# Surprise, Surprise: It's a false positive

Jarrod Hadfield

University of Edinburgh

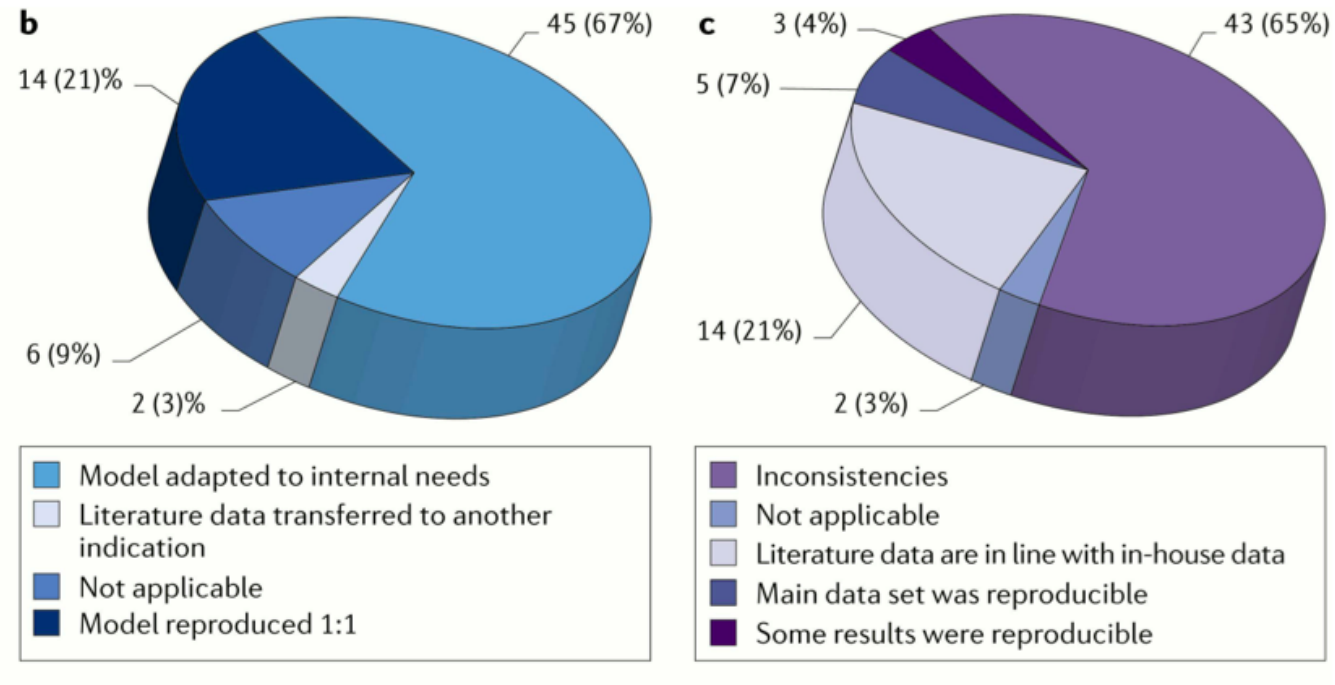
November 24, 2017

# Reproducibility in Medicine

- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10

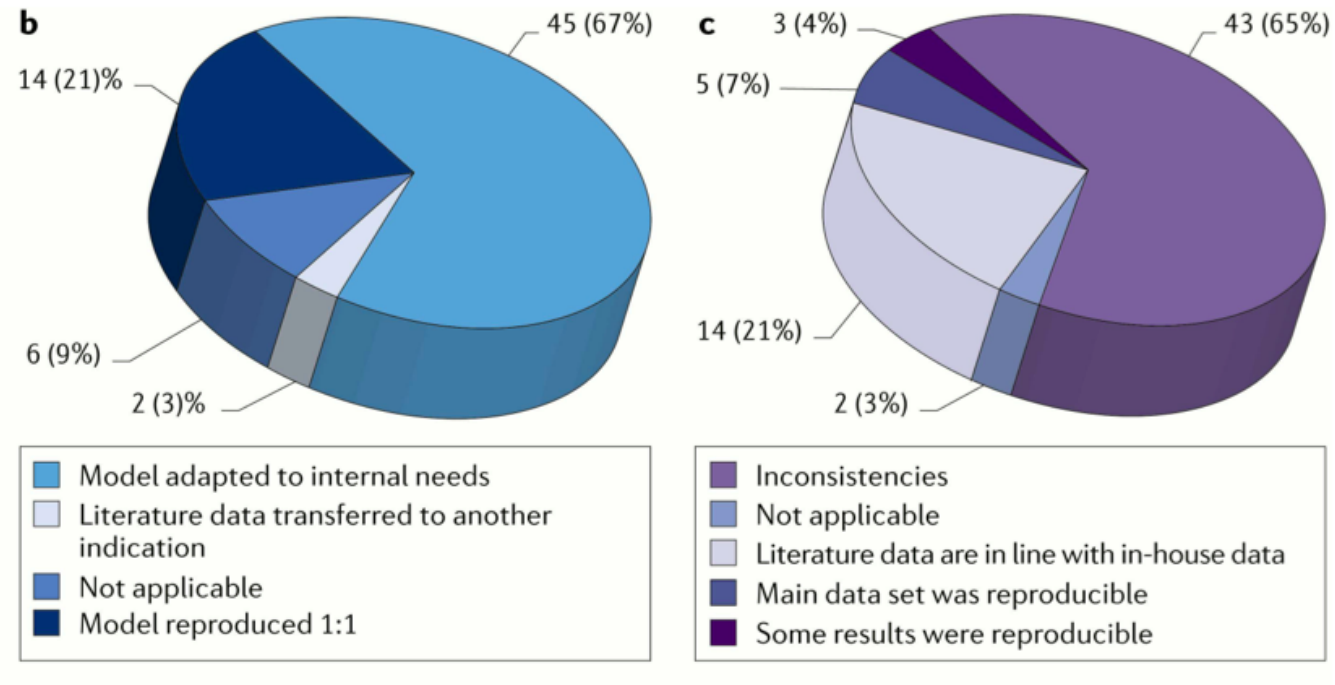
# Reproducibility in Medicine

- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10



# Reproducibility in Medicine

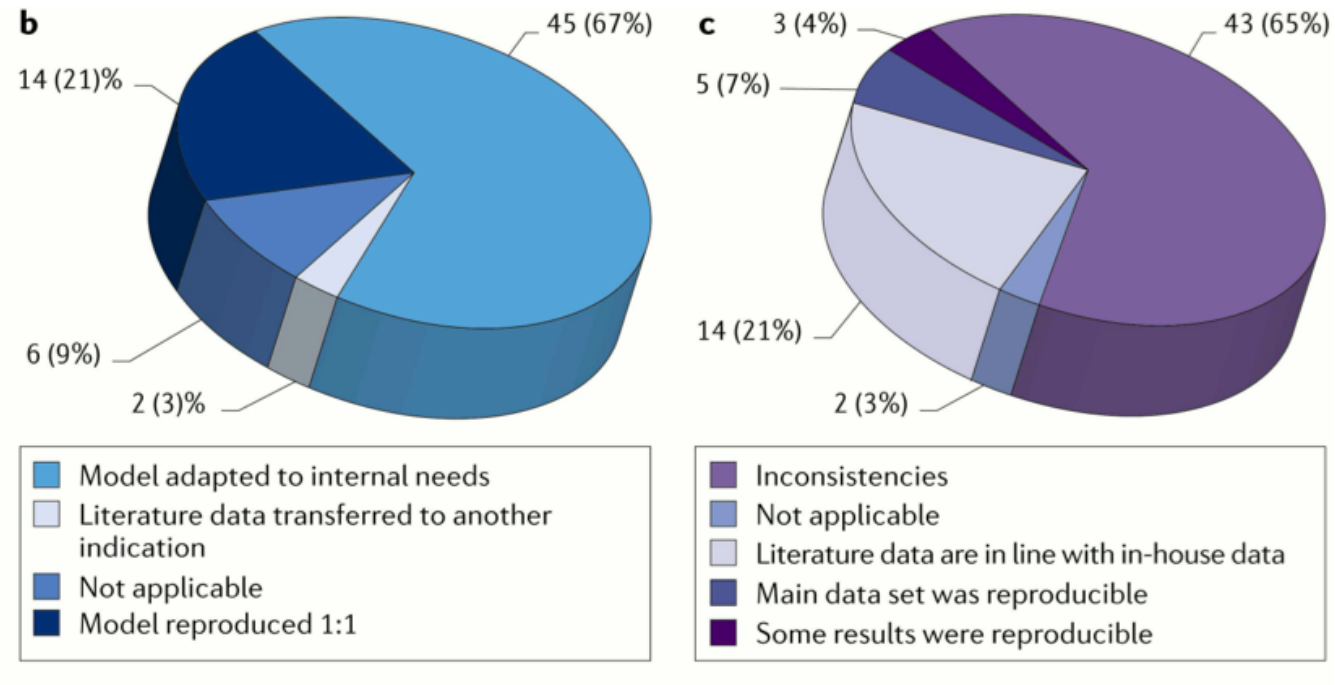
- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10




- Begley, CG. & Ellis, LM. (2012) Drug development: Raise standards for preclinical cancer research. Nature 483

# Reproducibility in Medicine

- Prinz, F. et al. (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 10




- Begley, CG. & Ellis, LM. (2012) Drug development: Raise standards for preclinical cancer research. Nature 483
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349

# Scientific Integrity and Transparency

- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

# Scientific Integrity and Transparency

- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

  Bruce Alberts (then the editor of Science):

# Scientific Integrity and Transparency

- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

  Bruce Alberts (then the editor of Science):

  - *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*

# Scientific Integrity and Transparency

- U.S. House of Representatives Hearing on 'Scientific Integrity and Transparency' (2013)

  Bruce Alberts (then the editor of Science):

  - *'Budding scientists must be taught technical skills, including statistics, and must be imbued with scepticism towards their own results and those of others. Researchers ought to be judged on the basis of the quality, not the quantity, of their work.'*

  - *'We need to develop a value system where simply moving on from one's mistakes without publicly acknowledging them severely damages, rather than protects, a scientific reputation.'*

# How Science Goes Wrong

# How Science Goes Wrong


*'Science still commands enormous - if sometimes bemused - respect. But its privileged status is founded on the capacity to be right most of the time and to correct its mistakes when it gets things wrong. And it is not as if the universe is short of genuine mysteries to keep generations of scientists hard at work. The false trails laid down by shoddy research are an unforgivable barrier to understanding.'*

# Surprise, Surprise: It's a false positive

- Background
- Problems
- Under-graduate Project
- Solutions
- Unacknowledged Issues

# Problems

- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.

# Problems

- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.

# Problems

- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.
- Statistical mistakes/misinterpretation are widespread.

# Problems

- The reward system favours those that publish statistically significant high-profile work quickly without future correction/validation.
- The quality of peer review is not adequate.
- Statistical mistakes/misinterpretation are widespread.
- A lot of scientific research is poorly thought through.

# How Science Goes Wrong

- Statistical mistakes/misinterpretation are widespread.
  - **Wacholder, S. et al.** (2004) *Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies.* Journal of the National Cancer Institute 96, 434-442
  - **Ioannidis, JPA.** (2005) *Why Most Published Research Findings Are False* PLoS Medicine 2 e124

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)

| | Significant | Not Signficant |
|---|---|---|
| No Association | [False Positive] $\alpha$ | [True Negative] $(1 - \alpha)$ |

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true ($\beta$ is the Type II error rate)

| | Significant | Not Signficant |
|---|---|---|
| No Association | [False Positive] $\alpha$ | [True Negative] $(1 - \alpha)$ |

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true ($\beta$ is the Type II error rate)

| | Significant | Not Signficant |
|---|---|---|
| No Association | [False Positive] $\alpha$ | [True Negative] $(1 - \alpha)$ |
| True Association | [True Positive] $(1 - \beta)$ | [False Negative] $\beta$ |

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true ($\beta$ is the Type II error rate)

| | Significant | Not Signficant |
|---|---|---|
| No Association | [False Positive] $\alpha$ | [True Negative] $(1 - \alpha)$ |
| True Association | [True Positive] $(1 - \beta)$ | [False Negative] $\beta$ |

Probability the alternative hypothesis is true *given* a significant result:

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true ($\beta$ is the Type II error rate)

| | Significant | Not Signficant |
|---|---|---|
| No Association | [False Positive] $\alpha$ | [True Negative] $(1 - \alpha)$ |
| True Association | [True Positive] $(1 - \beta)$ | [False Negative] $\beta$ |

Probability the alternative hypothesis is true *given* a significant result:

$$\frac{(1 - \beta)}{(1 - \beta) \quad + \alpha}$$

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)
- $1 - \beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true ($\beta$ is the Type II error rate)
- $\pi$ is the prior probability that the alternative hypothesis is true

| | Significant | Not Signficant |
|---|---|---|
| No Association | [False Positive] $\alpha$ | [True Negative] $(1 - \alpha)$ |
| True Association | [True Positive] $(1 - \beta)$ | [False Negative] $\beta$ |

Probability the alternative hypothesis is true *given* a significant result:

$$\frac{(1 - \beta)}{(1 - \beta) \quad + \alpha}$$

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)
- $1-\beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true ($\beta$ is the Type II error rate)
- $\pi$ is the prior probability that the alternative hypothesis is true

| | Significant | Not Signficant |
|---|---|---|
| No Association | [False Positive] $\alpha(1-\pi)$ | [True Negative] $(1-\alpha)(1-\pi)$ |
| True Association | [True Positive] $(1-\beta)\pi$ | [False Negative] $\beta\pi$ |

Probability the alternative hypothesis is true *given* a significant result:

$$\frac{(1-\beta)}{(1-\beta) \quad +\alpha}$$

# False Positive Report Probability

- $\alpha$ is the probability of a statistically significant finding, given that the null hypothesis is true ($\alpha$ is the Type I error rate)
- $1-\beta$ is the power; the probability of a statistically significant finding given the alternative hypothesis is true ($\beta$ is the Type II error rate)
- $\pi$ is the prior probability that the alternative hypothesis is true

| | Significant | Not Signficant |
|---|---|---|
| No Association | [False Positive] $\alpha(1-\pi)$ | [True Negative] $(1-\alpha)(1-\pi)$ |
| True Association | [True Positive] $(1-\beta)\pi$ | [False Negative] $\beta\pi$ |

Probability the alternative hypothesis is true *given* a significant result:

$$\frac{(1-\beta)\pi}{(1-\beta)\pi+\alpha(1-\pi)}$$

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

| Scenario | Effect Size | Power | 1-FPRP |
|---|---|---|---|
| Ideal | | 0.80 | 0.94 |

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

| Scenario | Effect Size | Power | 1-FPRP |
|---|---|---|---|
| Ideal | | 0.80 | 0.94 |

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

| Scenario | Effect Size | Power | 1-FPRP |
|---|---|---|---|
| Ideal | | 0.80 | 0.94 |
| Experiment (n=30) | Small | 0.08 | 0.61 |
| Experiment (n=30) | Medium | 0.26 | 0.84 |

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

| Scenario | Effect Size | Power | 1-FPRP |
|---|---|---|---|
| Ideal | | 0.80 | 0.94 |
| Experiment (n=30) | Small | 0.08 | 0.61 |
| Experiment (n=30) | Medium | 0.26 | 0.84 |

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)
- Moller & Jennions (2002) and Jennions & Moller (2003): report average effect size and power in ecological and evolutionary studies.

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- $\alpha = 0.05$ is usually fixed.
- assume $\pi = 0.5$: the null and alternate hypothesis are equally likely.
- $1 - \beta =$ power and depends on sample size and effect size.

| Scenario | Effect Size | Power | 1-FPRP |
|---|---|---|---|
| Ideal | | 0.80 | 0.94 |
| Experiment (n=30) | Small | 0.08 | 0.61 |
| Experiment (n=30) | Medium | 0.26 | 0.84 |
| Average | Small | 0.20 | 0.80 |
| Average | Medium | 0.50 | 0.91 |

- Cohen (1988) effect size: small (r=0.1, d=0.2) medium (r=0.3, d=0.5) large (r=0.5, d=0.8)
- Moller & Jennions (2002) and Jennions & Moller (2003): report average effect size and power in ecological and evolutionary studies.

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

| Scenario | Effect Size | $\pi = 0.5$ |
|---|---|---|
| Ideal | | 0.94 |
| Experiment (n=30) | Small | 0.61 |
| Experiment (n=30) | Medium | 0.84 |
| Average | Small | 0.80 |
| Average | Medium | 0.91 |

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

| Scenario | Effect Size | $\pi = 0.5$ | $\pi = 0.1$ |
|---|---|---|---|
| Ideal | | 0.94 | 0.64 |
| Experiment (n=30) | Small | 0.61 | 0.15 |
| Experiment (n=30) | Medium | 0.84 | 0.37 |
| Average | Small | 0.80 | 0.31 |
| Average | Medium | 0.91 | 0.53 |

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

| Scenario | Effect Size | $\pi = 0.5$ | $\pi = 0.1$ | $\pi = 0.01$ |
|---|---|---|---|---|
| Ideal | | 0.94 | 0.64 | 0.14 |
| Experiment (n=30) | Small | 0.61 | 0.15 | 0.02 |
| Experiment (n=30) | Medium | 0.84 | 0.37 | 0.05 |
| Average | Small | 0.80 | 0.31 | 0.04 |
| Average | Medium | 0.91 | 0.53 | 0.09 |

# False Positive Report Probability

$$1 - FPRP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

| Scenario | Effect Size | $\pi = 0.5$ | $\pi = 0.1$ | $\pi = 0.01$ | $\alpha = 0.23$ |
|---|---|---|---|---|---|
| Ideal | | 0.94 | 0.64 | 0.14 | 0.034 |
| Experiment (n=30) | Small | 0.61 | 0.15 | 0.02 | 0.003 |
| Experiment (n=30) | Medium | 0.84 | 0.37 | 0.05 | 0.011 |
| Average | Small | 0.80 | 0.31 | 0.04 | 0.009 |
| Average | Medium | 0.91 | 0.53 | 0.09 | 0.021 |

# How Science Goes Wrong

# How Science Goes Wrong

- **Simmons, JP.** (2011) *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant* Psychological Science 22 1359-1366
- *Researcher degrees of freedom*: how many decisions were made during the course of data collection, analysis, presentation and publication.

# Under-graduate project: Gary Cameron

# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

- 12 top zoology journals.

# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.

# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test

# Under-graduate project: Gary Cameron

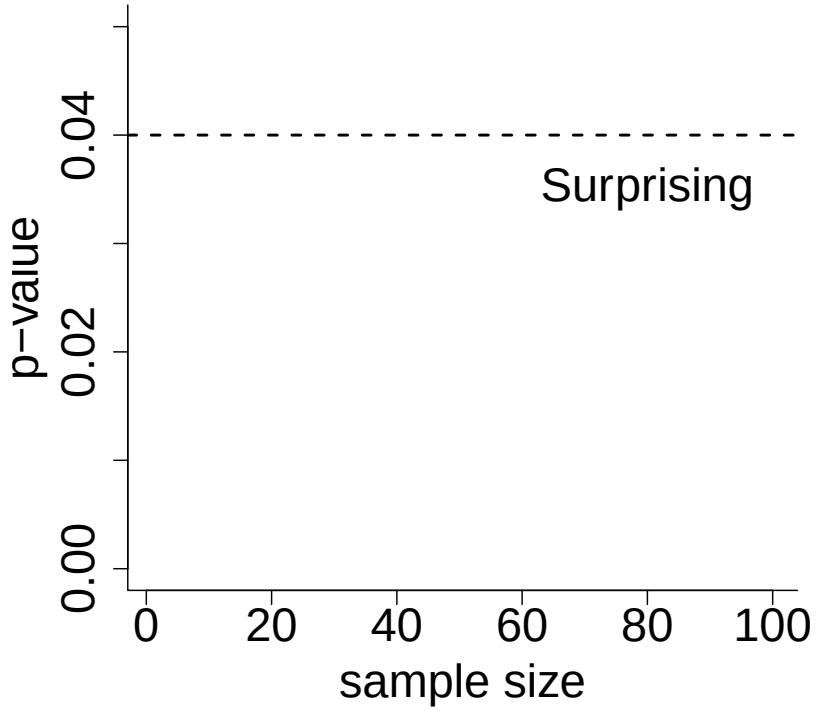
Are qualitative assessments of $\pi$ useful?

- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test

# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

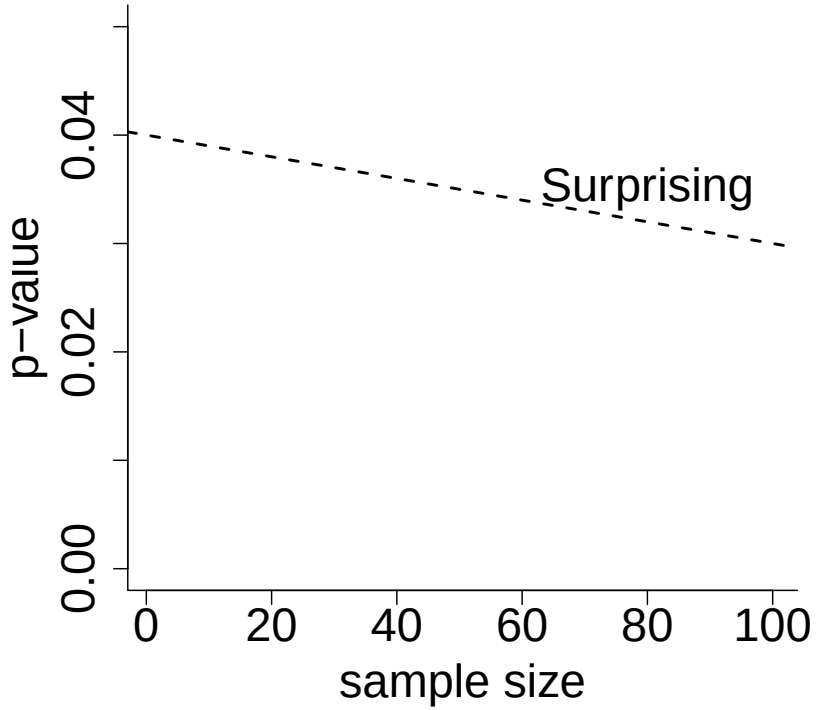
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

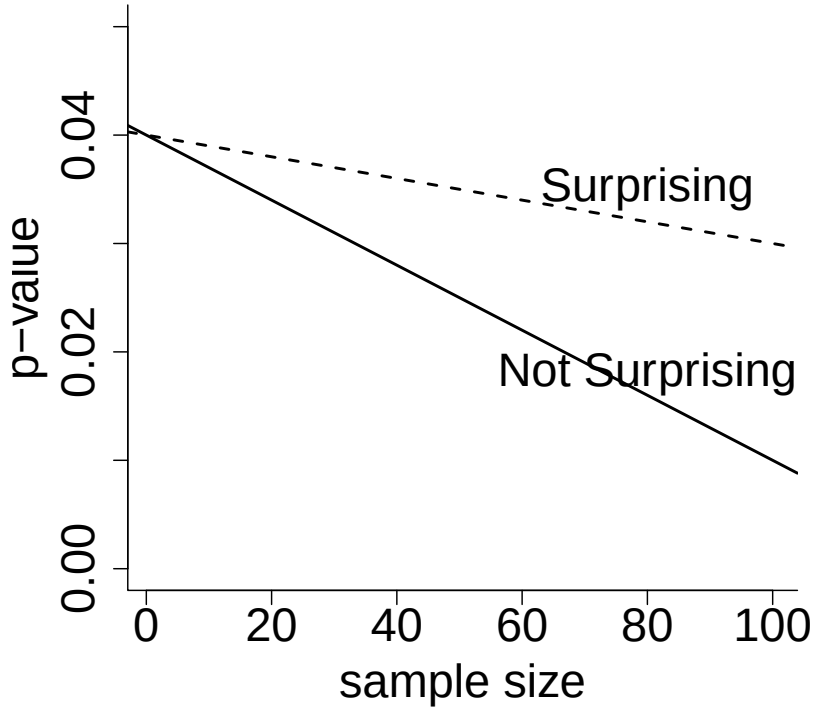
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

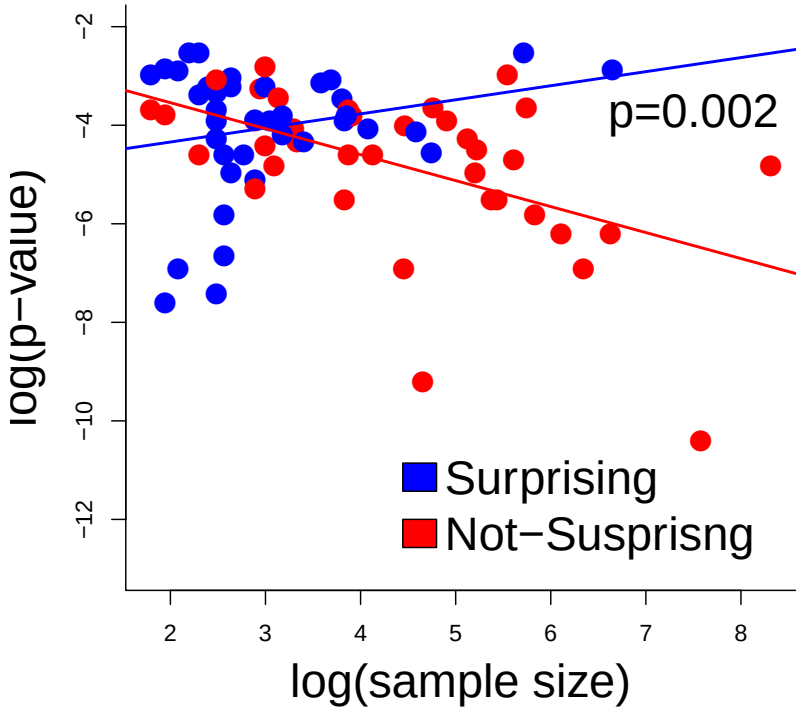
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

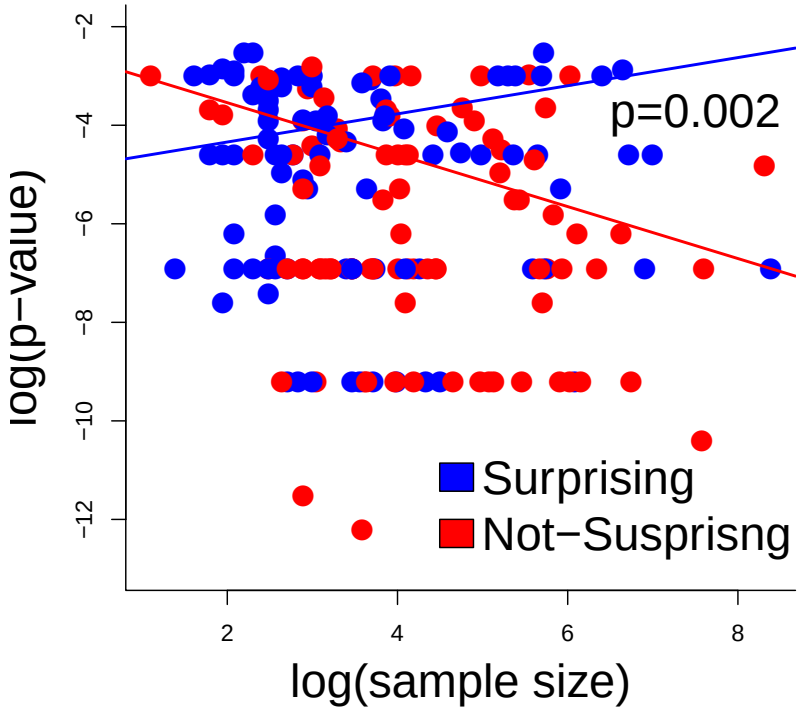
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

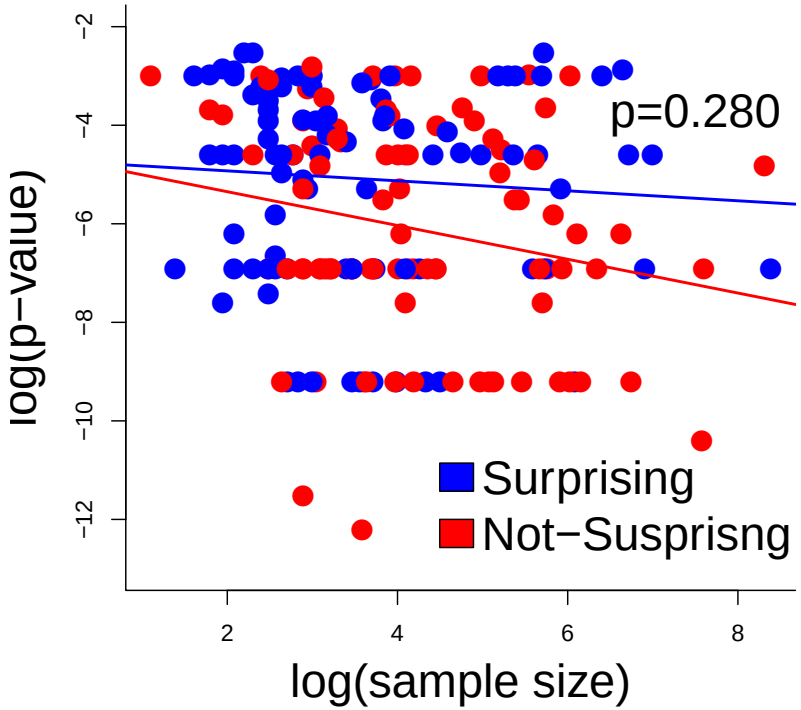
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



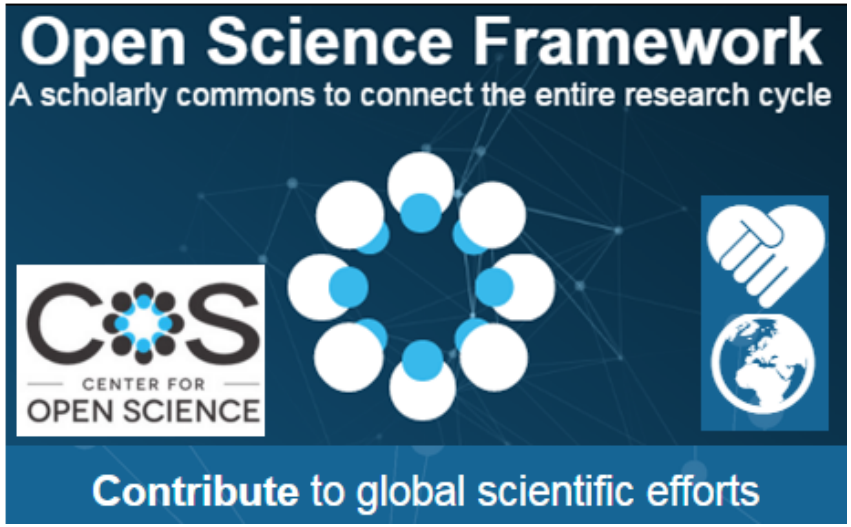
# Under-graduate project: Gary Cameron

Are qualitative assessments of $\pi$ useful?

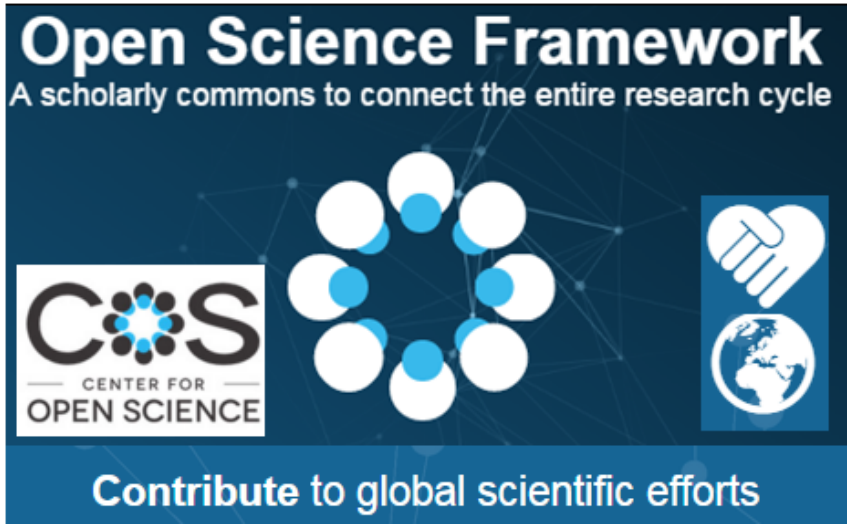
- 12 top zoology journals.
- Find 8 articles in each that self report a surprising result in the abstract.
- Record sample size and p-value for that test
- Go to the next paper in the journal and find an equivalent result.
- Record sample size and p-value for that test



# Solutions


Open Science Framework
A scholarly commons to connect the entire research cycle
CENTER FOR
OPEN SCIENCE
Contribute to global scientific efforts

# Solutions




- Digitally pre-register study hypotheses, data collection and analysis plans as a record of intent.

# Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

# Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

# Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

- Yes - a surprising result is less likely to be true.

# Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible

# Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible
- Not Sure - the expected information content of a true positive $= -log(\pi)$. Surprising results that are true are worth more.

# Unacknowledged problems

- Reducing the Type I error rate also increases the Type II error: what are the relative costs of the two types of error?

Should we bring (qualitative) priors into it?

- Yes - a surprising result is less likely to be true.
- No - fair and precise assessments are not possible
- Not Sure - the expected information content of a true positive $= -log(\pi)$. Surprising results that are true are worth more.