Stats done wrong

2021/05/11

Questionable Research Practices

What are Questionable Research Practices?

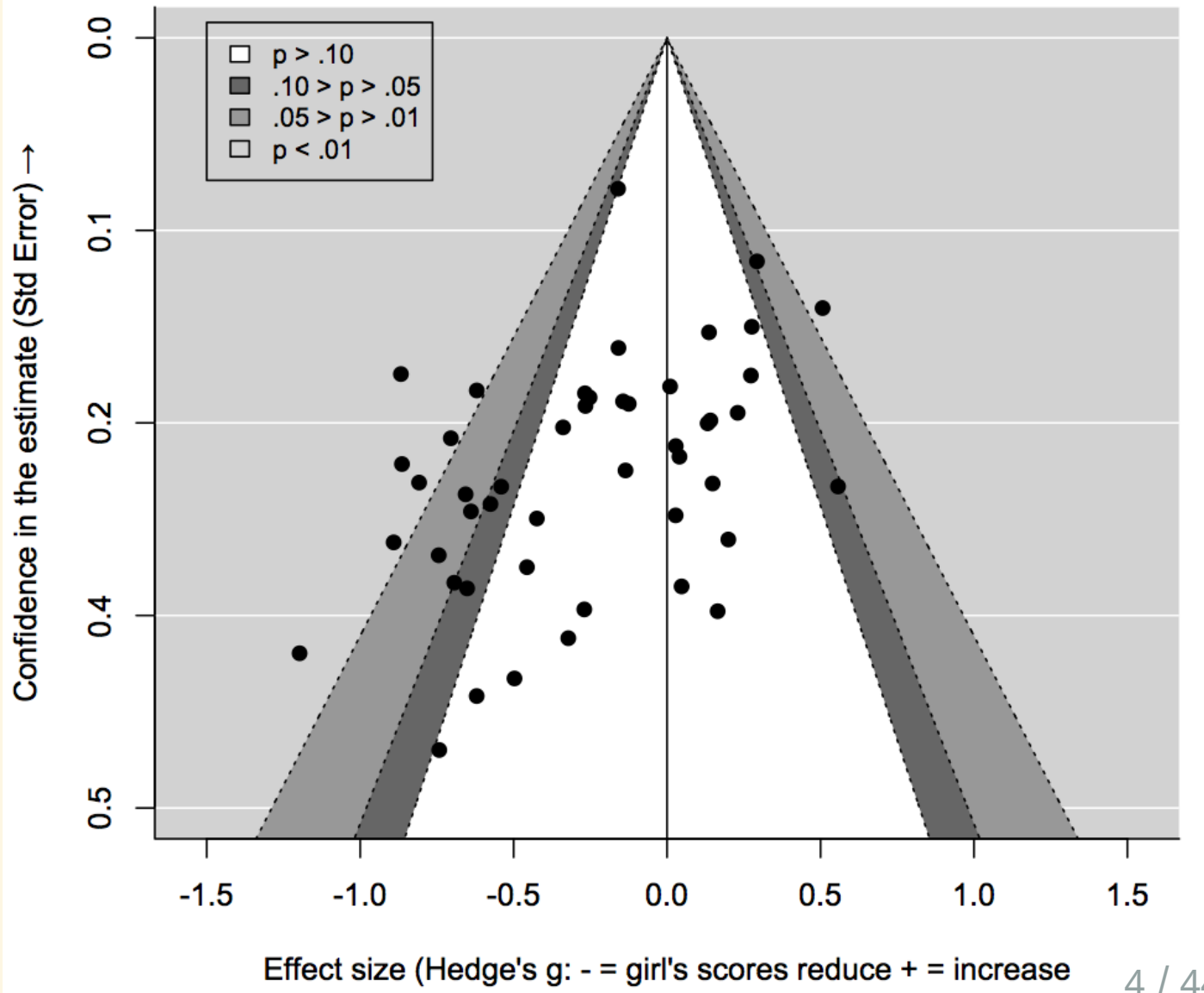
Questionable Research Practices (QRPs) are common, flawed research practices that are not outright fraud but can lead to false positives and a distorted picture of the true pattern of results.

- publication bias (the file drawer problem)
- selective reporting (cherry picking results you *want*)
- selective stopping (stopping when you get the result you want)
- flexible use of outliers
- Hypothesising after the results are known (HARKing)
- and more...

Publication bias

The "file drawer" problem - only significant results tend to be published.

Non-significant results go in the "file drawer".



Selective reporting

Selective reporting is reporting only those outcomes that suit the story you want to tell.

An example:

The US biotech company **InterMune** ran a clinical trial of a new drug for pulmonary fibrosis.

They found no overall effect, but found a small subset of participants with mild-to-moderate for whom mortality was significantly reduced.

The CEO of the company issued a press release reporting only the data from this small subset of participants; a later, larger trial found no benefit for these patients.

(The CEO ended up with a criminal conviction for defrauding the company's investors!)

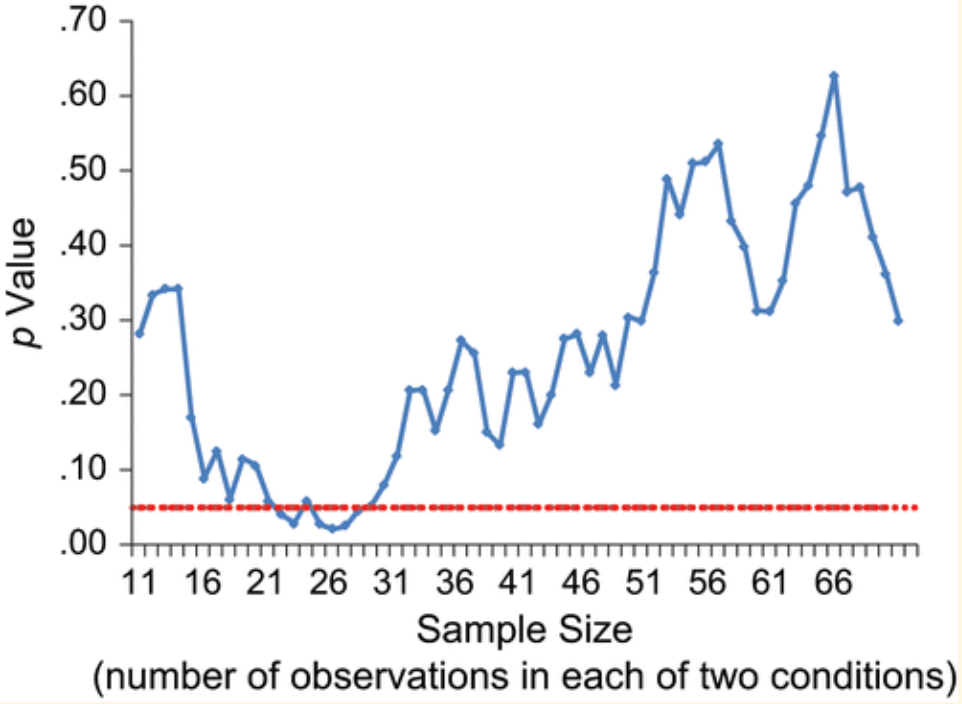
Selective stopping

Selective stopping or "peeking" is when you repeatedly check for significance every few observations.

Once you find a significant result, you stop collecting data.

This can *greatly* increase the rate of false positives.

Here's a fantastic simulation of Selective stopping by Lisa DeBruine of the University of Glasgow.



False positive psychology

Simmons, Nelson, & Simonsohn, 2011 demonstrated how these problems can all come together to produce spurious results.

They ran a study in which participants listened to either a children's song ("Hot potato" by the Wiggles) or a control song ("Kalimba", by Mr Scruff).

Participants reported that they felt older after listening to the children's song than the control song.

So they ran a second study...

False positive psychology

If listening to children's songs made people *feel* younger, can listening to a song about being *older* make people *actually* younger.

In their second study, participants listened to "When I'm Sixty-Four" by the Beatles or the control song. They also provided their birth date and their father's age.

"An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “When I’m Sixty-Four” (adjusted M = 20.1 years) rather than to “Kalimba” (adjusted M = 21.5 years), F(1, 17) = 4.92, p = .040."

False positive psychology

The authors used *every trick in the book* to get this effect. Here's an honest account of the second study:

Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked ~~20~~ 34 **University of Pennsylvania undergraduates to listen** only **to either "When I'm Sixty-Four" by The Beatles or "Kalimba"** or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated** only **their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," **their father's age**, their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. **We used father's age to control for variation in baseline age across participants**.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted *M* = 20.1 years) rather than to "Kalimba" (adjusted *M* = 21.5 years), *F*(1, 17) = 4.92, *p* = .040. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), $F(1, 18) = 1.01$, $p = .33$.

How common are these problems?

A 2012 study of over 2000 US psychologists found that

- 35% said they'd reported an unexpected finding as having been predicted beforehand (HARKing)
- 58% said they'd carried on collecting more data after seeing whether results were significant (optional stopping)
- 67% said they had failed to report all of a study's outcomes (selective reporting)

Countering QRPs

Preregistering designs and protocols

To guard against many of these practices, preregistration is often considered the gold standard.

Clinical trials generally need to publicly preregistered - the outcomes that will be measured are declared in advance.

Note that trials frequently still end up reporting different outcomes - but at least we can see that something suspicious is going on...

More information about clinical trial registration can be found here

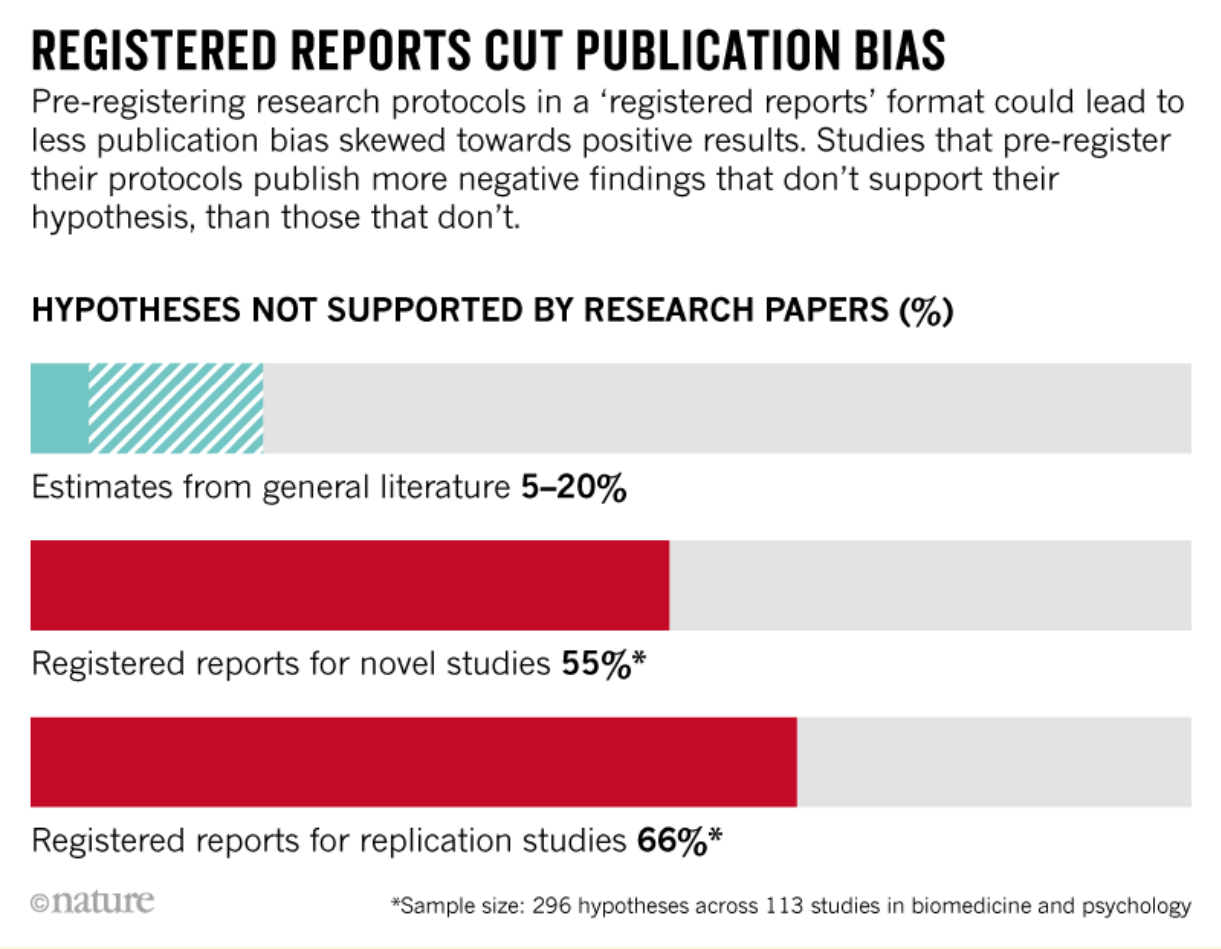
Preregistering designs and protocols

Many journals now offer *Registered Reports* (e.g. Cortex)

In this format, the experimental methods and analysis plans are reviewed before the data is collected.

This increases transparency, allows for feedback to be given before people run the study, and decouples the decision to publish from the significance of the results.

Preregistering designs and protocols



...but preregistration is not a panacea

Preregistration helps to solve some poor statistical practices, such as cherry-picking and outcome-switching, and guards against publication bias.

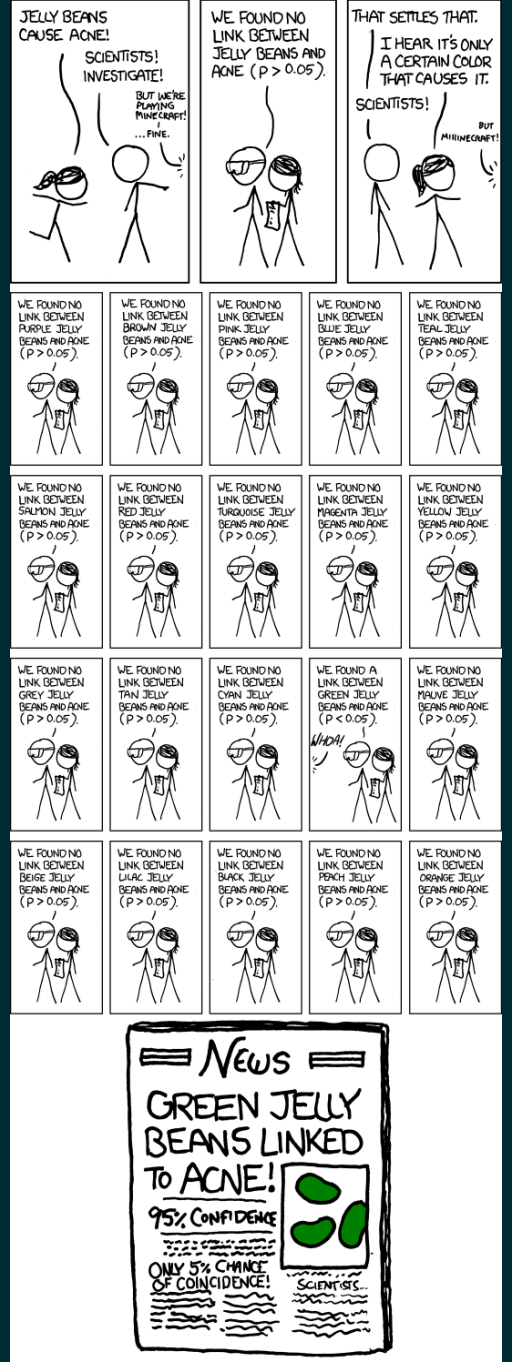
It doesn't necessarily help to generate better hypotheses, to develop better theories, or to ensure use of appropriate statistical methods.

- Is Pregistration Worthwhile? - Szollosi et al (2020)
- The case for formal methodology in scientific reform - Devezer et al., 2021

A plethora of problems

Multiple comparisons

image from Xkcd



Multiple comparisons

An fMRI study (Bennett et al.) examined the neural correlates of perspective taking.

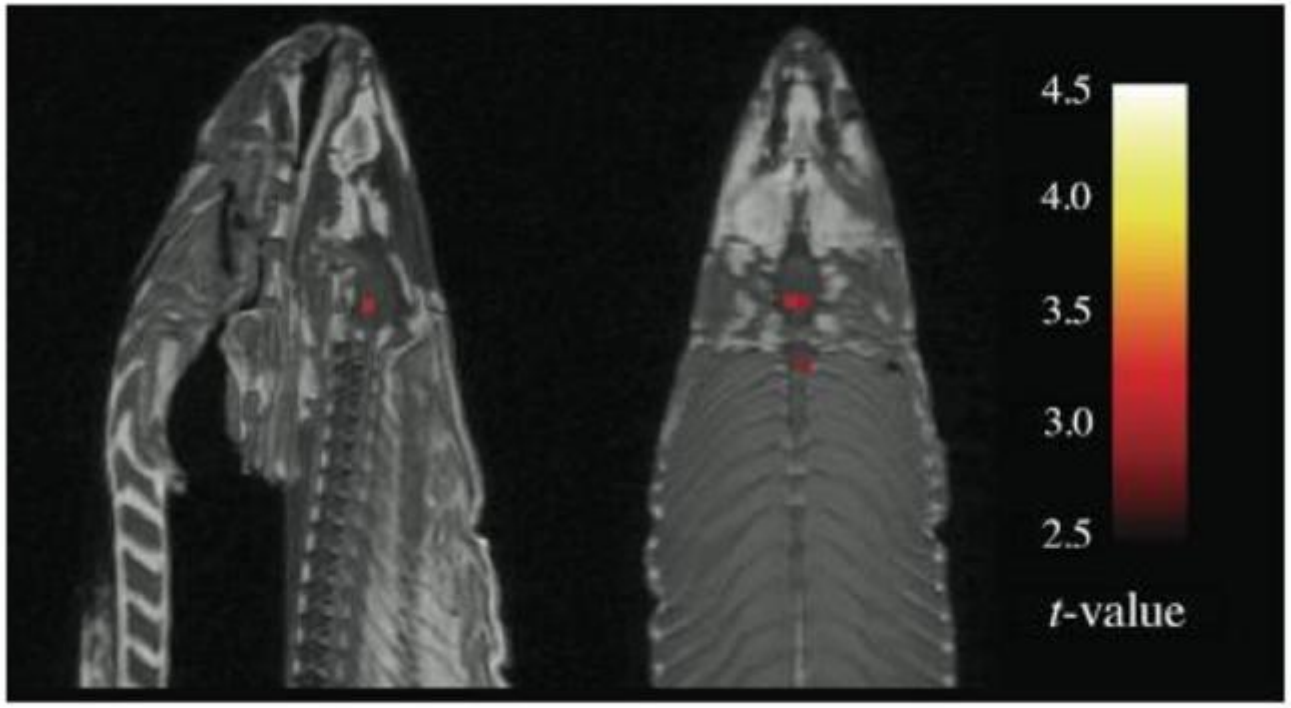
The subject was placed in the scanner and shown photographs of "human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive."

The task was "to determine which emotion the individual in the photo must have been experiencing."

"A t-contrast was used to test for regions with significant BOLD signal change during the presentation of photos as compared to rest. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold."

Multiple comparisons

So where was this cluster?



Multiple comparisons

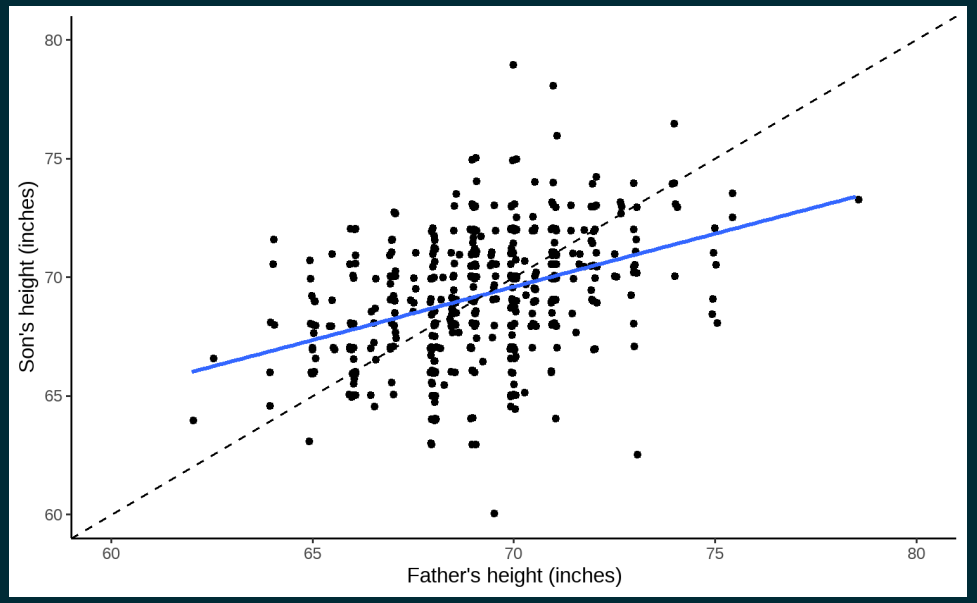
fMRI analyses involve running many, many, many tests at once.

The authors of "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction" deliberately did not use correction for multiple comparisons.

With appropriate corrections, the spurious activity disappeared!

"Statistics that were uncorrected for multiple comparisons showed active voxel clusters in the salmon’s brain cavity and spinal column. Statistics controlling for the familywise error rate (FWER) and false discovery rate (FDR) both indicated that no active voxels were present, even at relaxed statistical thresholds."

Regression to the mean



Scatter plot of heights of 465 fathers and sons.

The diagonal, dashed line on this plot indicates equality between the heights of fathers and sons.

The regression line (blue) is clearly lower for fathers who are taller than average, and higher for fathers who are shorter than average.

Tall fathers have slightly shorter sons; short fathers have slightly taller sons. This is **regression to the mean**.

Base rate fallacy

Base rate fallacy

Imagine we are performing tests for some kind of disease.

We have a test that is 90% sensitive: it correctly detects 90% of true cases.

It has a false positive rate of 5%: it falsely returns a positive result 5% of the time.

We run the test on 10000 people. What is the probability that a positive test is a *true* positive?

Base rate fallacy

To answer the question, we need to know the *base rate*.

If the disease affects 1 in 10 people, we'd expect 1000 true cases in 10000 people.

Out of those 1000 cases, the test would successfully detect **900** cases.

The test has a false positive rate of 5%, so we'd also get **50** false positives.

We would detect 950 cases in total; 900 of those would be true positives.

So the probability of a positive being a *true* positive is 900 / 950: **95%**.

Base rate fallacy

Now suppose that the disease affects 1 in 1000 people.

We'd expect 10 true cases in 10000 people.

Out of those 10, we'd detect 9 cases.

But we'd still get **50** false positives!

So the probability of a positive being a true positive is 9/59:

15%, not **90%**!

Base rate fallacy

Prevalence: 1 in 10

	Infected	Not infected	Total
Test positive	900	50	950
Test negative	100	8950	9050
Total	1000	9000	10000

When prevalence is high, a positive is **very likely** a true positive.

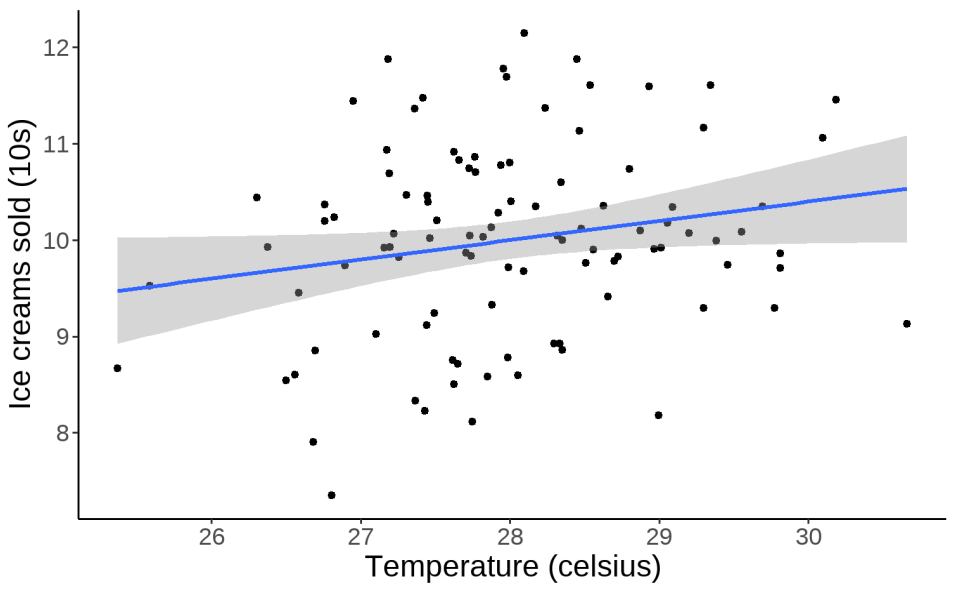
Base rate fallacy

Prevalence: 1 in 1000

	Infected	Not infected	Total
Test positive	9	50	59
Test negative	1	9940	9851
Total	10	9990	10000

When prevalence is low, a positive is **very unlikely** to be a true positive.

Erroneous analysis of interactions

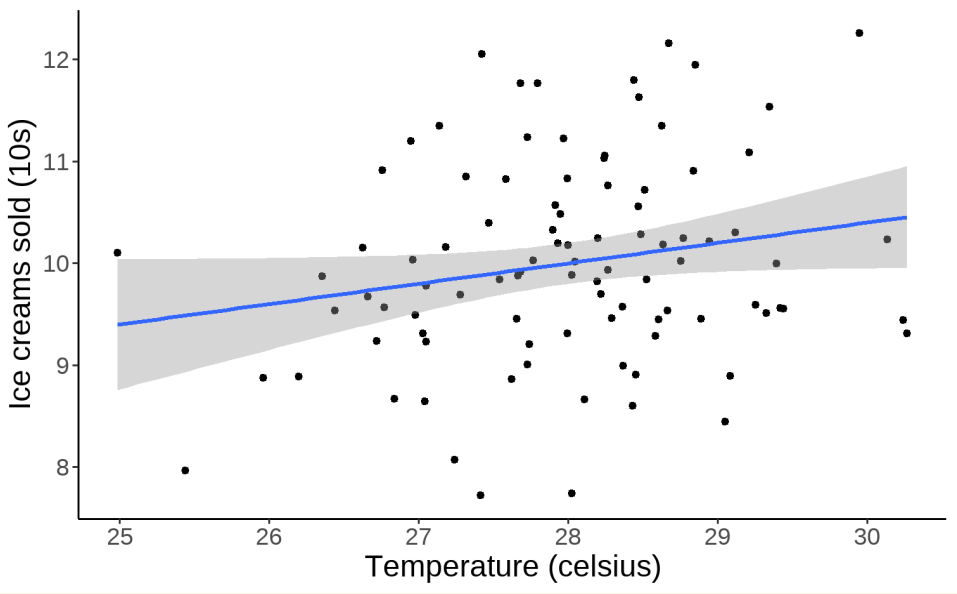


We check the ice cream sales from the MR WHIPPY VAN.

We find that there is a significant correlation between ice cream sales and temperature.

```
correlation::cor_test(icecreams,
                      "Temp", "ic_sales")
```

```
## Parameter1 | Parameter2 |    r |       95% CI | t(99) |      p
## ------------------------------------------------------------
## Temp       |   ic_sales | 0.20 | [0.00, 0.38] |  2.03 | < .05*
##
## Observations: 101
```

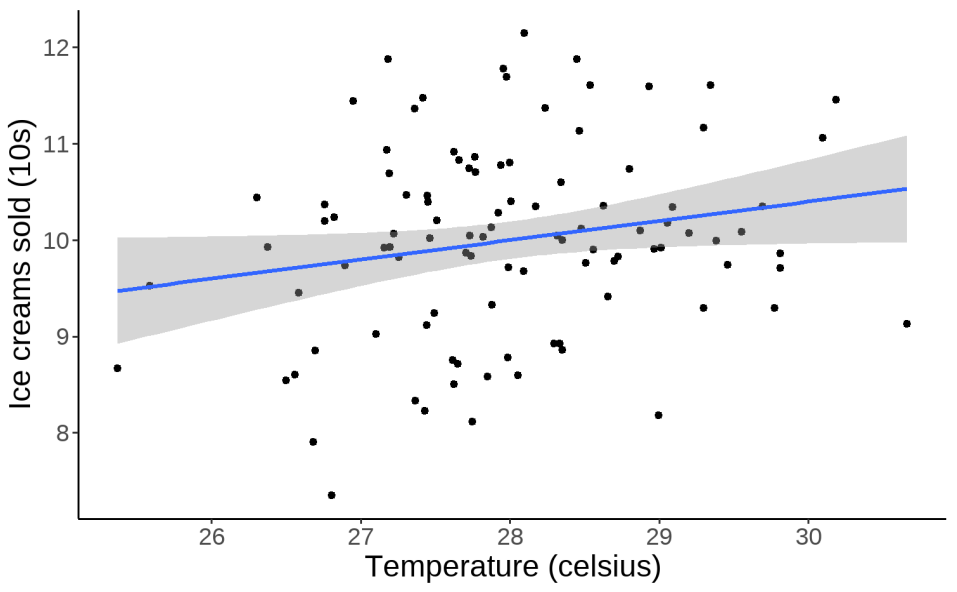


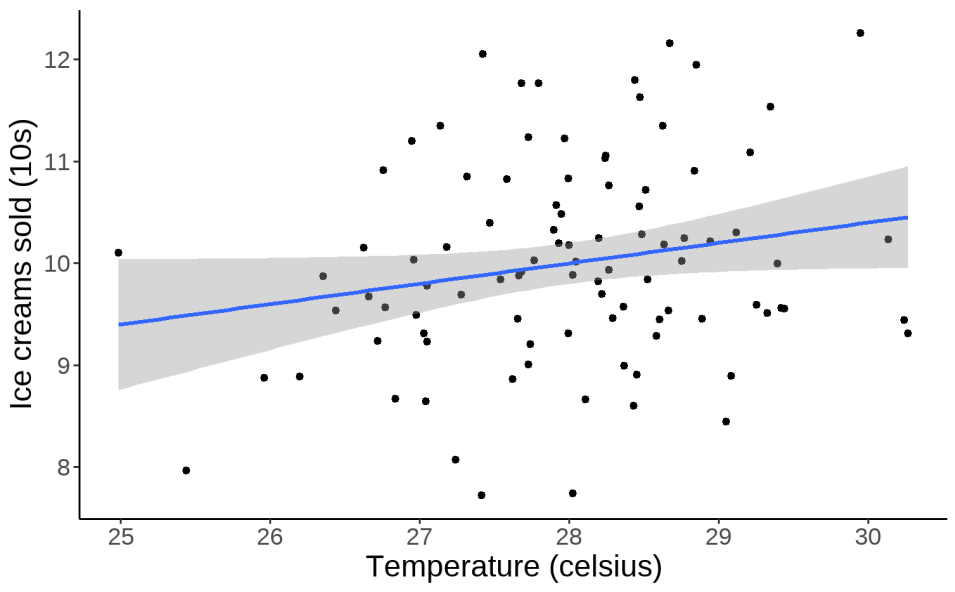
We now check the ice cream sales from MR FROSTY'S VAN.

We find that there is no significant correlation between ice cream sales and temperature.

```
correlation::cor_test(icecreams,
                      "Temp", "ic_sales")
```

```
## Parameter1 | Parameter2 |    r |        95% CI | t(93) |     p
## ------------------------------------------------------------
## Temp       |   ic_sales | 0.20 | [ 0.00, 0.39] |  1.97 | 0.052
##
## Observations: 95
```



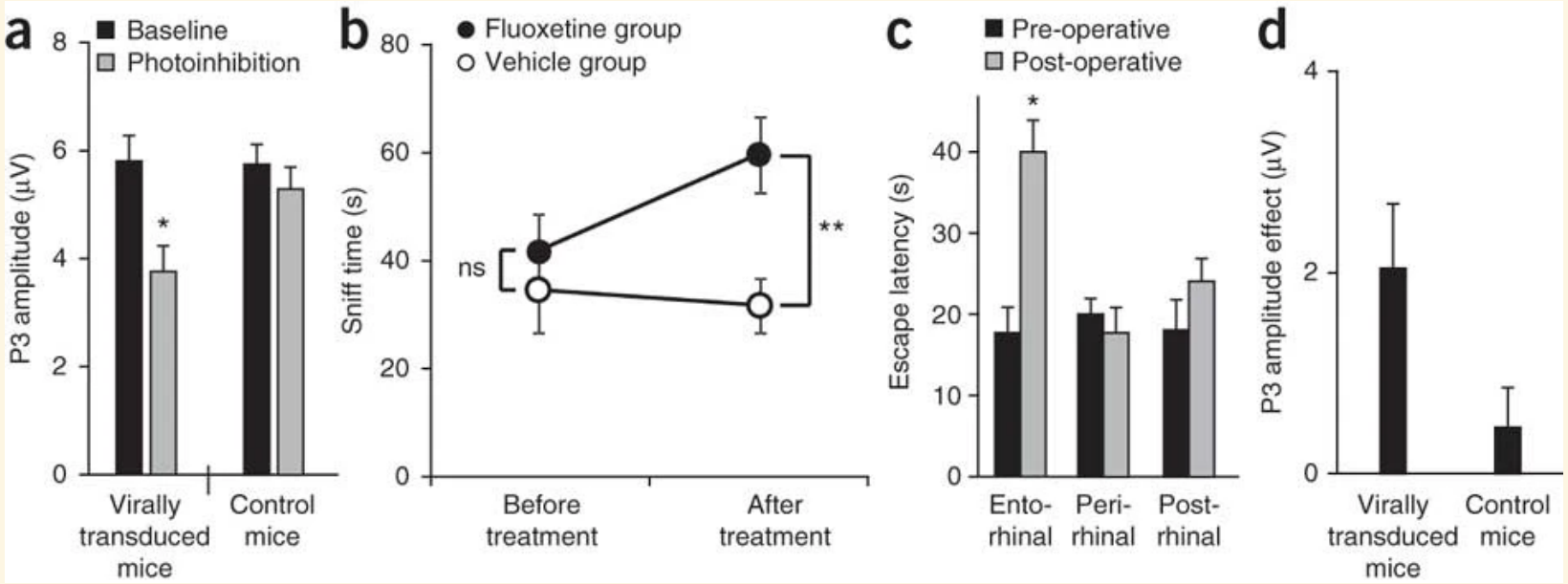


If we directly compare the correlations, there is no significant difference between them!

```
psych::paired.r(.2, .2, n = 101, n2 = 95)
```

```
## Call: psych::paired.r(xy = 0.2, xz = 0.2, n = 101, n2 = 95)
## [1] "test of difference between two independent correlations"
## z = 0  With probability =  1
```

Erroneous analysis of interactions



These are examples of comparisons between groups where an effect is significant and groups where it is not.

It's tempting to say the effect is there in one group but not the other.

Erroneous analysis of interactions

"We reviewed 513 behavioral, systems and cognitive neuroscience articles in five top-ranking journals (Science, Nature, Nature Neuroscience, Neuron and The Journal of Neuroscience) and found that 78 used the correct procedure and 79 used the incorrect procedure. An additional analysis suggests that incorrect analyses of interactions are even more common in cellular and molecular neuroscience."

Erroneous analyses of interactions in neuroscience: a problem of significance

The Difference between "Significant" and "Not Significant" is not Itself Statistically Significant

Selection bias

Selection bias

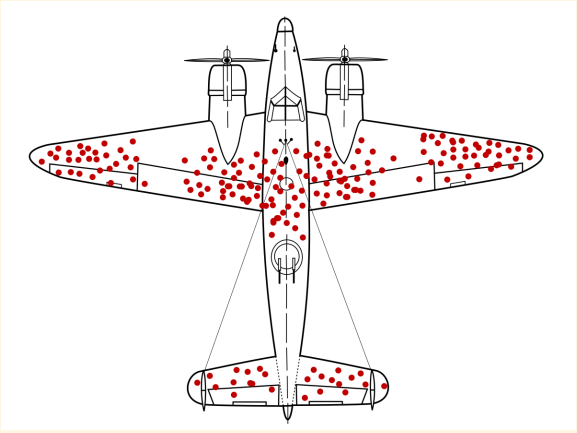
Selection bias is when the participants, groups, or data are selected in such a way as to make them unrepresentative of the population of interest.

Selection bias comes in many forms - for example:

- volunteer bias
- attrition bias
- susceptibility bias

These biases can undermine the validity of the results!

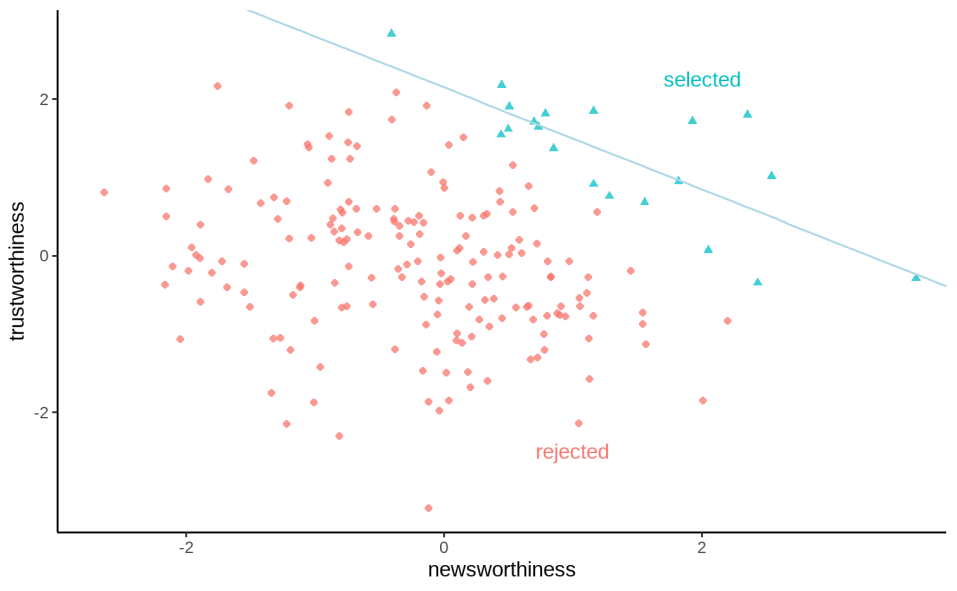
Survival bias



The US Military thought that the best place to add armour was where planes that returned home after missions had been shot the most often.

The statistician Abraham Wald pointed out that the planes that didn't make it back must have been shot in the *other* areas.

Collider bias

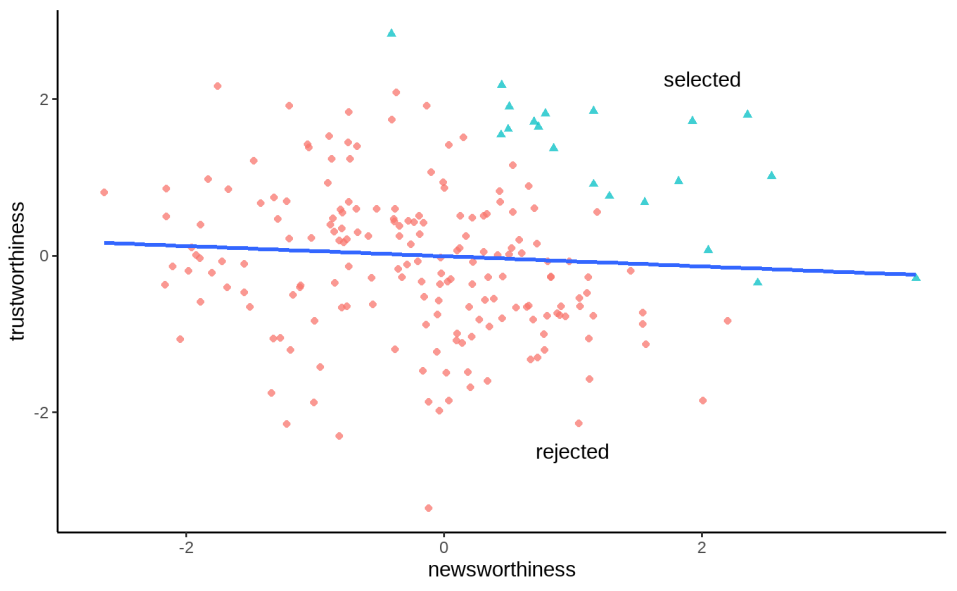


Trustworthiness and newsworthiness both *cause* publication.

The publication process tends to select papers that are either **very trustworthy** or **very newsworthy**.

After selecting a subset, there is a **negative** correlation between trustworthiness and newsworthiness.

Collider bias



When we don't select based on whether and article was published, what do we get?

No correlation between trustworthiness and newsworthiness.

Outright errors

Excel mistakes

There are a number of famous mistakes made when using Excel. An example:

Genes are given symbolic names. e.g. *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase]

Excel, by default, converts those to the *dates* '2-Sep' and '1-Mar' respectively.

Gene name errors are widespread in the scientific literature - Ziemann, Eren, El-Osta, 2016

Reporting mistakes

Nuijten, Hartgerink, van Assen, et al. (2016) looked at the prevalence of simple reporting errors in psychological journals:

"we found that half of all published psychology papers that use NHST contained at least one p-value that was inconsistent with its test statistic and degrees of freedom. One in eight papers contained a grossly inconsistent p-value that may have affected the statistical conclusion"

statcheck.io

Nuijten, Hartgerink, van Assen, et al. The prevalence of statistical reporting errors in psychology (1985–2013), 2016

What to do about all this?

Statistics is HARD.

Mistakes are inevitable.

Try not to fool yourself.

Think carefully about how to handle bias

Make your work transparent!

Additional resources

the100.ci blog

Collider bias: http://www.the100.ci/2017/03/14/that-one-weird-third-variable-problem-nobody-ever-mentions-conditioning-on-a-collider/

Multiverse analysis: http://www.the100.ci/2021/03/07/mulltiverse-analysis/

Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data, Rohrer, J., 2018

Statistical rethinking