Is Science Done According to the Scientific Method?

Should it be?

Adam M. Johansen

a.m.johansen@warwick.ac.uk
http://go.warwick.ac.uk/amjohansen/talks/

MASCDT Lunch Seminar
17th February 2016

Warwick
Statistics

Context: The “Reproducibility Crisis”

Increasingly, published findings cannot be reproduced:

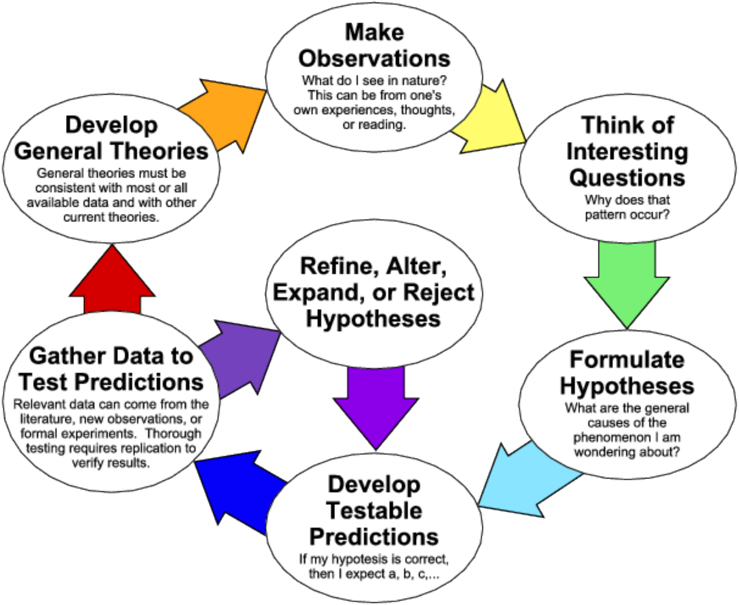
- Ioannidis, John PA. “Why most published research findings are false.” *PLoS Medicine* 2.8 (2005).
- Vul, Edward, et al. “Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition.” *Perspectives on psychological science* 4.3 (2009): 274-290.
- Open Science Collaboration. “Estimating the reproducibility of psychological science.” *Science* 349.6251 (2015).

Superficial mitigation:

- Explicit description of *all* relevant (whatever that means) details.
- Including experimental conditions, data preprocessing and analysis.
- Making data and code available.

are good things, but not a solution.

The Scientific Method as an Ongoing Process


Make Observations
What do I see in nature? This can be from one's own experiences, thoughts, or reading.
Think of Interesting Questions
Why does that pattern occur?
Formulate Hypotheses
What are the general causes of the phenomenon I am wondering about?
Develop Testable Predictions
If my hypotesis is correct, then I expect a, b, c,...
Gather Data to Test Predictions
Relevant data can come from the literature, new observations, or formal experiments. Thorough testing requires replication to verify results.
Develop General Theories
General theories must be consistent with most or all available data and with other current theories.
Refine, Alter, Expand, or Reject Hypotheses

Source: `https://commons.wikimedia.org/wiki/File:The_Scientific_Method_as_an_Ongoing_Process.svg`

Hypothesis Testing

- Given two hypotheses:

  H_0 The universe is as we thought, e.g. $\theta = 0$.

  H_1 The universe more complicated in some specified sense, $e.g. \theta \neq 0$.

- gather some data, $x^\star$,
- and compute some *statistic*, $T(x^\star)$.
- Compute $p = \mathbb{P}(T(X) \text{ at least as extreme as } T(x^\star) | H_0)$.
- If $p < \alpha$ reject H_0 at signifance level α.

Issue 1: Multiple Testing / Selective Reporting

If tests are carried out between H_0 and $H_1, \ldots, H_n$ then the significance level is different to their individual levels.

Multiple Groups Lots of research groups around the world are trying to answer the same questions.

Selective Publication Only *significant* findings are normally published.

Data Mining `https://xkcd.com/882/` Jelly Beans aren't associated with acne; nor are purple, brown, blue, teal, salmon, red, turqoise, magenta, yellow, grey, tan,. . . jelly beans. But green jelly beans are.

Genuine Multiple Testing GWAS, Neuroimaging,. . .

Some partial solutions: registration of studies; publication of *negative* results; controlling false discovery rate (reduces power).

Interpreting p values

Clearly if $p \ll 1$ then there is strong evidence to reject H_0.

- Consider this experiment:

  H_0 $X_1, \ldots, X_n$ are samples from a normal population of mean 0 and variance 1.

  H_1 $X_1, \ldots, X_n$ are samples from a normal population of mean 0.001 and variance 1.

- We observe $T = \frac{1}{4}(X_1, \ldots, X_4) = 7.124$
- $p = \mathbb{P}(T \geq 7.124 | H_0 \text{ true}) = 1.0018 \times 10^{-46}$
- $\mathbb{P}(T \geq 7.124 | H_1 \text{ true}) = 1.0308 \times 10^{-46}$.

- Age of the universe: 4.35×10^{17} s. Perhaps its not so clear.
- We'd need $n \gg 10^6$ to distinguish between H_0 and H_1.
- $p \ll 1$ means: the null model doesn't describe the data; if the alternative doesn't either that's not very informative.

Interpreting p values

Clearly if $p \ll 1$ then there is strong evidence to reject H_0.

- Consider this experiment:

  H_0 $X_1, \dots, X_n$ are samples from a normal population of mean 0 and variance 1.

  H_1 $X_1, \dots, X_n$ are samples from a normal population of mean 0.001 and variance 1.

- We observe $T = \frac{1}{4}(X_1, \dots, X_4) = 7.124$
- $p = \mathbb{P}(T \geq 7.124 | H_0 \text{ true}) = 1.0018 \times 10^{-46}$
- $\mathbb{P}(T \geq 7.124 | H_1 \text{ true}) = 1.0308 \times 10^{-46}$.
- Age of the universe: 4.35×10^{17} s. Perhaps its not so clear.
- We'd need $n \gg 10^6$ to distinguish between H_0 and H_1.
- $p \ll 1$ means: the null model doesn't describe the data; if the alternative doesn't either that's not very informative.

Interpreting p values

Clearly if $p \ll 1$ then there is strong evidence to reject H_0.

- Consider this experiment:

  H_0 $X_1, \ldots, X_n$ are samples from a normal population of mean 0 and variance 1.

  H_1 $X_1, \ldots, X_n$ are samples from a normal population of mean 0.001 and variance 1.

- We observe $T = \frac{1}{4}(X_1, \ldots, X_4) = 7.124$
- $p = \mathbb{P}(T \geq 7.124 | H_0 \text{ true}) = 1.0018 \times 10^{-46}$
- $\mathbb{P}(T \geq 7.124 | H_1 \text{ true}) = 1.0308 \times 10^{-46}$.
- Age of the universe: $4.35 \times 10^{17}\ \mathrm{s}$. Perhaps its not so clear.
- We'd need $n \gg 10^6$ to distinguish between H_0 and H_1.
- $p \ll 1$ means: the null model doesn't describe the data; if the alternative doesn't either that's not very informative.

Some More Context

From a (published psychology) paper considering the 2012 US Presidential Election:

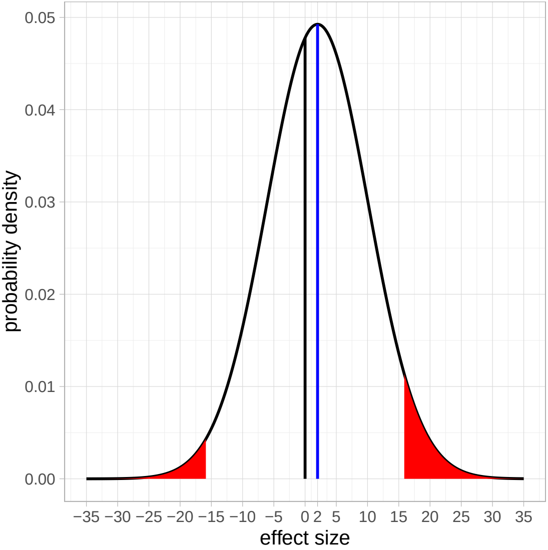
> *Ovulation led single women to become more liberal, less religious and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more conservative, more religious and more likely to vote for Mitt Romney.*

Findings:

- Were statistically significant.
- Had an associated effect size of around 20% (!).
- Were wildly at odds with existing evidence that any such effect could plausibly be at most $\sim 2\%$.

The innocent shall remain anonymous to protect the guilty.

Issue 2: Testing without Power



H_0 T is normally distributed with mean $\mu = 0$ and standard deviation 8.1.

H_1 T is normally distributed with mean $\mu \neq 0$ and standard deviation 8.1.

$P(\text{Reject } H_0 | \mu = 0) = 0.05$

True $\mu = 2$

$P(\text{Reject } H_0 | \mu = 2) \approx 0.06$

$P(T < 0 | \mu = 2, \text{ reject}) \approx 0.24$

$P(|T| > 16 | \mu = 2, \text{ reject}) = 1$

Small studies of small effects...

Consequences of Under-powered tests

If power is small[1] then:

- Significant effects are roughly equally probably under null and alternative hypotheses.
- Effect size is exaggerated whenever it's significant non-zero.
- Effect sign could be wrong even when it's significantly non-zero.

So, if we:

- Carry out *many* independent tests,
- determine which produce *significant* findings,
- report those which are. . . the results are unlikely to be what was intended.

[1] Usually the case for small effects, moderate sample sizes and especially multiple testing.

The Bonferroni principle

- Consider carrying out a large number, m, of tests.
- We expect αm false positives by chance.
- If there are k real effects to find, and the power of each associated test is β_i for $i = 1, \ldots, k$, we'll find about $\beta_1 + \ldots + \beta_k$ of them.
- Of the significant findings, a proportion $\beta_1 + \ldots + \beta_k/(\alpha m + \beta_1 + \ldots + \beta_k)$ of them will be real.
- "If the number of *true* positives is small compared to the expected number of *false* positives, then you won't find anything useful."

Issue 3: The Cox-Jaynes Axioms and Bayesian Inference I

Approximately, if degrees of belief:

1. Can be represented with Real Numbers;
2. show a qualitative corresponds with common sense;
3. and exhibit internal self-consistency,

then:

- they behave as probabilities;
- in light of new data they *must* be updated using Bayes rule;

Naïve combination of p values violates this principle.

Issue 3: The Cox-Jaynes Axioms and Bayesian Inference II

Three formal axioms:

1. Degrees of belief obey a transitive order relationship, i.e.:

$$B(x) > B(y) \text{ and } B(y) > B(z) \Rightarrow B(x) > B(z) \tag{1}$$

   This allows degrees of belief to be mapped to real numbers.

2. There exists a relationship between the degree of belief in any proposition, x and that in its negation, $\neg x$:

$$\exists f \text{ such that } f[B(\neg x)] = B(x) \tag{2}$$

3. The degree of belief in the conjunction of two propositions, x and y is related to the degree of belief in the conditional proposition $x|y$ and the degree of belief in the proposition y:

$$\exists g \text{ such that } B(x, y) = g[B(x|y), B(y)] \tag{3}$$

Bayesian Inference in One Slide

- Data $y_1, \ldots, y_n$ and data model $f(y_1, \ldots, y_n|\theta)$ where θ is some parameter of interest.
- In the frequentist framework θ is a parameter, not a random variable.
- In the Bayesian framework θ is a random variable with prior distribution $f^{\text{prior}}(\theta)$. After observing $y_1, \ldots, y_n$ the posterior density of f is

$$\begin{aligned} f^{\text{post}}(\theta) &= f(\theta|y_1, \ldots, y_n) \\ &= \frac{f^{\text{prior}}(\theta) f(y_1, \ldots, y_n|\theta)}{\int_\Theta f^{\text{prior}}(\vartheta) f(y_1, \ldots, y_n|\vartheta)\, d\vartheta} \\ &\propto f^{\text{prior}}(\theta) f(y_1, \ldots, y_n|\theta) \end{aligned}$$

- But what is f^{prior}? Can it be specified objectively?

Some Questions

- Have many recent studies deviated from the scientific method?
- Or does the method itself need revising to be suitable for modern science?
- How useful are testing and *p*-values obtained from single studies?
- Is *registration of studies* a good idea?
- Was The *Journal of Basic and Applied Social Psychology* right to ban "null hypothesis testing and *p* values" in 2015?
- Might Bayesian approaches be useful in this context?
- How can we reconcile prior distibutions with science?
- Is absolute objectivity possible?
- Is objectivity or transparency more important?
- What's the solution to the *reproducibility crisis*?