UNDERSTANDING STATISTICS & EXPERIMENTAL DESIGN

Content

1. Basic Probability Theory
2. Signal Detection Theory (SDT)
3. SDT and Statistics I and II
4. Statistics in a nutshell
5. Multiple Testing
6. ANOVA
7. Experimental Design & Statistics
8. Correlations & PCA
9. Meta-Statistics: Basics
10. Meta-Statistics: Too good to be true
11. Meta-Statistics: How big a problem is publication bias?
12. Meta-Statistics: What do we do now?

How big a problem is publication bias?

Greg Francis – lecture III

Does bias really matter?

- I ran three sets of simulated experiments
  - Two sample *t*-tests
- **Set 1**
  - True effect size = 0
  - Sample size: data peeking, starting with $n_1 = n_2 = 10$ and going to 30
  - 20 experiments
  - Reported only the 5 experiments that rejected the null hypothesis
- **Set 2**
  - True effect size = 0.1
  - Sample size randomly chosen between 10 and 30
  - 100 experiments
  - Reported only the 5 experiments that rejected the null hypothesis

Does bias really matter?

- I ran three sets of simulated experiments
  - Two sample *t*-tests
- **Set 3**
  - True effect size = 0.8
  - Sample size randomly chosen between 10 and 30
  - 5 experiments
  - All experiments rejected the null and were reported

- The following tables give you information about the reported experiments. Which is the valid set?

Does bias really matter?

n1=n2	t	p	g
10	2.48	0.03	1.06
28	2.10	0.04	0.55
10	3.12	0.01	1.34
15	2.25	0.04	0.80
12	2.34	0.03	0.92

n1=n2	t	p	g
21	2.67	0.01	0.81
27	4.72	<0.01	1.26
22	3.66	<0.01	1.08
26	2.74	0.01	0.75
24	2.06	0.05	0.58

n1=n2	t	p	g
16	2.10	0.04	0.72
19	2.19	0.04	0.70
25	2.22	0.03	0.62
14	2.24	0.04	0.82
23	2.49	0.02	0.72

Does bias really matter?

n1=n2	t	p	g
10	2.48	0.03	1.06
28	2.10	0.04	0.55
10	3.12	0.01	1.34
15	2.25	0.04	0.80
12	2.34	0.03	0.92

$g^* = 0.82$

n1=n2	t	p	g
21	2.67	0.01	0.81
27	4.72	<0.01	1.26
22	3.66	<0.01	1.08
26	2.74	0.01	0.75
24	2.06	0.05	0.58

$g^* = 0.89$

n1=n2	t	p	g
16	2.10	0.04	0.72
19	2.19	0.04	0.70
25	2.22	0.03	0.62
14	2.24	0.04	0.82
23	2.49	0.02	0.72

$g^* = 0.70$

Does bias really matter?

n1=n2	t	p	g
10	2.48	0.03	1.06
28	2.10	0.04	0.55
10	3.12	0.01	1.34
15	2.25	0.04	0.80
12	2.34	0.03	0.92

$g^* = 0.82$
Prob(all 5 reject) = 0.042

n1=n2	t	p	g
21	2.67	0.01	0.81
27	4.72	<0.01	1.26
22	3.66	<0.01	1.08
26	2.74	0.01	0.75
24	2.06	0.05	0.58

$g^* = 0.89$
Prob(all 5 reject) = 0.45

n1=n2	t	p	g
16	2.10	0.04	0.72
19	2.19	0.04	0.70
25	2.22	0.03	0.62
14	2.24	0.04	0.82
23	2.49	0.02	0.72

$g^* = 0.70$
Prob(all 5 reject) = 0.052

Does bias really matter?

n1=n2	t	p	g
10	2.48	0.03	1.06
28	2.10	0.04	0.55
10	3.12	0.01	1.34
15	2.25	0.04	0.80
12	2.34	0.03	0.92

$g^* = 0.82$
Prob(all 5 reject) = 0.042
r = -0.86

n1=n2	t	p	g
21	2.67	0.01	0.81
27	4.72	<0.01	1.26
22	3.66	<0.01	1.08
26	2.74	0.01	0.75
24	2.06	0.05	0.58

$g^* = 0.89$
Prob(all 5 reject) = 0.45
r = 0.25

n1=n2	t	p	g
16	2.10	0.04	0.72
19	2.19	0.04	0.70
25	2.22	0.03	0.62
14	2.24	0.04	0.82
23	2.49	0.02	0.72

$g^* = 0.70$
Prob(all 5 reject) = 0.052
r = - 0.83

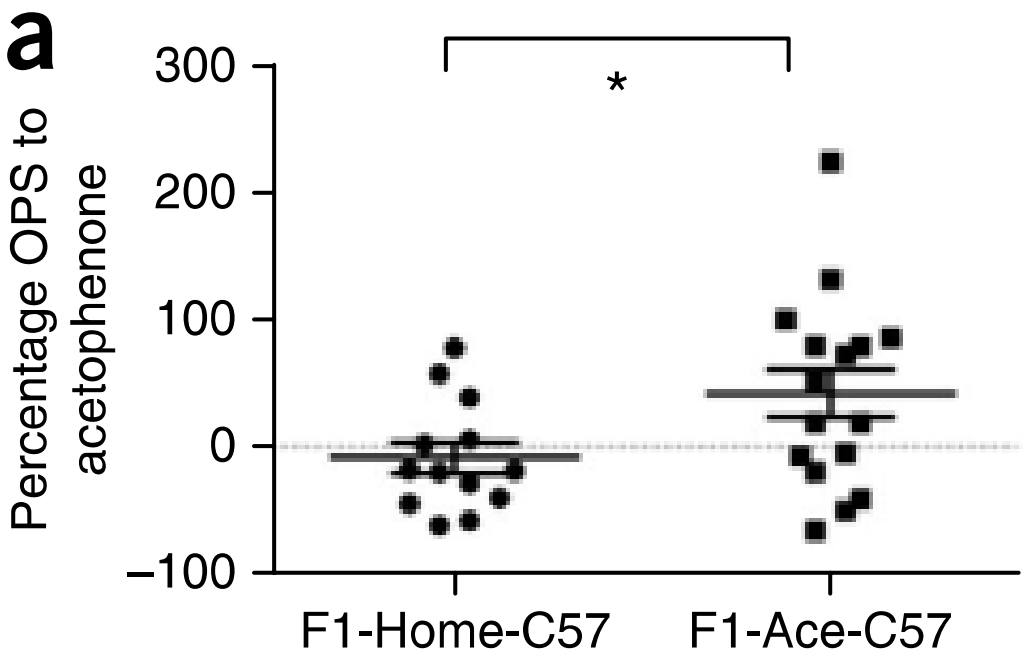
Is Bias A Problem For Important Findings?

- We might not care so much about bias if it is for effects that matter very little
- We can explore bias in “important” findings
- We can explore bias in “prominent” journals

- I call the analysis the “Test for Excess Success” (TES)
  - This generalizes what we discussed last time
  - Do not have to do a meta-analysis of reported studies to compute power
  - Use *post hoc* (observed) power estimates (carefully)

Dias & Ressler (2014)

- "Parental olfactory experience influences behavior and neural structure in subsequent generations," *Nature Neuroscience*
- Experiment in Figure 1a is representative
  - Male mice subjected to fear conditioning in the presence of the odor acetophenone
  - Their *offspring* exhibited significantly enhanced sensitivity to acetophenone
    - Compared to the offspring of unconditioned controls
  - n_1=16, n_2=13, t=2.123, p=.043, g=0.770, power=0.512



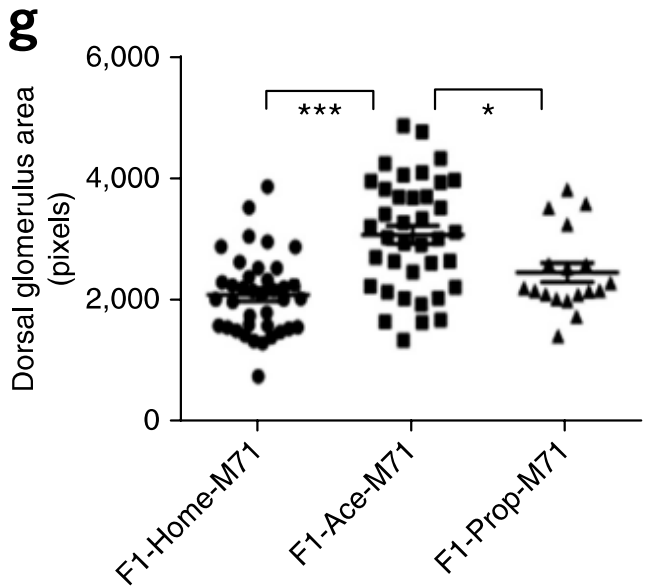
Dias & Ressler (2014)

- All the experiments were successful
- Probability of all 10 experiments being successful is P_{TES}=.023
- Indicates the results seem “too good to be true”
- Researchers should be skeptical about the results and/or the conclusions

Exp.	Sample sizes	Reported Inference	Probability of success
Figure 1a	16, 13	μ1≠μ2	0.512
Figure 1b	7, 9	μ1=μ2	0.908
Figure 1c	11, 13, 19	ANOVA, μ1≠μ2, μ2≠μ3, μ1 ≥μ3	0.662
Figure 1d	10, 11, 8	ANOVA, μ1=μ2, μ2≠μ3	0.712
Figure 2a	16, 16	μ1≠μ2	0.663
Figure 2b	16, 16	μ1≠μ2	0.928
Figure 4a	8, 12	μ1≠μ2	0.675
Figure 4b	8, 11	μ1≠μ2	0.545
Figure 5a	13, 16	μ1≠μ2	0.6
Figure 5b	4, 7, 6, 5	ANOVA, μ1≠μ2, μ3≠μ4	0.775

Dias & Ressler (2014)

- Further support provided by 12 neuroanatomy studies (staining of olfactory bulb for areas sensitive to acetophenone)
- The experiment in Figure 3g is representative
  - Group 1: Control
  - Group 2: Offspring from male mice subjected to fear conditioning in the presence of the odor acetophenone
  - Group 3: Offspring from male mice subjected to fear conditioning in the presence of the odor propanol
- Three tests reported as being important (post hoc power)
  - ANOVA (0.999)
  - $\mu_1 \neq \mu_2$ (0.999)
  - $\mu_2 \neq \mu_3$ (0.782)
- Joint success (0.782)



Dias & Ressler (2014)

- Probability of all 12 neuroanatomy experiments being successful is P_{TES}=.189
- This is above the criterion (.1)
- Researchers do not need to be skeptical about the results and/or the conclusions

Exp.	Sample sizes	Reported Inference	Probability of success
Figure 3g	38, 38, 18	ANOVA, μ1≠μ2, μ2≠μ3	0.782
Figure 3h	31, 40, 16	ANOVA, μ1≠μ2, μ2≠μ3	≈1.00
Figure 3i	6, 6, 4	ANOVA, μ1≠μ2, μ2≠μ3	0.998
Figure 4g	7, 8	μ1≠μ2	0.999
Figure 4h	6, 10	μ1≠μ2	0.974
Figure 4i	23, 16	μ1≠μ2	0.973
Figure 4j	16, 19	μ1≠μ2	≈1.00
Figure 5g	6, 4, 5, 3	ANOVA, μ1≠μ2, μ3≠μ4, μ1=μ3	0.892
Figure 5h	4, 3, 8, 4	ANOVA, μ3≠μ4, μ1=μ3	0.824
Figure 6a	12, 10	μ1≠μ2	0.574
Figure 6c	12, 10	μ1=μ2	0.901
Figure 6e	8, 8	μ1≠μ2	0.681

Dias & Ressler (2014)

- Success for the theory about epigenetics required both the behavioral and neuroanatomy findings to be successful
- Probability of all 22 experiments being successful is
  - P_{TES}= P_{TES}(Behavior) x P_{TES}(Neuroanatomy)
  - P_{TES}= 0.023 x 0.189 = 0.004
- Indicates the results seem “too good to be true”
- Francis (2014) “Too Much Success for Recent Groundbreaking Epigenetic Experiments” *Genetics*.

Dias & Ressler (2014)

- Reply by Dias and Ressler (2014):
  - “we have now replicated these effects multiple times within our laboratory with multiple colleagues as blinded scorers, and we fully stand by our initial observations.”
- More successful replication only makes their results **less** believable
- It is not clear if they “stand by” the magnitude of the reported effects or by the rate of reported replication (100%)
  - These two aspects of the report are in conflict so it is difficult to stand by both findings

Psychological Science

- Flagship journal of the Association for Psychological Science
- Presents itself as an outlet for the very best research in the field
- Sent to 20,000 APS members
- Acceptance rate of around 11%

- The editor-in-chief, Eric Eich, asked me to apply the Test for Excess Success analysis to all articles that had four or more experiments and report the findings

TES Analysis

- I downloaded all 951 articles published in Psychological Science for 2009-2012.
- There were 79 articles that had four or more experiments
- The analysis requires calculation of the probability of experimental success
  - 35 articles did not meet this requirement (for a variety of reasons)
- The remaining 44 articles were analyzed to see if the rate of reported experimental success matched the rate that should appear if the experiments were run properly and fully reported
  - The analysis is within each article, not across articles

TES Analysis For *PSCI*

2012: 7 out of 10 articles have $P_{TES} \leq .1$

Authors	Short title	P_{TES}
Anderson, Kraus, Galinsky & Keltner	Sociometric Status and Subjective Well-Being	.167
Bauer, Wilkie, Kim & Bodenhausen	Cuing Consumerism	.062
Birtel & Crisp	Treating Prejudice	.133
Converse, Risen & Carter	Karmic Investment	.043
Converse & Fishbach	Instrumentality Boosts Appreciation	.110
Keysar, Hayakawa & An	Foreign-Language Effect	.091
Leung, Kim, Polman, Ong, Qiu, Goncalo & Sanchez-Burks	Embodied Metaphors and Creative "Acts"	.076
Rounding, Lee, Jacobson & Ji	Religion and Self-Control	.036
Savani & Rattan	Choice and Inequality	.064
van Boxtel & Koch	Visual Rivalry Without Spatial Conflict	.071

TES Analysis For *PSCI*

2011: 5 out of 6
articles have $P_{TES} \leq .1$

Authors	Short title	P_{TES}
Evans, Horowitz & Wolfe	Weighting of Evidence in Rapid Scene Perception	.426
Inesi, Botti, Dubois, Rucker & Galinsky	Power and Choice	.026
Nordgren, Morris McDonnell, & Loewenstein	What Constitutes Torture?	.090
Savani, Stephens & Markus	Interpersonal and Societal Consequences of Choice	.063
Todd, Hanko, Galinsky & Mussweiler	Difference Mind-Set and Perspective Taking	.043
Tuk, Trampe & Warlop	Inhibitory Spillover	.092

TES Analysis For PSCI

2010: 12 out of 14 articles have $P_{TES} \leq .1$

Authors	Short title	P_{TES}
Balcetis & Dunning	Wishful Seeing	.076
Bowles & Gelfand	Status and Workplace Deviance	.057
Damisch, Stoberock & Mussweiler	How Superstition Improves Performance	.057
de Hevia & Spelke	Number-Spacing Mapping in Human Infants	.070
Ersner-Hershfield, Galinsky, Kray & King	Counterfactual Reflection	.073
Gao, McCarthy & Scholl	The Wolfpack Effect	.115
Lammers, Stapel & Galinsky	Power and Hypocrisy	.024
Li, Wei & Soman	Physical Enclosure and Psychological Closure	.079
Maddux, Yang, Falk, Adam, Adair, Endo, Carmon & Heine	Culture and the Endowment Effect	.014
McGraw & Warren	Benign Violations	.081
Sackett, Meyvis, Nelson, Converse & Sackett	When Time Flies	.033
Savani, Markus, Naidu, Kumar & Berlia	What Counts as a Choice?	.058
Senay, Albarracín & Noguchi	Interrogative Self-Talk and Intention	.090
West, Anderson, Bedwell & Pratt	Red Diffuse Light Suppresses Fear Prioritization	.157

TES Analysis For PSCI

2009: 12 out of 14 articles have $P_{TES} \leq .1$

Authors	Short title	P_{TES}
Alter & Oppenheimer	Fluencey and Self-Disclosure	.071
Ashton-James, Maddux, Galinsky & Chartrand	Affect and Culture	.035
Fast & Chen	Power, Incompetence, and Aggression	.072
Fast, Gruenfeld, Sivanathan & Galinsky	Power and Illusory Control	.069
Garcia & Tor	The N-Effect	.089
González & McLennan	Hemispheric Differences in Sound Recognition	.139
Hahn, Close & Graf	Transformation Direction	.348
Hart & Albarracín	Describing Actions	.035
Janssen & Caramazza	Phonology and Grammatical Encoding	.083
Jostmann, Lakens & Schubert	Weight and Importance	.090
Labroo, Lambotte & Zhang	The Name-Ease Effect and Importance Judgments	.008
Nordgren, van Harreveld & van der Pligt	Restraint Bias	.0998
Wakslak & Trope	Construal Level and Subjective Probability	.061
Zhou, Vohs & Baumeister	Symbolic Power of Money	.041

TES Analysis For *PSCI*

- In all, 36 of the 44 articles (82%) produce $P_{TES} \leq .1$
  - Details in Francis (2014, *Psychonomic Bulletin & Review*)
- I do **not** believe these authors deliberately misled the field
- I do believe that these authors did not make good scientific arguments to support their theoretical claims
  - They may have inappropriately sampled their data
  - They may have practiced “p-hacking”
  - They may have interpreted unsuccessful experiments as being methodologically flawed rather than as evidence against their theory
  - They may have “over fit” the data by building a theory that perfectly matched the reported significant and non-significant findings
- To me, these findings indicate serious problems with standard scientific practice

Science

- Flagship journal of the American Association for the Advancement of Science
- Presents itself as: “The World’s Leading Journal of Original Scientific Research, Global News, and Commentary”
- Sent to “over 120,000” subscribers
- Acceptance rate of around 7%
- I downloaded all 133 original research articles that were classified as Psychology or Education for 2005-2012
- 26 articles had 4 or more experiments
- 18 articles provided enough information to compute success probabilities for 4 or more experiments
- Francis, Tanzman & Williams (2014, *Plos One*)

TES Analysis For *Science*

2005-2012: 15 out of 18 articles (83%) have $P_{TES} \leq .1$

Authors	Short title	P_{TES}
Dijksterhuis et al. (2006)	Deliberation-Without-Attention Effect	0.051
Vohs et al. (2006)	Psychological Consequences of Money	0.002
Zhong & Lijenquist (2006)	Washing Away Your Sins	0.095
Wood et al. (2007)	Perception of Goal-Directed Action in Primates	0.031
Whitson & Galinsky (2008)	Lacking Control Increases Illusory Pattern Perception	0.008
Mehta & Zhu (2009)	Effect of Color on Cognitive Performance	0.002
Paukner et al. (2009)	Monkeys Display Affiliation Toward Imitators	0.037
Weisbuch et al. (2009)	Race Bias via Televised Nonverbal Behavior	0.027
Ackerman et al. (2010)	Incidental Haptic Sensations Influence Decisions	0.017
Bahrami et al. (2010)	Optimally Interacting Minds	0.332
Kovács et al. (2010)	Susceptibility to Others' Beliefs in Infants and Adults	0.021
Morewedge et al. (2010)	Imagined Consumption Reduces Actual Consumption	0.012
Halperine et al. (2011)	Promoting the Middle East Peace Process	0.210
Ramirez & Beilock (2011)	Writing About Worries Boosts Exam Performance	0.059
Stapel & Lindenberg (2011)	Disordered Contexts Promote Stereotyping	0.075
Gervais & Norenzayan (2012)	Analytic Thinking Promotes Religious Disbelief	0.051
Seeley et al. (2012)	Stop Signals Provide Inhibition in Honeybee Swarms	0.957
Shah et al. (2012)	Some Consequences of Having Too Little	0.091

What Does It All Mean?

- I think it means there are some fundamental misunderstandings about the scientific method
- It highlights that doing good science is really difficult
- Consider four statements that seem like principles of science, but often do not apply to psychology studies
    1) Replication establishes scientific truth
    2) More data are always better
    3) Let the data define the theory
    4) Theories are proven by validating predictions

(1) Replication

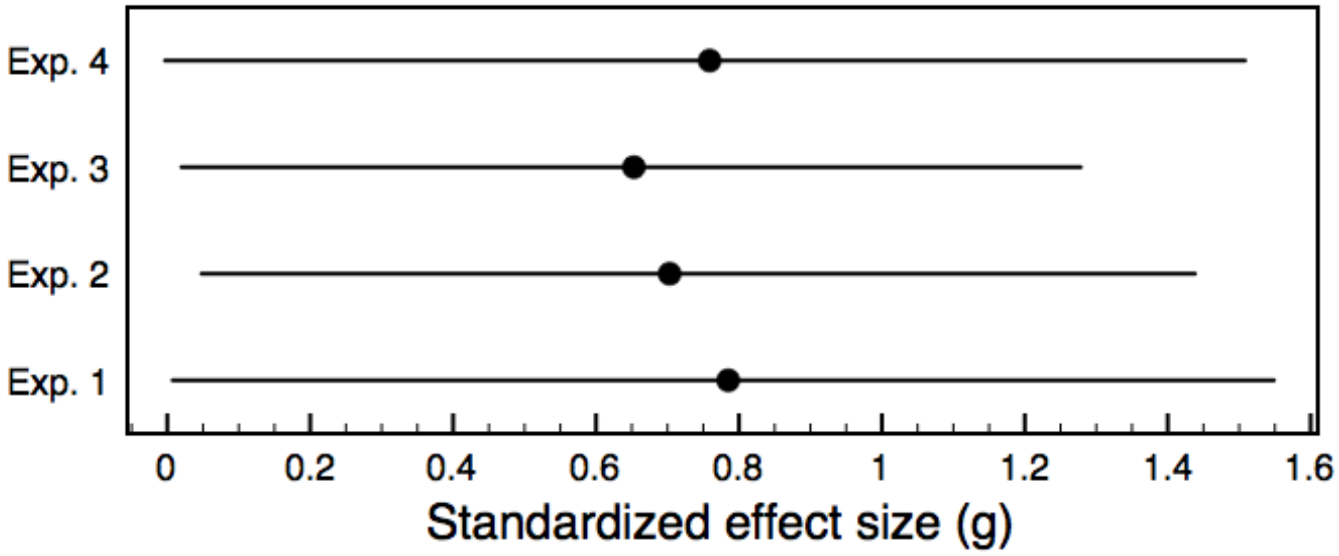
- Successful replication is often seen as the “gold standard” of empirical work
- But when success is defined statistically (e.g., significance), proper experiment sets show successful replication at a rate that matches experimental power
- Experiments with moderate or low power that always reject the null are a cause for concern
- Recent reform efforts are calling for more replication, but this call misunderstands the nature of our empirical investigations

(2) More Data

- Our statistics improve with more data, and it seems that more data brings us closer to scientific truth
- Authors might add more data when they get *p*=.07 but not when they get *p*=.03 (optional stopping)
- Similar problems arise across experiments, where an author adds Experiment 2 to check on the marginal result in Experiment 1
- Collecting more data is not wrong in principle, but it leads to a loss of Type I error control
  - The problem exists even if you get *p*=.03 but would have added subjects had you gotten *p*=.07. It is the **stopping**, not the **adding**, that is a problem

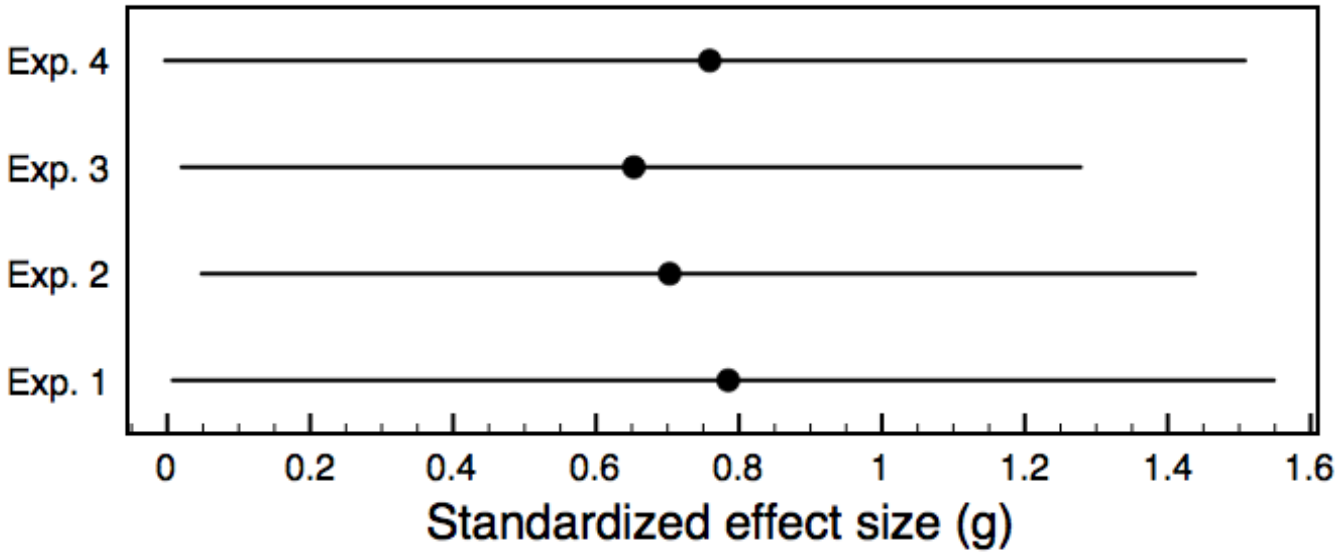
(3) Let The Data Define The Theory

- Theories that do not match data must be changed or rejected
- But the effect of data on theory depends on the precision of the data and the precision of the theory
- Consider the precision of the standardized effect sizes in one of the studies in *Psychological Science*



(3) Let The Data Define The Theory

- The data tell us very little about the measured effect
- Such data cannot provide strong evidence for *any* theory
  - A theory that perfectly matches the data is matching both signal *and* noise
- **H**ypothesizing **a**fter the **r**esults are **k**nown (HARKing)
  - Kerr (1988)



(3) Let The Data Define The Theory

- Scientists can try various analysis techniques until getting a desired result
  - Transform data (e.g., log, inverse)
  - Remove outliers (e.g., > 3 sd, >2.5 sd, ceiling/floor effects)
  - Combine measures (e.g., blast of noise: volume, duration, volume*duration, volume+duration)

- Causes loss of Type I error control
  - A 2x2 ANOVA where the null is true has a 14% chance of finding at least one p<.05 from the main effects and interaction (higher if also consider various contrasts)

(4) Theory Validation

- Scientific arguments are very convincing when a theory predicts a novel outcome that is then verified
- A common phrase in the *Psychological Science* and *Science* articles is “as predicted by the theory…”

- We need to think about what it means for a theory to predict the outcome of a hypothesis test
  - Even if an effect is real, not every sample will produce a significant result
  - At best, a theory can predict the *probability* (power) of rejecting the null hypothesis

(4) Theory Validation

- To predict power, a theory must indicate an effect size for a given experimental design and sample size

- **None** of the articles in *Psychological Science* or *Science* included a discussion of predicted effect sizes and power
  - So, none of the articles formally predicted the outcome of the hypothesis tests
  - The fact that essentially every hypothesis test matched the “prediction” is bizarre
  - It implies success at a fundamentally impossible task

(4) Theory Validation

There are two problems with how many scientists theorize

1) Not trusting the data:

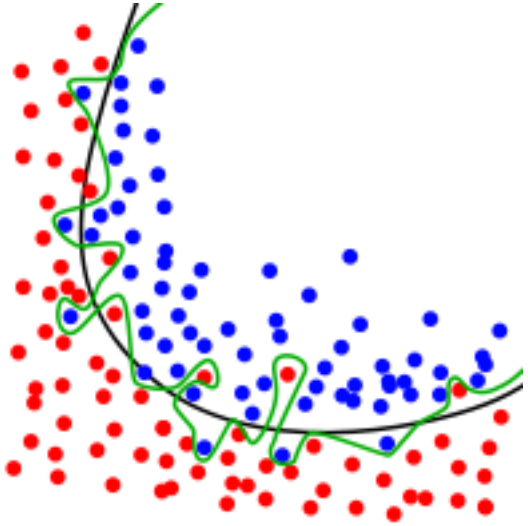
Search for confirmation of ideas (publication bias, p-hacking)

“Explain away” contrary results (replication failures)

2) Trusting data too much:

Theory becomes whatever pattern of significant and non-significant results are found in the data

Some theoretical components are determined by “noise”



Conclusions

- Faulty statistical reasoning (publication bias and related issues) misrepresent reality
- Faulty statistical reasoning appears to be present in reports of important scientific findings
- Faulty statistical reasoning appears to be common in top journals

Take home messages

Take Home Messages

1. There is too much replication in many fields including medicine, biology, psychology, and likely many more.
2. It seems that many scientists use techniques they should better avoid: optional stopping, publication bias, HARKing, flexibility in the analysis, and many more.

END Class 11