.50		
Intention to accept screen	no leaflet	11.61	0.12	11.36	11.86	0.003	0.22
	leaflet	12.15	0.12	11.91	12.39		

Standardised mean effect size – Exercise

- 20 athletes rate their personal playing ability, ***M* = 3.4 (*SD* = .6)**
(on a scale of 1 to 5)
- After an intensive training program, the players rate their personal playing ability again, ***M* = 3.8 (*SD* = .6)**
- What is the ES? How good was the intervention?

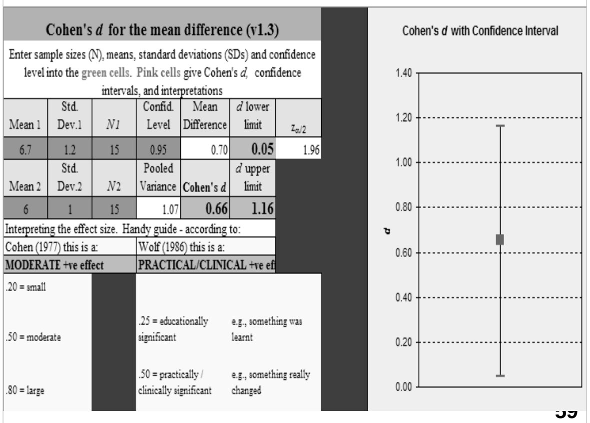
Standardised mean effect size - Answer

Standardised mean effect size

- = $(M_2 - M_1) / SD_{pooled}$
- = (3.8 - 3.4) / .6
- = .4 / .6
- = .67
- = a moderate-large change over time

For simplicity, this example uses the same *SD* for both occasions.

58



59

Effect sizes – Answer
Using spreadsheet calculator

Mean 1	Std. Dev.1	*N1*	Confid. Level	Mean Difference	*d* lower limit
3.8	0.6	20	0.95	0.40	0.40
Mean 2	Std. Dev.2	*N2*	Pooled Variance	**Cohen's *d***	*d* upper limit
3.4	0.6	20	0.60	**0.67**	0.93

Using effect size calculator (Cohensd.xls)

Effect sizes: Summary

- ES indicates amount of difference or strength of relationship - underutilised
- Inferential tests should be accompanied by ESs and CIs
- Common ESs include Cohen's *d*, *r*
- *d*: .2 = small, .5 = moderate, .8 = large
- Cohen's *d* - not in SPSS – use a spreadsheet calculator

61

Power & effect sizes in psychology

Ward (2002) examined articles in 3 psych. journals to assess the current status of statistical power and effect size measures.

- Journal of Personality and Social Psychology
- Journal of Consulting and Clinical Psychology
- Journal of Abnormal Psychology

62

Power & effect sizes in psychology

- 7% of studies estimate or discuss statistical power.
- 30% calculate ES measures.
- A medium ES was discovered as the average ES across studies
- Current research designs typically do not have sufficient power to detect such an ES.

63

Confidence Intervals

Confidence intervals

- Very useful, underutilised
- Gives 'range of certainty' or 'area of confidence'
  e.g., true *M* is 95% likely to lie between -1.96 SD and +1.96 of the sample *M*
- Based on the *M*, *SD*, *N*, and critical α, o calculate:
  - Lower-limit
  - Upper-limit

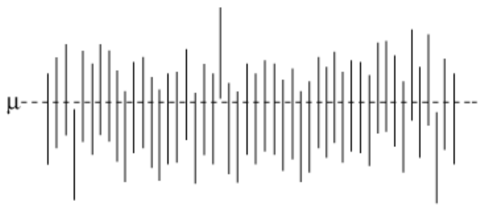
Confidence intervals

- Confidence intervals can be reported for:
  - *M*s
  - Mean differences ($M_2 - M_1$)
  - ESs
- CIs can be examined statistically and graphically (e.g., error-bar graphs)

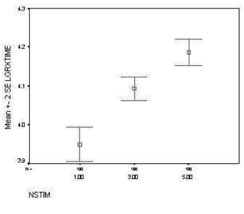
CIs & error bar graphs

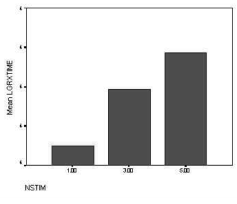
- CIs around means can be presented as error bar graphs
- More informative alternatives to bar graphs or line graphs
- For representing the central tendency and distribution of continuous data for different groups

Confidence intervals – error bars



CIs & error bar graphs





Confidence intervals: Review question 1

Question

If I have a sample M = 5, with 95% CI of 2.5 to 7.5, what would I conclude?

A. Accept H_0 that the M is equal to 0.

B. Reject H_0 that the M is equal to 0.

Confidence intervals: Review question 2

Question

If I have a sample M = 5, with 95% CI of -.5 to 11.5, what would I conclude?

A. Accept H_0 that the M is equal to 0.

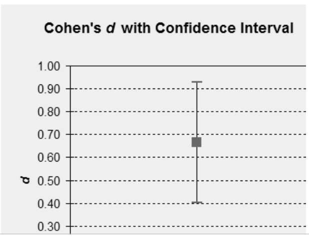
B. Reject H_0 that the M is equal to 0.

Effect size confidence interval

- In addition to getting CIs for Ms, we can obtain and should report CIs for M differences and for ESs.

d = .67

d lower limit
0.40
d upper limit
0.93



Confidence interval of the mean difference

Independent Samples Test

t for ances	t-test for Equality of Means						
						95% Confidence Interval of the Difference	
Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
.897	.764	489	.445	5.401E-02	7.067E-02	-8.48E-02	.1929
	.778	355.220	.437	5.401E-02	6.944E-02	-8.26E-02	.1906

•Lower 95% CI = -.08
•Upper 95% CI = .19

Publication Bias

Two counter-acting biases

- **Low Power**:
  → under-estimation of real effects
- **Publication Bias or File-drawer effect**:
  → over-estimation of real effects

Publication bias

- When publication of results depends on their nature and direction.
- Studies that show sig. effects are more likely to be published.
- Type I publication errors are underestimated to the extent that they are: “frightening, even calling into question the scientific basis for much published literature.” (Greenwald, 1975, p. 15)

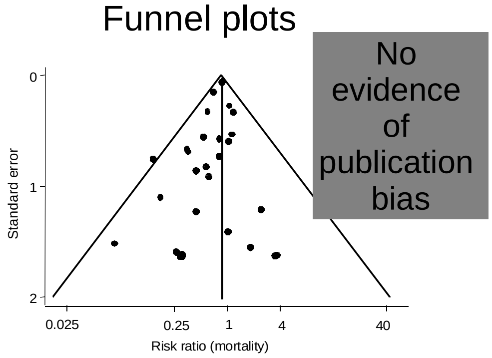
76

Funnel plots

- A scatterplot of treatment effect against study size.
- Precision in estimating the true treatment effect ↑s as *N* ↑s.
- Small studies scatter more widely at the bottom of the graph.
- In the absence of bias the plot should resemble a *symmetrical* inverted funnel.

77

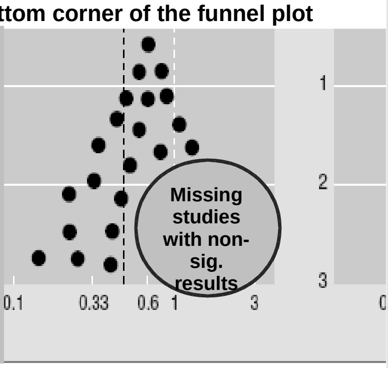
Funnel plots



Publication Bias:

Asymmetrical appearance of the funnel plot with a gap in a bottom corner of the funnel plot

As studies become less precise, results should be more variable, scattered to both sides of the more precise larger studies … unless there is publication bias.

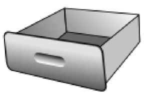


Publication bias

- If there is publication bias this will cause meta-analysis to overestimate effects.
- The more pronounced the funnel plot asymmetry, the more likely it is that the amount of bias will be substantial.

File-drawer effect

- Tendency for non-sig. results to be ‘filed away’ (hidden) and not published.
- # of null studies which would have to ‘filed away’ in order for a body of significant published effects to be considered doubtful.



Countering the bias

Journal of Articles in Support of the Null Hypothesis

INDEX ABOUT MANUSCRIPT SUBMISSION REVIEWER SUBMISSION EDITORIAL BOARD LINKS CONTACT

Welcome to the *Journal of Articles in Support of the Null Hypothesis*. In the past other journals and reviewers have exhibited a bias against articles that did not reject the null hypothesis. We seek to change that by offering an outlet for experiments that do not reach the traditional significance levels ($p < .05$). Thus, reducing the file drawer problem, and reducing the bias in psychological literature. Without such a resource researchers could be wasting their time examining empirical questions that have already been examined. We collect these articles and provide them to the scientific community free of cost.

Academic Integrity

Academic Integrity: Students

(Marsden, Carroll, & Neill, 2005)

- N = 954 students enrolled in 12 faculties of 4 Australian universities
- Self-reported:
  - –Cheating (41%),
  - –Plagiarism (81%)
  - –Falsification (25%).

Summary

- Counteracting biases in scientific publishing; tendency:
  - –towards low-power studies which underestimate effects
  - –to publish sig. effects over non-sig. Effects
- Violations of academic integrity are prevalent, from students through researchers

Recommendations

- Decide on H_0 and H_1 (1 or 2 tailed)
- Calculate power beforehand & adjust the design to detect a min. ES
- Report power, sig., ES, CIs
- Compare results with meta-analyses and/or meaningful benchmarks
- Take a balanced, critical approach, striving for objectivity and scientific integrity

Further resources

- Statistical significance (Wikiversity)
- http://en.wikiversity.org/wiki/Statistical_significance
- Effect sizes (Wikiversity): http://en.wikiversity.org/wiki/Effect_size
- Statistical power (Wikiversity): http://en.wikiversity.org/wiki/Statistical_power
- Confidence interval (Wikiversity)
- http://en.wikiversity.org/wiki/Confidence_interval
- Academic integrity (Wikiversity)
- http://en.wikiversity.org/wiki/Academic_integrity
- Publication bias
- http://en.wikiversity.org/wiki/Publication_bias

References

1. Marsden, H., Carroll, M., & Neill, J. T. (2005). Who cheats at university? A self-report study of dishonest academic behaviours in a sample of Australian university students. *Australian Journal of Psychology*, *57*, 1-10. http://wilderdom.com/abstracts/MarsdenCarrollNeill2005
2. Ward, R. M. (2002). *Highly significant findings in psychology: A power and effect size survey*. http://digitalcommons.uri.edu/dissertations/AAI3053127/
3. Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Open Office Impress

- This presentation was made using Open Office Impress.
- Free and open source software.
- http://www.openoffice.org/product/impress.html