Power, Effect Sizes, Confidence Intervals, & Academic Integrity

Lecture 9

Survey Research & Design in Psychology

James Neill, 2018

Overview

1 Significance testing
2 Inferential decision making
3 Statistical power
4 Effect size
5 Confidence intervals
6 Publication bias
7 Academic integrity
8 Statistical method guidelines

2

Readings

1. **Howitt & Cramer (2014)**
   - Ch 35: The size of effects in statistical analysis: Do my findings matter?
   - Ch 36: Meta-analysis: Combining and exploring statistical findings from previous research
   - Ch 38: Confidence intervals
   - Ch 40: Statistical power
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

Significance testing: Overview

- Logic
- History
- Criticisms
- Decisions
- Practical significance

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

Based on the statistical properties of sample data, we can extrapolate (guesstimate) about the probability of the observed differences or relationships occurring in the target population. In so doing, we are assuming that the sample data is representative and that the data meets the assumptions associated with the inferential test.

Logic of significance testing

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

8

History of significance testing

- Developed by Ronald Fisher (1920s-1930s)
- To determine which agricultural methods yielded greater output
- Were variations in output between two plots attributable to chance or not?

History of significance testing

- Agricultural research designs couldn't be fully experimental because natural variations such as weather and soil quality couldn't be fully controlled.
- Therefore, it was needed to determine whether variations in the DV were due to the IV(s) or to chance.

10

History of significance testing

- ST spread to other fields, including social sciences.
- Spread was aided by the development of computers and training.
- ST became widely used during the 2nd half of the 20th century.
- So widely used that, in the latter 20th century, ST attracted critique for its over-use and misuse.

11

Criticisms of significance testing

- Critiqued as early as 1930.
- Cohen's (1980s-1990s) critique helped a critical mass of awareness to develop.
- Led to changes in publication guidelines and teaching about over-reliance on ST and alternative and adjunct techniques.

12

Criticisms of significance testing

1. The null hypothesis is rarely true.
2. ST provides:
   - a binary decision (yes or no) and
   - direction of the effect

   But mostly we are interested in the size of the effect – i.e., *how much* of an effect?
3. Statistical vs. practical significance
4. Sig. is a function of ES, *N*, and α

13

Criticisms of significance testing

- **Statistical significance** simply means that the observed effect (relationship or differences) are unlikely to be due to sampling error
- Statistical significance can be evident for very small (trivial) effects if *N* and/or critical alpha are large enough

14

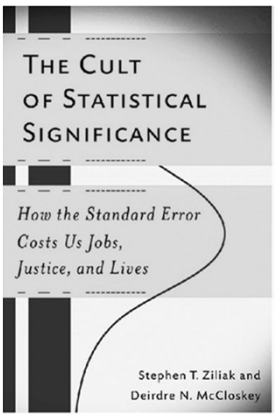
Practical significance

- **Practical significance** is about whether the difference is large enough to be of value in a real world sense:
  - Is an effect worth being concerned about?
  - Is the effect noticeable or worthwhile?
  - e.g., a 5% increase in well-being probably starts to have practical value

15

Criticisms of significance testing

Ziliak, S. T. & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error cost us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.



Criticisms of significance testing

ears. For example, Frank Yates (1951), a contemporary of Fisher, observe
iat 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)

Kirk, R. E. (2001). Promoting good statistical practices: Some Suggestions. *Educational and Psychological Measurement*, *61*, 213-218. doi: 10.1177/00131640121971185

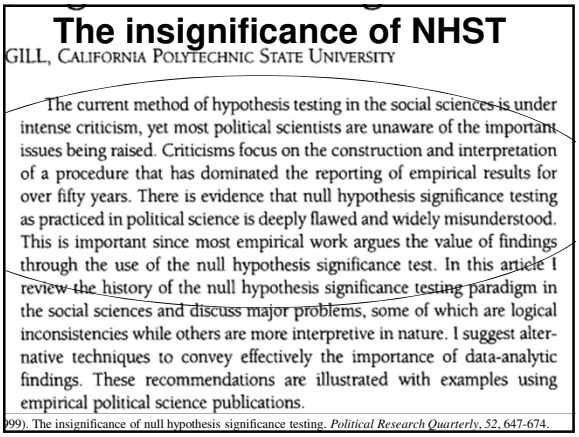
Criticisms of significance testing

A more strongly worded criticism of null hypothesis significance testi
as 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)

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.

The insignificance of NHST



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.

99). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*, 647-674.

APA publication manual 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.

20

Significance testing 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*." [my italics] (APA Publication Manual (6th ed., 2009, p. 33))

21

American Statistical Association Statement on Significance Testing and *p*-Values

(Wasserstein & Lazar, 2016)

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

22

Significance testing: Recommendations

- Use traditional NHST (Fisherian logic / inferential testing)
- Also use complementary techniques (ESs and CIs)
- Emphasise practical significance
- Recognise merits and shortcomings of each approach

23

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" - At what point do you reject H_0?
- **History**:
  - Developed by Fisher for agricultural experiments in early 20th century
  - During the 1980s and 1990s, ST was increasingly criticised for over-use and mis-application.

24

Significance testing: Summary

- **Criticisms**:
  - Binary
  - Depends on *N*, ES, and critical alpha
  - Need practical significance
- **Recommendations**:
  - Wherever you report a significance test (*p*-level), also report an ES
  - Also consider reporting power and CIs

25

Inferential decision making

Hypotheses in inferential testing

Null Hypothesis (H_0):
No differences / No relationship

Alternative Hypothesis (H_1):
Differences / Relationship

27

Inferential decisions

In inferential testing, a conclusion about a target population is made based on sample data. Either:

Do not reject H_0
p is not significant
(i.e., not below the critical α)

Reject H_0
p is significant
(i.e., below the critical α)

28

Inferential decisions: Correct

We hope to make a correct inference based on the sample data; i.e., either:

Do not reject H_0:
Correctly retain H_0 (i.e., when there is no real difference/effect in the population)

Reject H_0 (Power):
Correctly reject H_0 (i.e., when there is a real difference/effect in the population)

29

Inferential decisions: Type I and Type II errors

However, we risk making these 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)

30

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
- Correct rejection of H_0 (Power) = 1- β
- False rejection of H_0 (Type I error) = α
- False acceptance of H_0 (Type II error) = β
- Traditionally, emphasis has been:
  - too much on limiting Type I errors and
  - not enough on limiting Type II error
  - balance is needed

32

Statistical power

Statistical power

Statistical power is the:

- probability of correctly rejecting a false H_0
  (i.e. getting a sig. result when there is a real difference in the population)

Image source: https://commons.wikimedia.org/wiki/File:Emoji_u1f4aa.svg

34

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 ~ .60
  (in the social sciences)
- Power becomes higher when any of these increase:
  - –Sample size (*N*)
  - –Critical alpha (α)
  - –Effect size (Δ)

36

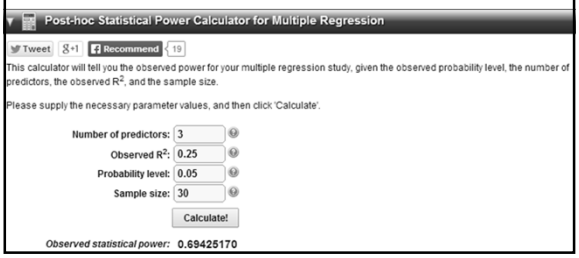
Power analysis

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

37

Power analysis for MLR

- Free, online post-hoc power calculator for MLR
- http://www.danielsoper.com/statcalc3/calc.aspx?id=9



Summary: Statistical power

1. Power = probability of detecting a real effect as statistically significant
2. Increase by:
   - ↑ N
   - ↑ critical α
   - ↑ ES
- Power
  - > .8 "desirable"
  - ~ .6 is more typical
- Can be calculated prospectively and retrospectively

39

Effect sizes

What is an effect size?

- A measure of the **strength** (or size) of a relationship or effect.
- Where *p* is reported, also present an effect size.
- "reporting and interpreting effect sizes in the context of previously reported effects is essential to good research"
  (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599)

41

Why use an effect size?

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

42

Commonly used effect sizes

Correlational

- r, r^2, sr^2
- R, R^2

Mean differences

- Standardised mean difference e.g., Cohen's *d*
- Eta squared (η^2), Partial eta squared (η_p^2)

43

Standardised mean difference

The difference between two means in standard deviation units:

- -ve = negative difference
- 0 = no difference
- +ve = positive difference

44

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*
- e.g., Cohen's *d*, Hedges' *g*
- Not readily available in SPSS; use a separate calculator e.g., https://www.danielsoper.com/statcalc/calculator.aspx?id=48

45

Example effect sizes

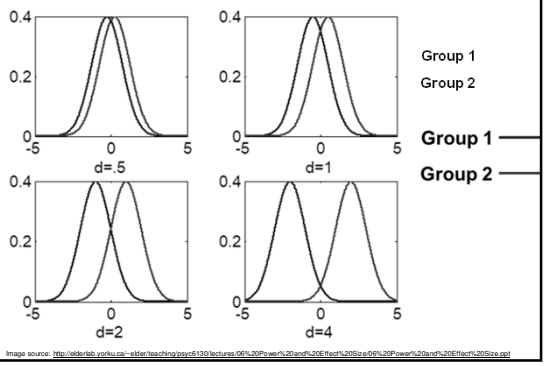


Image source: http://elderlab.yorku.ca/~elder/teaching/psyc6130/lectures.06%20Power%20and%20Effect%20Size/06%20Power%20and%20Effect%20Size.ppt

Interpreting effect size

- No agreed standards
- Ultimately subjective
- Best approach is to compare with other similar studies

47

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. 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)

48

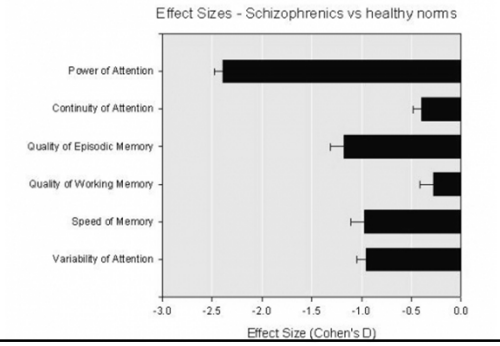
Rules of thumb for interpreting standardised mean differences

- Cohen (1977): .2 = small
  .5 = moderate
  .8 = large
- Wolf (1986): .25 = educationally significant
  .50 = practically significant
  (therapeutic)

Standardised Mean ESs are proportional

49

Standardised mean effect size - Graphing



Standardised mean effect size - Table

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	Plevel	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		

Power and effect sizes in psychology

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

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

52

Power and effect sizes in psychology

- 7% of studies estimated or discuss statistical power.
- 30% provided ESs.
- Average ES was medium
- Current research designs typically do not have sufficient power to detect medium ESs.

53

Summary: Effect size

1. ES = Standardised difference or strength of relationship
2. Inferential tests should be accompanied by ESs and CIs
3. Common bivariate ESs include:
   1. Cohen's *d*
   2. Correlation *r*

- Cohen's *d* - not in SPSS - use an effect size calculator

54

Confidence intervals

Confidence intervals

- Very useful, underutilised
- Gives "range of certainty" or "area of confidence"
  e.g., a population *M* is 95% likely to be between -1.96 and +1.96 *SD* of the sample *M*
- Expressed as a:
  - Lower-limit
  - Upper-limit

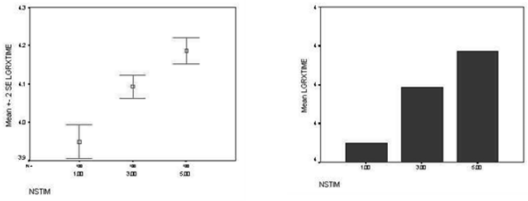
56

Confidence intervals

- CIs can be reported for:
  - *B* (unstandardised regression coefficient) in MLR
  - *M*s
  - ESs (e.g., *r*, *R*, *d*)
- CIs can be examined statistically and graphically (e.g., error-bar graphs)

57

Confidence intervals and error bar graphs



- CIs can be presented as error bar graphs
- Shows the mean and upper and lower CIs
- More informative alternative to bar graphs or line graphs

Confidence intervals in MLR

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	8.826	2.212		3.990	.000	4.454	13.199
	sex Sex of Student	.616	.767	.068	.803	.423	-.900	2.133
	sel Socio-Educ Level	.116	.372	.023	.313	.755	-.619	.852
	enghomwk Freq of English Homework	.696	.393	.149	1.770	.079	-.081	1.473
	attent Attentiveness in year 8	1.531	.316	.375	4.843	.000	.906	2.157

a. Dependent Variable: engach English Achievement

- In this example, CIs for *B*s indicate that we should not reject the possibility that the population *B*s are zero, except for Attentiveness (we are 95% sure that the true *B* for Attentiveness is between .91 and 2.16)

Confidence intervals: Practice question 1

Question
If a MLR predictor has a $B = .5$, with a 95% CI of .25 to .75, what should be concluded?
a) Do not reject H_0 (that $B = 0$)
b) Reject H_0 (that $B = 0$)

60

Confidence intervals: Practice question 2

Question
If a MLR predictor has a $B = .2$, with a 95% CI of -.2 to .6, what should be concluded?
a) Do not reject H_0 (that $B = 0$)
b) Reject H_0 (that $B = 0$)

61

Summary: Confidence intervals

1. Gives "range of certainty" when generalising from a sample to a target population
2. CIs be used for *M*, *B,* ES
3. Can be examined
   1. Statistically (upper and lower limits)
   2. Graphically (e.g., error-bar graphs)

62

Publication bias

63

Publication bias

- When the likelihood of publication of depends on their nature and direction of results.
- Significant effects are more likely to be published!
- Type I publication errors are underestimated to an extent that is: "frightening, even calling into question the scientific basis for much published literature." (Greenwald, 1975, p. 15)

Image source: https://commons.wikimedia.org/wiki/File:SMirC-shock.svg

64

File-drawer effect

- Tendency for non-significant 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.

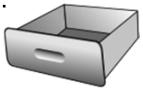


Image source: http://commons.wikimedia.org/wiki/File:Edge-drawer.png

65

Two counteracting biases

Two counteracting biases in social science research:

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

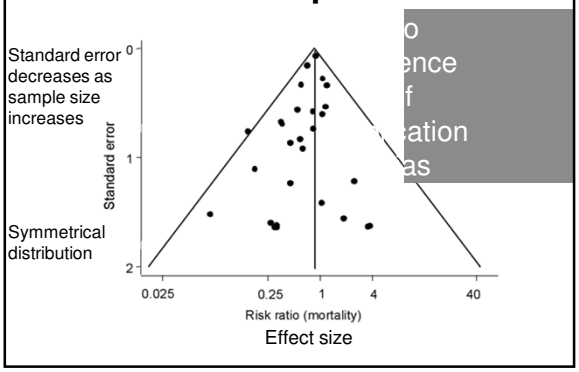
66

Funnel plots

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

67

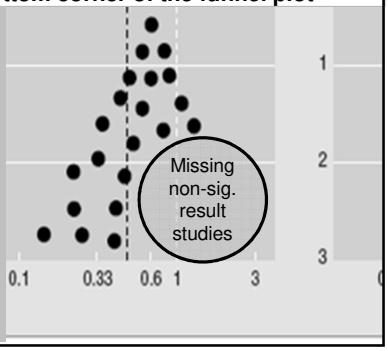
Funnel plots



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.

70

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.

http://www.jasnh.com

Countering the bias

JOURNAL OF NEGATIVE RESULTS

- ECOLOGY & EVOLUTIONARY BIOLOGY -

HOME ABOUT LOG IN REGISTER SEARCH CURRENT ARCHIVES

JOURNAL OF NEGATIVE RESULTS

The primary intention of Journal of Negative Results is to provide an online-medium to publish peer-reviewed, sound scientific work in ecology and evolutionary biology that is scientifically rigorous but does not rely upon arbitrary significance thresholds to support conclusions. In recent years, the trend has been to publish only studies with 'significant' results and to ignore studies that seem uneventful. This may lead to a biased, perhaps untrue, representation of what exists in nature. By counter-balancing such selective reporting, JNR aims to expand the capacity for formulating generalizations. The work to be published in JNR will include studies that 1) test novel or established hypotheses/theories that yield negative or dissenting results, or 2) replicate work published previously (in either cognate or different systems). Short notes on studies in which the data are biologically interesting but lack statistical power are also welcome. JNR also intends to present the results of studies in a format suitable for formal meta-analysis. Research quality is of highest importance for JNR. Manuscripts will be assessed for publication on this basis - positive results or support for current scientific dogma are not essential.

http://www.jnr-eeb.org/index.php/jnr

Summary: Publication bias

1. Tendency for statistically significant studies to be published over non-significant studies
2. Indicated by gap in funnel plot → file-drawer effect
3. Counteracting biases in scientific publishing; tendency:
   - towards low-power studies which underestimate effects
   - to publish sig. effects over non-sig. effects

73

Academic integrity

Academic integrity: Students
(Marsden, Carroll, & Neill, 2005)

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

75

Retraction watch

Tracking retractions
as a window
into the scientific process

https://retractionwatch.com/

Research investigations mounting for embattled University of New South Wales Professor Levon Khachigian

http://www.abc.net.au/news/2014-04-17/research-investigations-mounting-for-embattled-professor/5397516

Updated Fri 18 Apr 2014, 8:22am AEST

Millions of dollars in research money for the University of New South Wales has been frozen as multiple investigations into alleged research misconduct are launched.

The National Health and Medical Research Centre is withholding almost $8.4 million in funding it had awarded Professor Levon Khachigian following an investigation into the veracity of research papers about a skin cancer drug called DZ13.

Two investigations are currently being run into that research, and the university is about to establish another two inquiries.

Last year, the ABC revealed that the human clinical trial using DZ13 on skin cancer patients was stopped due to concerns about the science leading up to the trial.

UNSW Professor Levon Khachigian faces mounting investigations

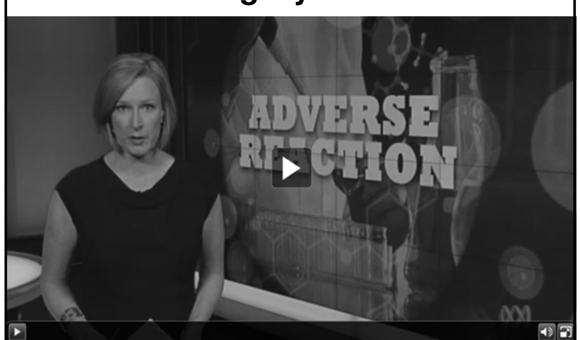
VIDEO: UNSW Professor Levon Khachigian faces mounting investigations (7pm TV News NSW)

PHOTO: More questions over research: Professor Levon Khachigian

RELATED STORY: Journal corrects paper by NSW academic facing misconduct probe

RELATED STORY: Fresh concerns over research conducted by UNSW professor Levon Khachigian

Academic integrity: Academic staff



http://www.abc.net.au/7.30/content/2013/s3823977.htm

Second study from University of Queensland research pair retracted from journal after misconduct probe http://www.abc.net.au/news/2014-04-04/uq-research-retraction-barwood-murdoch/5368800

By Elise Worthington

Updated Sat 5 Apr 2014, 11:45am AEDT

A second study by two controversial former University of Queensland researchers has been retracted, after statistical problems were found with the data.

The university examined 93 papers by Professor Bruce Murdoch and Dr Caroline Barwood, after problems with an article on Parkinson's disease last year.

Its investigation has now found concerns with the statistical methodology used in a June 2013 multiple sclerosis study by the pair.

The study, titled Cognitive Linguistic Deficits in Relapsing-Remitting Multiple Sclerosis, had been published in the journal Aphasiology.

A retraction notice on the publisher's website says the study claimed to have a control group of 15, but

PHOTO: Professor Murdoch and Dr Barwood no longer work at the university. (ABC News: Giulio Saggin)

RELATED STORY: More concerns over academics accused of misconduct

RELATED STORY: University investigates fresh claims of research misconduct

RELATED STORY: University investigates study with ins

Academic integrity: Academic staff

Richard Horton (2015), editor of "The Lancet" (one of the world's oldest and best-known medical journals):

1. "A lot of what is published is incorrect"
2. "Much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance."
3. "Scientists too often sculpt data to fit their preferred theory of the world. Or they retrofit hypotheses to fit their data."
4. "Our love of "significance" pollutes the literature with many a statistical fairy-tale."

http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(15)60696-1/fulltext

Summary: Academic integrity

1. Violations of academic integrity are most prevalent amongst those with incentives to cheat: e.g.,
   1. Students
   2. Competitively-funded researchers
   3. Commercially-sponsored researchers
2. Adopt a balanced, critical approach, striving for objectivity and academic integrity

81

Statistical Methods in Psychology Journals: Guidelines and Explanations

(Wilkinson, 1999)

https://www.apa.org/pubs/journals/releases/amp-54-8-594.pdf

Provides useful tips for good scientific writing e.g., for lab reports

Method - Design

Method

Design

Make clear at the outset what type of study you are doing. Do not cloak a study in one guise to try to give it the assumed reputation of another. For studies that have multiple goals, be sure to define and prioritize those goals.

(Wilkinson, 1999)

Method - Population

Population

The interpretation of the results of any study depends on the characteristics of the population intended for analysis. Define the population (participants, stimuli, or studies) clearly. If control or comparison groups are part of the design, present how they are defined.

(Wilkinson, 1999)

Method - Sample

Sample

Describe the sampling procedures and emphasize any inclusion or exclusion criteria. If the sample is stratified (e.g., by site or gender), describe fully the method and rationale. Note the proposed sample size for each subgroup.

(Wilkinson, 1999)

Method - Random assignment

Random assignment. *For research involving causal inferences, the assignment of units to levels of the causal variable is critical. Random assignment (not to be confused with random selection) allows for the strongest possible causal inferences free of extraneous assumptions. If random assignment is planned, provide enough information to show that the process for making the actual assignments is random.*

(Wilkinson, 1999)

Method - Nonrandom assignment

Nonrandom assignment. *For some research questions, random assignment is not feasible. In such cases, we need to minimize effects of variables that affect the observed relationship between a causal variable and an outcome. Such variables are commonly called confounds or covariates. The researcher needs to attempt to determine the relevant covariates, measure them adequately, and adjust for their effects either by design or by analysis. If the effects of covariates are adjusted by analysis, the strong assumptions that are made must be explicitly stated and, to the extent possible, tested and justified. Describe methods used to attenuate sources of bias, including plans for minimizing dropouts, noncompliance, and missing data.*

(Wilkinson, 1999)

Method - Instruments

Instruments. *If a questionnaire is used to collect data, summarize the psychometric properties of its scores with specific regard to the way the instrument is used in a population. Psychometric properties include measures of validity, reliability, and any other qualities affecting conclusions. If a physical apparatus is used, provide enough information (brand, model, design specifications) to allow another experimenter to replicate your measurement process.*

(Wilkinson, 1999)

Method - Variables

Variables. *Explicitly define the variables in the study, show how they are related to the goals of the study, and explain how they are measured. The units of measurement of all variables, causal and outcome, should fit the language you use in the introduction and discussion sections of your report.*

(Wilkinson, 1999)

Method - Procedure

Procedure. *Describe any anticipated sources of attrition due to noncompliance, dropout, death, or other factors. Indicate how such attrition may affect the generalizability of the results. Clearly describe the conditions under which measurements are taken (e.g., format, time, place, personnel who collected data). Describe the specific methods used to deal with experimenter bias, especially if you collected the data yourself.*

(Wilkinson, 1999)

Method - Power and sample size

Power and sample size. *Provide information on sample size and the process that led to sample size decisions. Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations. Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size. Once the study is analyzed, confidence intervals replace calculated power in describing results.*

(Wilkinson, 1999)

Results - Complications

Results

Complications

Before presenting results, report complications, protocol violations, and other unanticipated events in data collection. These include missing data, attrition, and nonresponse. Discuss analytic techniques devised to ameliorate these problems. Describe nonrepresentativeness statistically by reporting patterns and distributions of missing data and contaminations. Document how the actual analysis differs from the analysis planned before complications arose. The use of techniques to ensure that the reported results are not produced by anomalies in the data (e.g., outliers, points of high influence, nonrandom missing data, selection bias, attrition problems) should be a standard component of all analyses.

(Wilkinson, 1999)

Results - Min. sufficient analysis

Choosing a minimally sufficient analysis. *The enormous variety of modern quantitative methods leaves researchers with the nontrivial task of matching analysis and design to the research question. Although complex designs and state-of-the-art methods are sometimes necessary to address research questions effectively, simpler classical approaches often can provide elegant and sufficient answers to important questions. Do not choose an analytic method to impress your readers or to deflect criticism. If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. Occam's razor applies to methods as well as to theories.*

law of parsimony = all other things being equal, the simplest solution is the best

(Wilkinson, 1999)

Results - Use of software

Computer programs. *There are many good computer programs for analyzing data. More important than choosing a specific statistical package is verifying your results, understanding what they mean, and knowing how they are computed. If you cannot verify your results by intelligent "guesstimates," you should check them against the output of another program. You will not be happy if a vendor reports a bug after your data are in print (not an infrequent event). Do not report statistics found on a printout without understanding how they are computed or what they mean. Do not report statistics to a greater precision than is supported by your data simply because they are printed that way by the program. Using the computer is an opportunity for you to control your analysis and design. If a computer program does not provide the analysis you need, use another program rather than let the computer shape your thinking.*

(Wilkinson, 1999)

Results - Assumptions

Assumptions. *You should take efforts to assure that the underlying assumptions required for the analysis are reasonable given the data. Examine residuals carefully. Do not use distributional tests and statistical indexes of shape (e.g., skewness kurtosis) as a substitute for examining your residuals graphically.*

(Wilkinson, 1999)

Results - Hypothesis testing

Hypothesis tests. *It is hard to imagine a situation in which a dichotomous accept–reject decision is better than reporting an actual p value or, better still, a confidence interval. Never use the unfortunate expression "accept the null hypothesis." Always provide some effect-size estimate when reporting a p value.* Cohen

(Wilkinson, 1999)

Results - Effect sizes

Effect sizes. *Always present effect sizes for primary outcomes. If the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day), then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure (r or d). It helps to add brief comments that place these effect sizes in a practical and theoretical context.*

(Wilkinson, 1999)

Results - Interval estimates

Interval estimates. *Interval estimates should be given for any effect sizes involving principal outcomes. Provide intervals for correlations and other coefficients of association or variation whenever possible.*

(Wilkinson, 1999)

Results - Multiplicities

Multiplicities. *Multiple outcomes require special handling. There are many ways to conduct reasonable inference when faced with multiplicity (e.g., Bonferroni correction of p values, multivariate test statistics, empirical Bayes methods). It is your responsibility to define and justify the methods used.*

(Wilkinson, 1999)

Results - Causality

Causality. *Inferring causality from nonrandomized designs is a risky enterprise. Researchers using nonrandomized designs have an extra obligation to explain the logic behind covariates included in their designs and to alert the reader to plausible rival hypotheses that might explain their results. Even in randomized experiments, attributing causal effects to any one aspect of the treatment condition requires support from additional experimentation.*

(Wilkinson, 1999)

Results - Tables and figures

Tables and figures. *Although tables are commonly used to show exact values, well-drawn figures need not sacrifice precision. Figures attract the reader's eye and help convey global results. Because individuals have different preferences for processing complex information, it often helps to provide both tables and figures. This works best when figures are kept small enough to allow space for both formats. Avoid complex figures when simpler ones will do. In all figures, include graphical representations of interval estimates whenever possible.*

(Wilkinson, 1999)

Discussion - Interpretation

Discussion

Interpretation

When you interpret effects, think of credibility, generalizability, and robustness. Are the effects credible, given the results of previous studies and theory? Do the features of the design and analysis (e.g., sample quality, similarity of the design to designs of previous studies, similarity of the effects to those in previous studies) suggest the results are generalizable? Are the design and analytic methods robust enough to support strong conclusions?

(Wilkinson, 1999)

Discussion - Conclusions

Conclusions

Speculation may be appropriate, but use it sparingly and explicitly. Note the shortcomings of your study. Remember, however, that acknowledging limitations is for the purpose of qualifying results and avoiding pitfalls in future research. Confession should not have the goal of disarming criticism. Recommendations for future research should be thoughtful and grounded in present and previous findings. Gratuitous suggestions ("further research needs to be done ...") waste space. Do not interpret a single study's results as having importance independent of the effects reported elsewhere in the relevant literature. The thinking presented in a single study may turn the movement of the literature, but the results in a single study are important primarily as one contribution to a mosaic of study effects. (Wilkinson, 1999)

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 (Wikiversity) http://en.wikiversity.org/wiki/Publication_bias

104

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/MarsdenCarrollNeill2005WhoCheatsAtUniversity.htm

2 Ward, R. M. (2002). *Highly significant findings in psychology: A power and effect size survey*. http://digitalcommons.uri.edu/dissertations/AAI3053127/

3 Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, process, and purpose*. doi:10.1080/00031305.2016.1154108

4 Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604. https://www.apa.org/pubs/journals/releases/amp-54-8-594.pdf

105

Next lecture

Summary and conclusion

- Recap of previous 9 lectures
- Review of learning outcomes

106