Moving to a World beyond p < .05

An introduction to the philosophy of statistics

Ladislas Nalborczyk

LPC, LNC, CNRS, Aix-Marseille Univ

Slides available at https://github.com/lnalborczyk/intro_phil_stat_2022

Program 👋 👋

- Introduction to the philosophy of statistics: Theories, models, evidence, inference
  - Theoretical and statistical models
  - Statistical evidence and inference
- Correct and incorrect interpretations of common hypothesis tests
  - P-values and confidence intervals
  - Bayes factors
  - Problems induced by the mindless use of statistics
- How to move forward: A model comparison (and model criticism) approach
  - Statistical modelling and model comparison
  - A principled Bayesian workflow

Introduction to the philosophy of statistics (why do we need statistics in the first place?)

Scientific theories

A scientific theory can be defined as **a set of logical propositions that posits causal relationships between observable phenomena**.

- Initially broad and abstract: “Every object responds to the force of gravity in the same way”.
- Then, concrete (testable) predictions: “The falling speed of two objects A and B should be the same, all other things being equal”.

These logical propositions are originally abstract and broad (e.g., “every object responds to the force of gravity in the same way”) but lead to concrete and specific predictions that are empirically testable (e.g., “the falling speed of two objects A and B should be the same, all other things being equal”).

Scientific theories

The concept of a scientific theory is not a unitary concept though. As an example, Meehl (1986) lists three kinds of theories:

- **Functional-dynamic theories** which relate “states to states or events to events”. For instance, we say that when one variable changes, certain other variables change in such and such ways.
- **Structural-compositional theories** in which the main idea is to explain what something is composed of, or what kind of parts it has, and how they are put together.
- **Evolutionary theories** which are about the history and/or development of things (e.g., Darwin’s theory, Wegener’s theory of continental drift, the fall of Rome, etc).

First problem: We can not confirm theories

According to Campbell (1990), the (intuitive) logical argument of science has the following form:

- If Newton's theory A is true, then it should be observed that the tides have period B, the path of Mars shape C, the trajectory of a cannonball form D, etc.
- Observation confirms B, C, and D.
- Therefore Newton's theory A is "true".

However, this argument is a fallacious argument known as the **affirmation of the consequent**. The invalidity comes from the existence of the cross-hatched area, that is, other possible explanations for B, C, and D being observed (figure from Campbell, 1990).

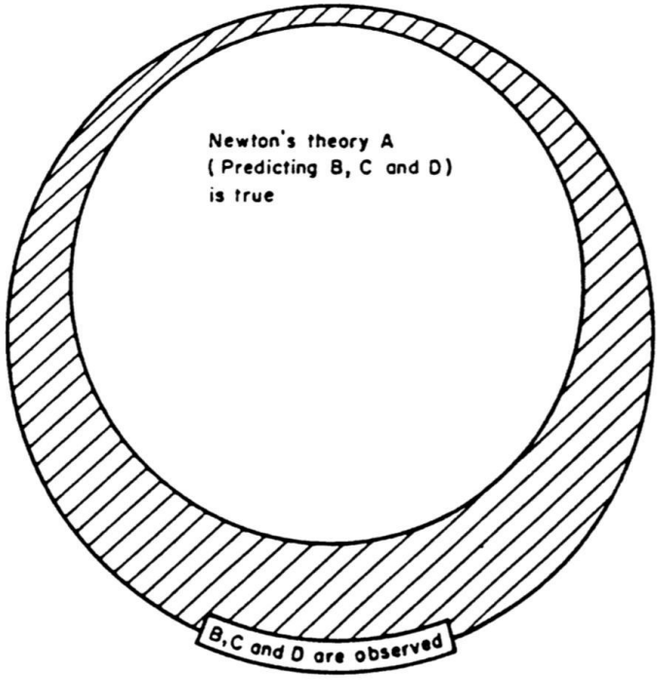


Figure 1. Newton's gravitational theory as an "incomplete induction."

Second problem: We can not (strictly) falsify theories

We can not confirm theories, but maybe we can at least think of a way of disproving them? According to Popper's view, a theory can be considered as falsifiable if it can be shown to be false. But what does it mean for a theory to be false?

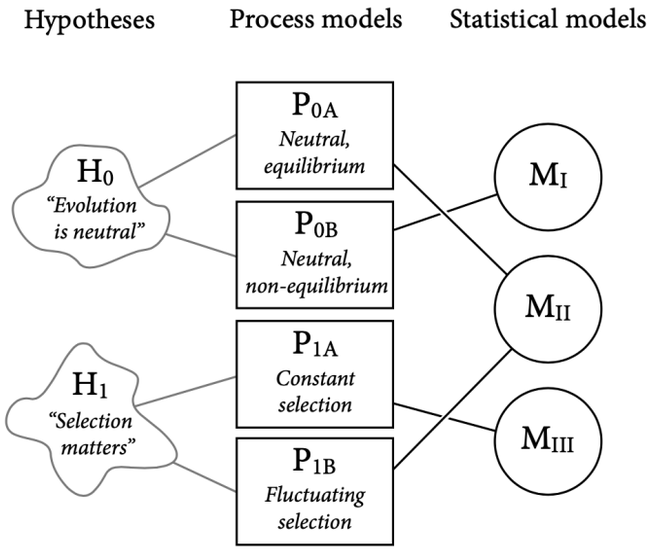
Here we should note that the falsifiability of early Popper concerns the problem of demarcation (i.e., what is science and what is pseudoscience), and defines pseudosciences as composed of non falsifiable theories (i.e., theories that do not allow the possibility of being disproved).

But when it comes to describe how science works (descriptive purposes) or to know how scientific enquiries should be lead (prescriptive purposes), science is usually not described by the falsification standard, as Popper himself recognised and argued. In fact, deductive falsification is impossible in nearly every scientific context (McElreath, 2016).

In the next sections, we discuss some of the reasons that prevent (almost) any scientific theory to be strictly falsified (in a logical sense), namely: i) the distinction between theoretical and statistical models ii) the problem of measurement iii) the problem of continuous hypotheses, and iv) the Duhem-Quine problem.

1) Theoretical and statistical models

A statistical model is a device that connect theories to data. It can be defined as an instantiation of a theory as a set of probabilistic statements (Rouder et al., 2016).



Theoretical models and statistical models are usually not equivalent as many different theoretical models can correspond to the same probabilistic description. Conversely, different probabilistic descriptions can be derived from the same theoretical model. In other words, there is no one-to-one mapping between the two worlds, which render the induction from the statistical model to the theoretical model quite tricky (figure from McElreath, 2020a).

1) Theoretical and statistical inference

Causal and inferential relations between substantive theory, statistical hypothesis, and observational data (figure from Meehl, 1990).

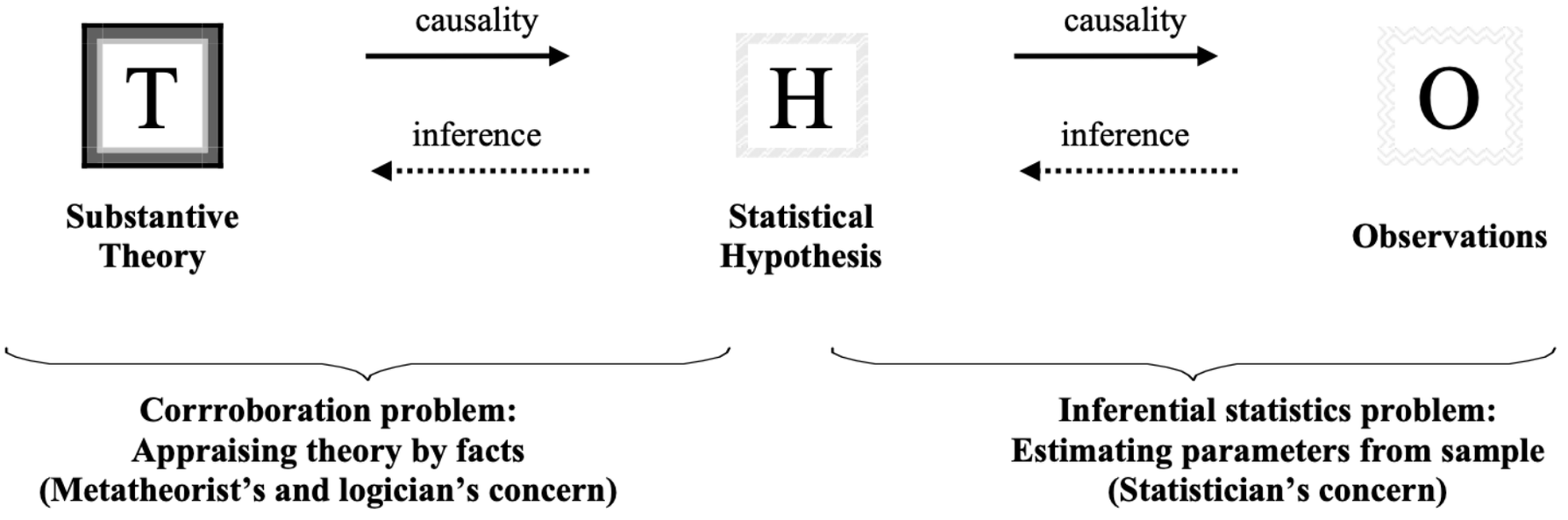


Figure 2. Causal and inferential relations between substantive theory, statistical hypothesis, and observational data.

Another problem yet, as stressed by Paul Meehl, is that while statistical methodology usually deals with the issue of assessing the validity of statistical hypotheses from observations, it does not address, and maybe can not address, the issue of assessing the validity of substantive theories from the corroboration or disconfirmation of statistical hypotheses.

2) Measurement error

The logic of falsification is pretty simple and rests on the power of the modus tollens. This argument (whose exposition, for some reason, usually involves swans) can be presented as follows:

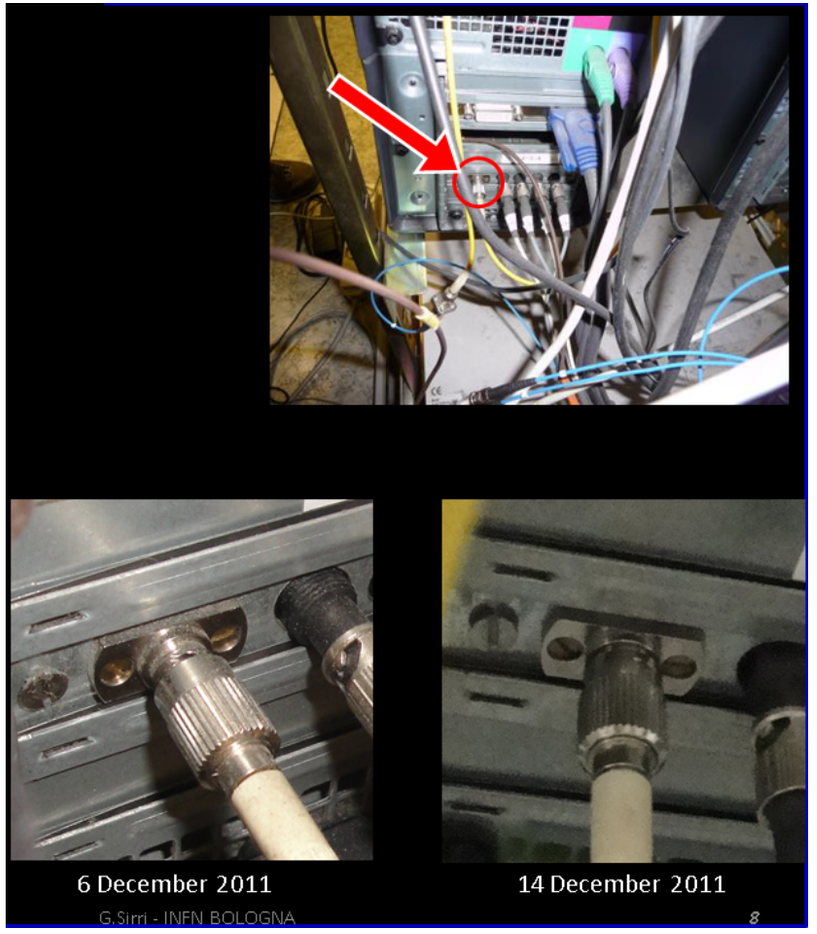
- If my theory T is right, then I should observe these data D
- I observe data that are not those I predicted $\neg D$
- Therefore, my theory is wrong $\neg T$

This argument is perfectly valid and works well for logical statements (statements that are either true or false). However, the first problem that arises when we try to apply this reasoning to the “real world” is the problem of observation error: observations are prone to error, especially at the boundaries of knowledge (McElreath, 2016).

2) Measurement error

According to Einstein, neutrinos can not travel faster than the speed of light. Thus, any observation of faster-than-light neutrinos would act as a strong falsifier of Einstein's special relativity. In 2011 however, a large team of respected physicists announced the detection of faster-than-light neutrinos (cf. the Wikipedia article).

What was the reaction of the scientific community? The dominant reaction was not to claim Einstein's theory to be falsified but was instead: "How did this team mess up the measurement?" (McElreath, 2016). And they were right to suspect something was wrong with the measurement: A fiber optic cable was attached improperly, and a clock oscillator was ticking too fast...



3) Probabilistic hypotheses

Another problem arises from a misapplication of deductive syllogistic reasoning (a misapplication of the modus tollens). The problem (the "permanent illusion," as put by Gigerenzer, 1993) is that most scientific hypotheses are not really of the kind "all swans are white" but rather of the form:

- Ninety percent of swans are white.
- If my hypothesis is correct, we should probably not observe a black swan.

Given this hypothesis, what can we conclude if we observe a black swan? Not much. To understand why, let's translate it first to a more common statement in psychological research (from Cohen, 1994):

- If the null hypothesis is true, then these data are highly unlikely.
- These data have occurred.
- Therefore, the null hypothesis is highly unlikely.

But because of the probabilistic premise (i.e., the "highly unlikely") this conclusion is invalid. Why?

3) Probabilistic hypotheses

Consider the following argument (Cohen, 1994; Pollard & Richardson, 1987a):

- If a person is an American, he is probably not a member of Congress.
- This person is a member of Congress.
- Therefore, he is probably not an American.

This conclusion is not sensible (the argument is invalid), because it fails to consider the alternative to the premise, which is that if this person were not an American, the probability of being a member of Congress would be 0.

This is formally exactly the same as:

- If the null hypothesis is true, then these data are highly unlikely.
- These data have occurred.
- Therefore, the null hypothesis is highly unlikely.

Which is as much invalid as the previous argument, because i) the premise (the hypothesis) is probabilistic/continuous rather than discrete/logical and ii) because it fails to consider the probability of the alternative. Thus, even without measurement/observation error, this problem would prevent us from applying the modus tollens to our hypothesis, thus preventing any possibility of strict falsification.

4) The underdetermination problem

Again another problem is known as the Duhem–Quine thesis/problem (aka the **underdetermination problem**). In practice, when a substantive theory T happens to be tested, some hidden assumptions, such as auxiliary theories about the instruments we use, are also put under examination (Meehl, 1990, 1978, 1997).

When we test a theory predicting that "if O_1" (some manipulation), "then O_2" (some observation), what we actually mean is that we should observe this relation, **if and only if** all of the above (i.e., the auxiliary theories, the instrument theories, the particulars, etc) are true.

4) The underdetermination problem

Thus, the logical structure of an empirical test of a theory T can be described as the following conceptual formula (Meehl, 1990, 1978, 1997):

$$(T \wedge A_t \wedge C_p \wedge A_i \wedge C_n) \rightarrow (O_1 \supset O_2)$$

where the $\wedge$ are conjunctions ("and"), the arrow $\rightarrow$ denotes deduction ("follows that ..."), and the horseshoe $\supset$ is the material conditional ("If O_1, Then O_2"). A_t is a conjunction of auxiliary theories, Cp is a "ceteribus paribus" clause (i.e., we assume there is no other factor exerting an appreciable influence that could obfuscate the main effect of interest), A_i is an auxiliary theory regarding instruments, and C_n is a statement about experimentally realised conditions (i.e., we assume that there is no systematic error/noise in the experimental settings).

4) The underdetermination problem

$$(T \wedge A_t \wedge C_p \wedge A_i \wedge C_n) \rightarrow (O_1 \supset O_2)$$

However, although the modus tollens is a valid figure of the implicative syllogism for logical statements (e.g., “all swans are black”), the neatness of Popper’s classic falsifiability concept is fuzzed up by the acknowledgement of the actual form of an empirical test. Obtaining falsificative evidence during an empirical test does not only falsify the substantive theory T, but it does falsify all the left-side of the above statement. In other words, what we have achieved by our laboratory or correlational “falsification” is a falsification of the combined claims $T \wedge A_t \wedge C_p \wedge A_i \wedge C_n$, which is probably not what we had in mind when we did the experiment (Meehl, 1990).

To sum up, failing to observe a predicted outcome does not necessarily mean that the theory itself is wrong, but rather that the conjunction of the theory and the underlying assumptions at hand are invalid (Lakatos, 1976; Meehl, 1978, 1997).

Consequences

Falsification in science is almost always consensual, not logical (McElreath, 2020b). A theoretical claim is considered to be falsified only when multiple lines of converging evidence have been obtained, by independent teams of researchers, and usually after several years or decades of critical discussion. The "falsification of a theory" then appears as a social result, issued from the community of scientists, and (almost) never as a deductive falsification.

How can we accumulate **evidence** in favour of or against a theory? That's were statistics comes into play. There are several philosophical frameworks for statistical inference, which differ by their assumptions and by their definition of what counts as evidence in favour or against a theory.

Correct and incorrect interpretations of common hypothesis tests: p-values and confidence intervals

Null Hypothesis Significance Testing (NHST)

Let's say we are interested in height differences between women and men...

```r
set.seed(19) # to get reproducible results
men <- rnorm(n = 100, mean = 175, sd = 10) # 100 men heights
women <- rnorm(n = 100, mean = 170, sd = 10) # 100 women heights
```

```r
t.test(x = men, y = women)
```

```
    Welch Two Sample t-test

data:  men and women
t = 2.4969, df = 197.98, p-value = 0.01335
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7658541 6.5209105
sample estimates:
mean of x mean of y
 175.1258  171.4825
```

None of these definitions is true... 🥺🥺

Null Hypothesis Significance Testing (NHST)

We are going to simulate t-values computed on samples generated under the assumption of no difference between women and men (the null hypothesis $\mathcal{H}_0$).

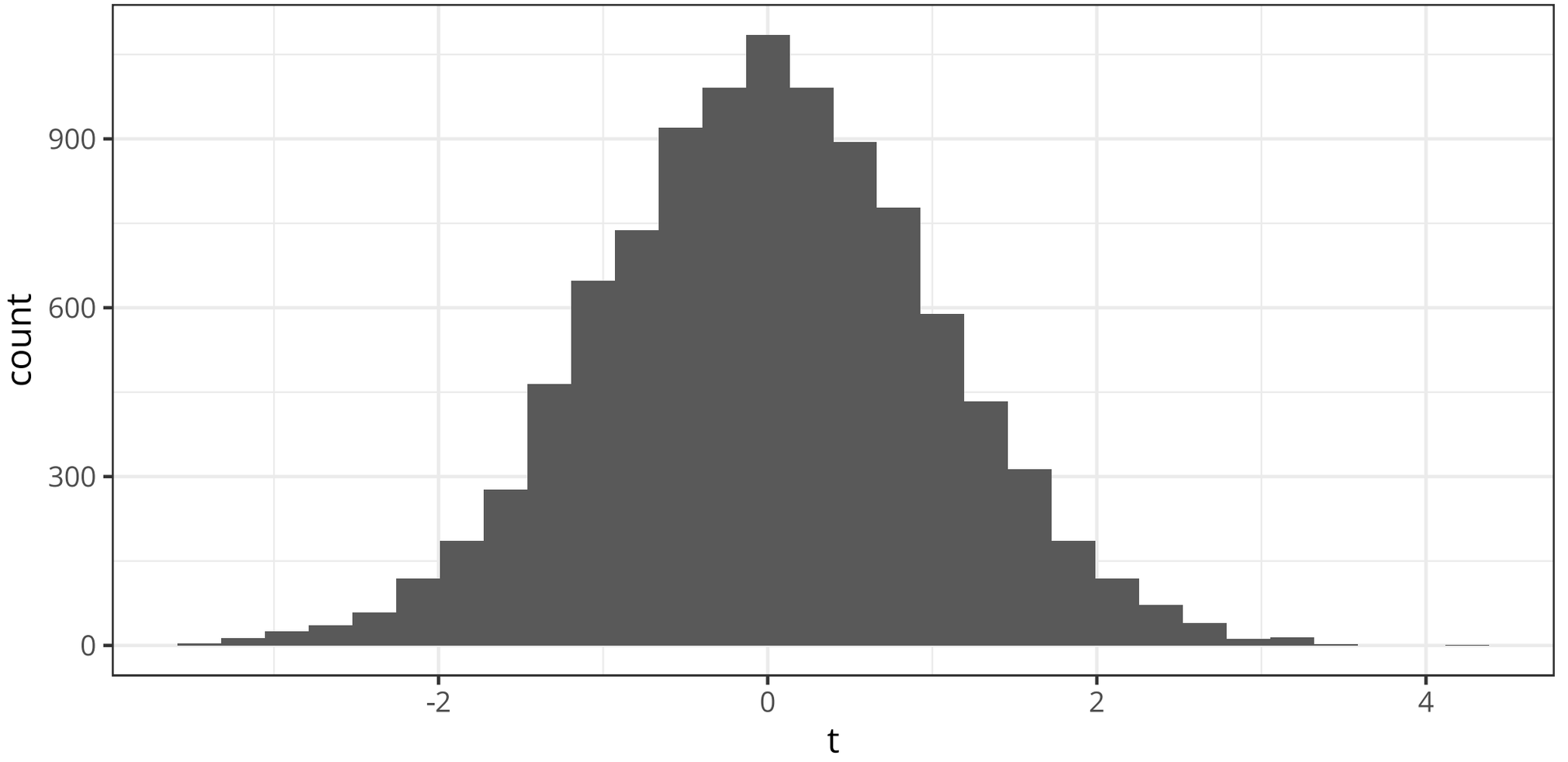
```r
nsims <- 1e4 # number of simulations
t <- rep(x = NA, times = nsims) # initialising an empty vector

for (i in 1:nsims) {

    men2 <- rnorm(n = 100, mean = 170, sd = 10)
    women2 <- rnorm(n = 100, mean = 170, sd = 10)
    t[i] <- t.test(x = men2, y = women2)$statistic

}
```

Or without for loops.

```r
t <- replicate(n = nsims, expr = t.test(x = rnorm(100, 170, 10), y = rnorm(100, 170, 10) )$statistic)
```

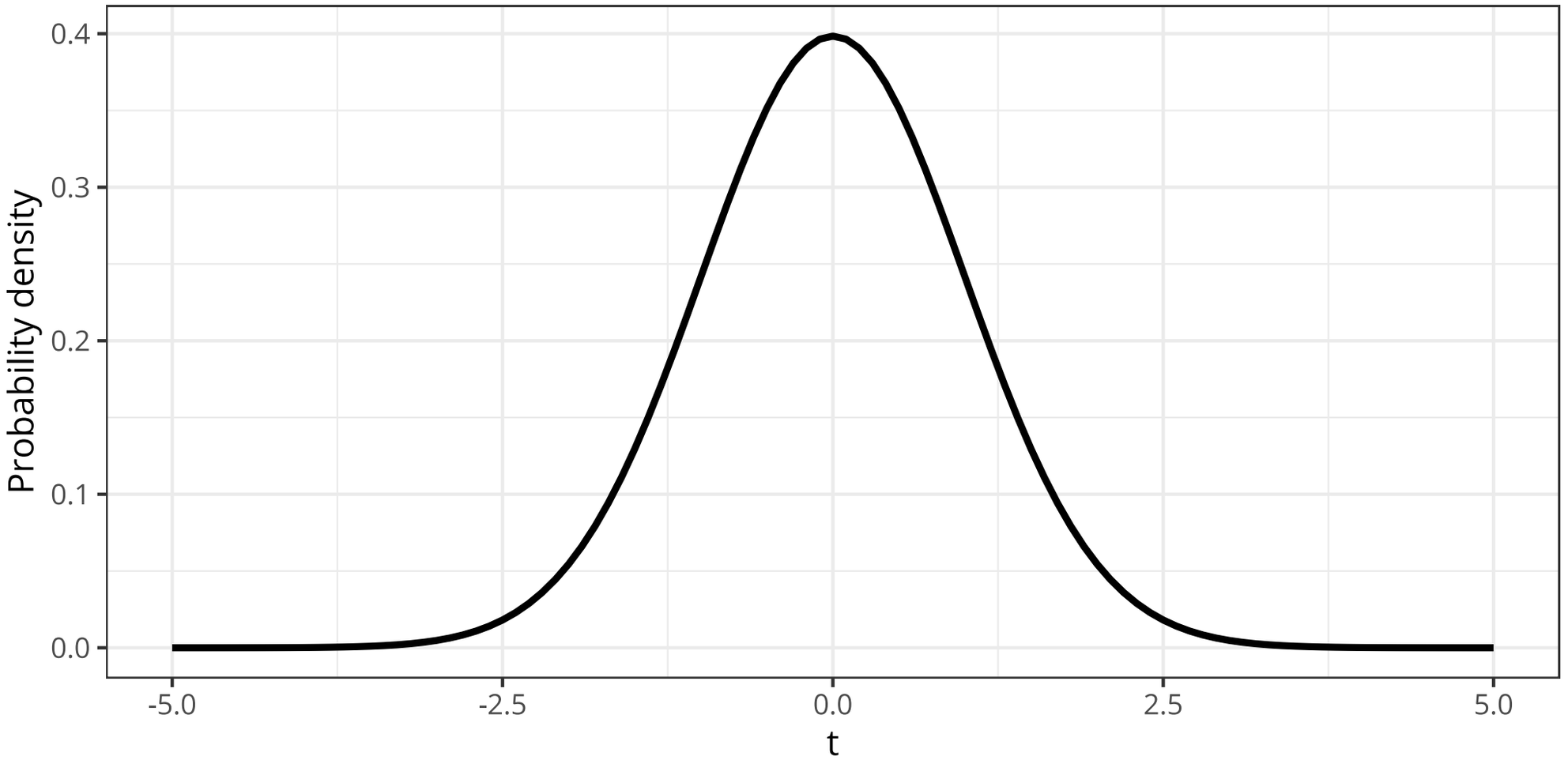
Null Hypothesis Significance Testing (NHST)

```
1  data.frame(t = t) %>%
2      ggplot(aes(x = t) ) +
3      geom_histogram()
```



Null Hypothesis Significance Testing (NHST)

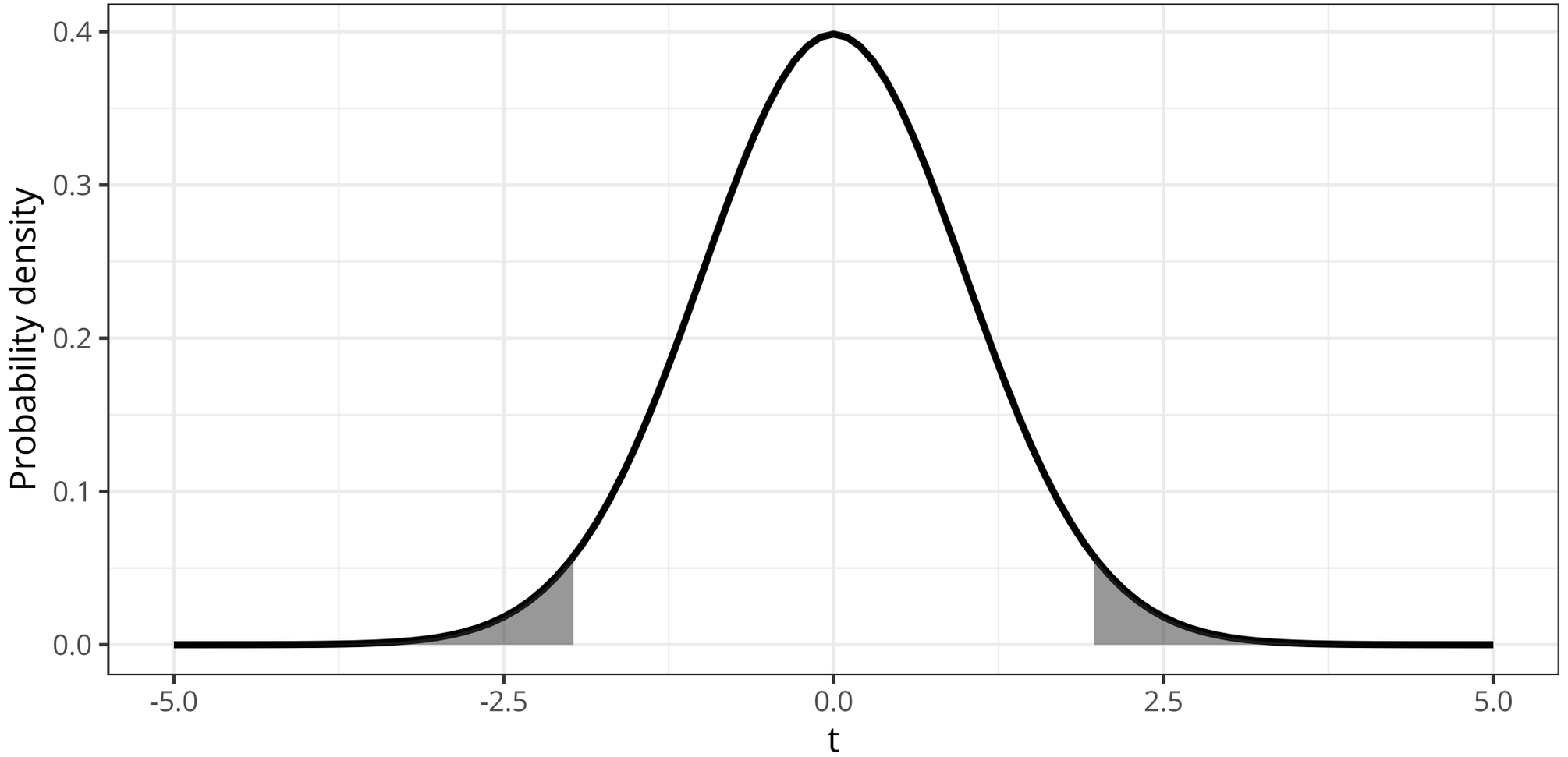
```r
data.frame(t = c(-5, 5) ) %>%
    ggplot(aes(x = t) ) +
    stat_function(fun = dt, args = list(df = t.test(men, women)$parameter), size = 1.5) +
    ylab("Probability density")
```



Null Hypothesis Significance Testing (NHST)

```r
alpha <- .05 # significance threshold (alpha)
abs(qt(alpha / 2, df = t.test(x = men, y = women)$parameter) ) # two-sided critical t-value
```

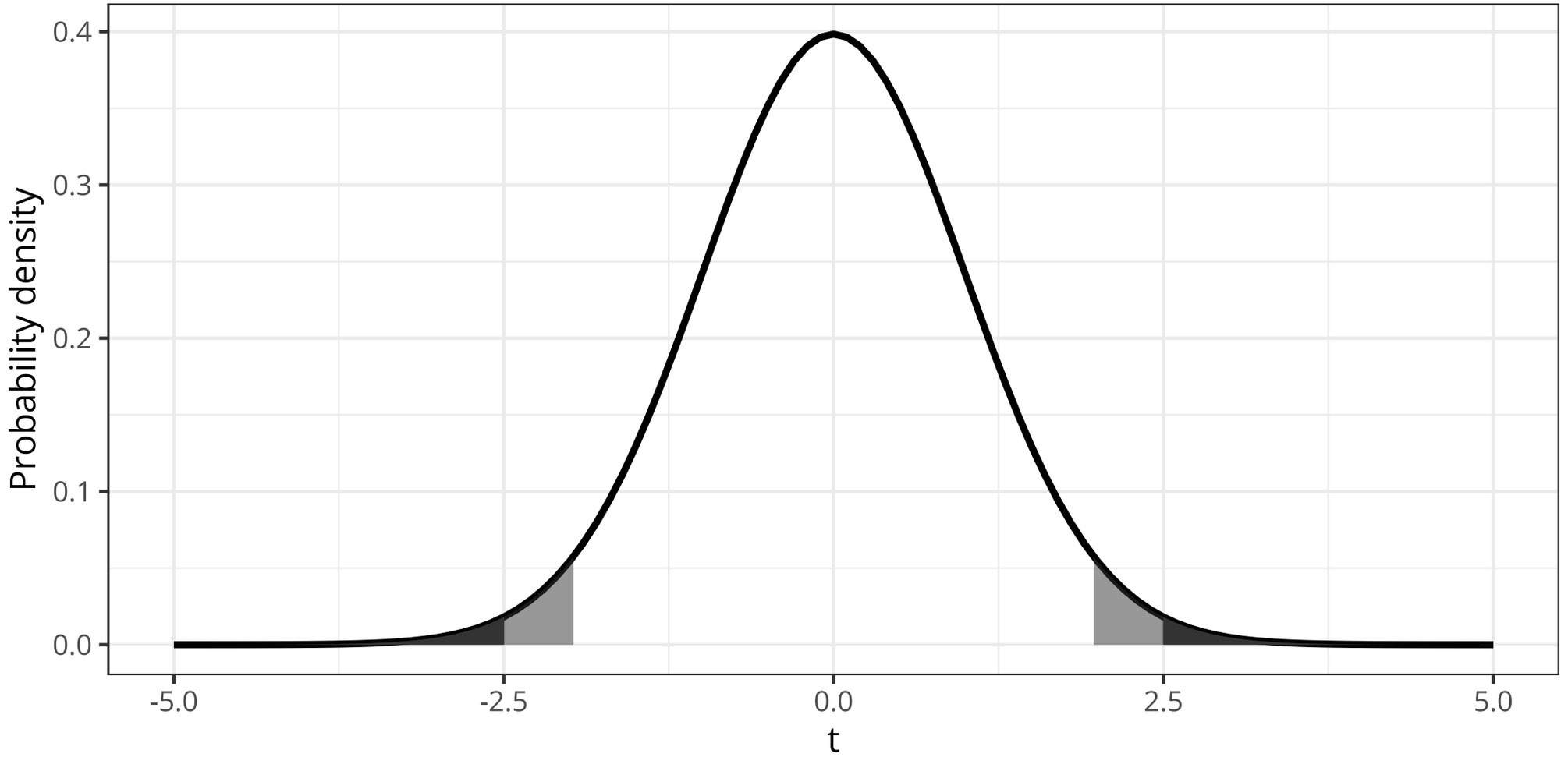
```
[1] 1.972019
```



Null Hypothesis Significance Testing (NHST)

```r
tobs <- t.test(x = men, y = women)$statistic # observed t-value
tobs %>% as.numeric
```

```
[1] 2.496871
```



P-values

A p-value is simply a tail area (an integral) computed from the distribution of test statistics under (given) the null hypothesis. It gives the probability of observing the data we observed **or more extreme data, given that the null hypothesis is true** (Wagenmakers, 2007).

$$p[\mathbf{t}(\mathbf{x}^{\text{rep}}; \mathcal{H}_0) \geq t(x)]$$

```r
t.test(x = men, y = women)$p.value
```

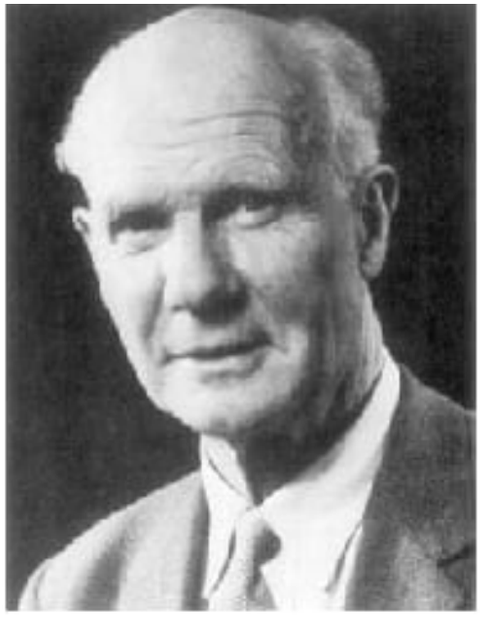
```
[1] 0.01334509
```

```r
tvalue <- abs(t.test(x = men, y = women)$statistic)
df <- t.test(x = men, y = women)$parameter
2 * integrate(f = dt, lower = tvalue, upper = Inf, df = df)$value
```

```
[1] 0.01334509
```

Fisher versus Neyman & Pearson





According to Fisher, the p-value is thought to measure the strength of evidence against the null hypothesis: the lower the p-value, the stronger the evidence against the null hypothesis. But we know that p-values at best **correlate** (in a loose meaning) with evidence (e.g., see Wagenmakers, 2007).

The Fisherian continuous interpretation of p-values has many problems (cf. next slide) and has been widely criticised.

Neyman & Pearson used p-values and significance thresholds as a way of **controlling error rates in the long run**. In this perspective, we don't interpret the p-value, we only "classify" results as **significant** or **non-significant**. This strict procedure allows keeping error rates at a fixed level (given that the null hypothesis is true, see this blogpost). However, this view also has serious problems. One of the biggest problem being **the domain problem** (e.g., Trafimow & Earp, 2017).

Logic, frequentism, and probabilistic reasoning

The modus tollens is one of the strongest rule of inference in logic. It works perfectly well in science when we deal with hypotheses of the following form: "*If* $\mathcal{H}_0$ is true, then we should not observe x. We observed x. Then, $\mathcal{H}_0$ is false".

BUT, most of the time, we deal with continuous, **probabilistic** hypotheses...

The Fisherian inference (induction) is of the form: "*If* $\mathcal{H}_0$ is true, then we should PROBABLY not observe x. We observed x. Then, $\mathcal{H}_0$ is PROBABLY false".

However, as we have seen previously, this argument is invalid. The modus tollens does not apply to probabilistic statements (e.g., Pollard & Richardson, 1987b).

Interpreting confidence intervals

Confidence intervals are basically regions of significance. Thus, they have to be interpreted as cautiously as p-values, and are submitted to the same flaws.

A 95% confidence interval **does not mean** that there is a 95% probability that the interval contains the population value of the parameter (remember the modus tollens fallacy).

The only correct interpretation is to think about it in terms of **coverage proportion** (see next slide and this blogpost).

A 95% confidence interval represents **a statement about the procedure**, not about the parameter. It means that, in the long run, 95% of the confidence intervals we could compute (in an exact replication of the experiment) would contain the population value of the parameter. But we can not say anything about the particular confidence interval we computed in this particular experiment...

Preliminary summary (statistics is hard)

Frequentist statistics (e.g., p-values and confidence intervals) make sense under the frequentist interpretation of probability: they refer to **long-run frequencies**.

P-values are simply tail areas in probability distributions. It means that they are conditional on some distribution. But it also means that computing a p-value is a generic statistical procedure, it's not inextricable from the null hypothesis (e.g., see Bayesian p-values).

Confidence intervals are basically regions of significance. Thus, they are prone to the very same limits as p-values.

Correct and incorrect interpretations of common hypothesis tests: Bayes factors

Bayes factors

Instead of testing only one hypothesis (the null hypothesis), Bayes factors allow comparing two hypotheses. For instance, let's say we are comparing two models:

- $\mathcal{H}_0 : \mu_1 = \mu_2 \rightarrow \delta = 0$
- $\mathcal{H}_1 : \mu_1 \neq \mu_2 \rightarrow \delta \neq 0$

$$\underbrace{\frac{p(\mathcal{H}_0|D)}{p(\mathcal{H}_1|D)}}_{\text{posterior odds}} = \underbrace{\frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\text{prior odds}}$$

$$\text{evidence} = p(D|\mathcal{H}) = \int p(\theta|\mathcal{H})p(D|\theta,\mathcal{H})\mathrm{d}\theta$$

The **evidence** in favour of a model corresponds to the **marginal likelihood** of a model. In other words, it is an averaged likelihood weighted by the prior predictions of the model, which makes the Bayes factor a kind of Bayesian likelihood ratio.

Bayes factors are the new p-values...

Be careful not to interpret Bayes factors as **posterior odds**... Bayes factors indicate how much we should update our **prior odds**, in the light of new incoming data. They **do not tell us what is the most probable hypothesis**, given the data (unless the prior odds are 1:1).

Let's take another example:

- $\mathcal{H}_0$: there is no such thing as precognition
- $\mathcal{H}_1$: precognition does really exist

We run an experiment and observe a $\mathrm{BF}_{10} = 27$. What are the posterior odds in favour of $\mathcal{H}_1$?

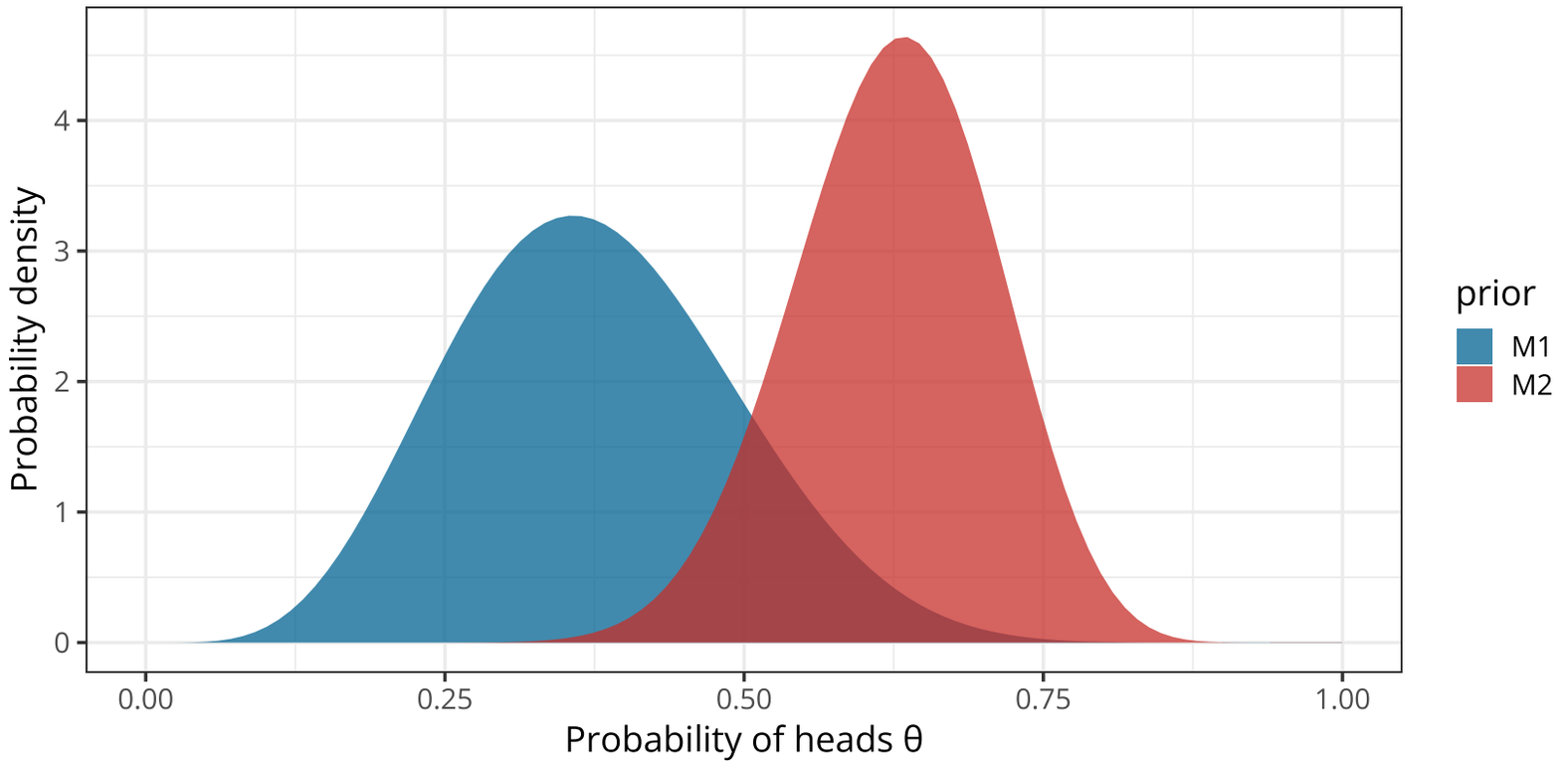
$$\underbrace{\frac{p(\mathcal{H}_1|D)}{p(\mathcal{H}_0|D)}}_{\text{posterior odds}} = \underbrace{\frac{27}{1}}_{\text{Bayes factor}} \times \underbrace{\frac{1}{1000}}_{\text{prior odds}} = \frac{27}{1000} = 0.027$$

What does a Bayes factor look like?

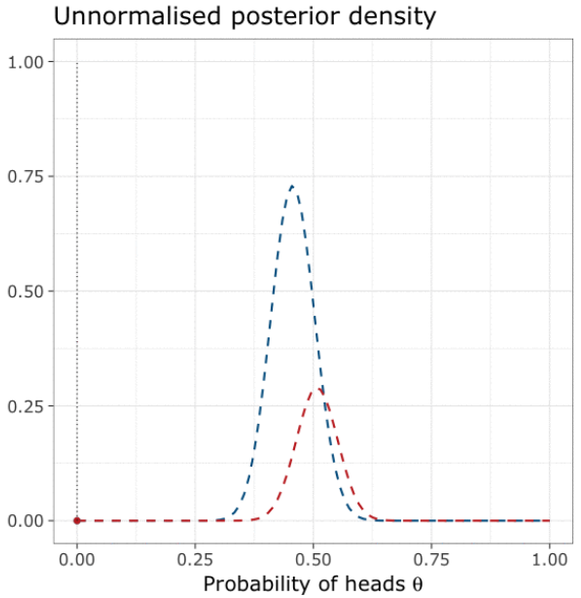
Let's say we want to estimate the bias θ of a coin. For convenience, we can write our predictions as two Beta-Binomial models.

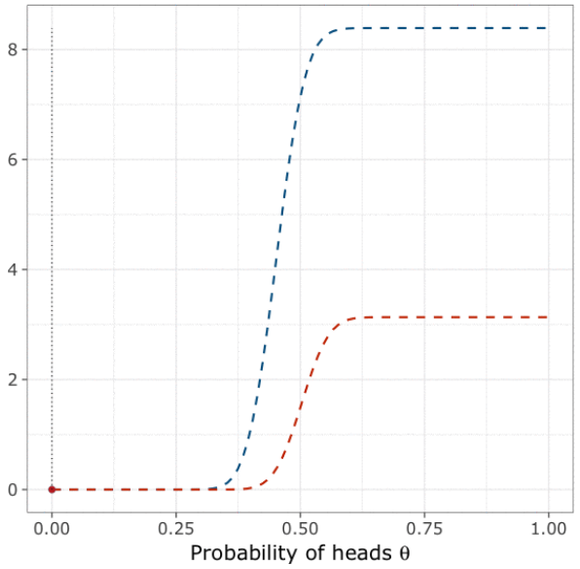
$$\begin{aligned} \mathcal{M}_1 : y_i &\sim \text{Binomial}(n, \theta) \\ \theta &\sim \text{Beta}(6, 10) \end{aligned}$$

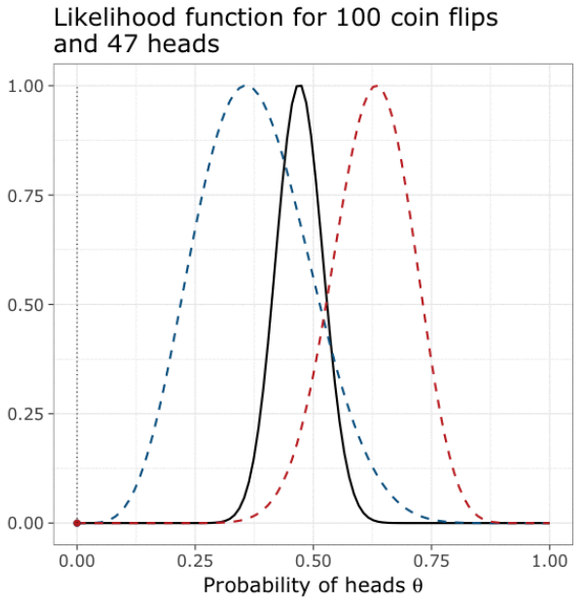
$$\begin{aligned} \mathcal{M}_2 : y_i &\sim \text{Binomial}(n, \theta) \\ \theta &\sim \text{Beta}(20, 12) \end{aligned}$$

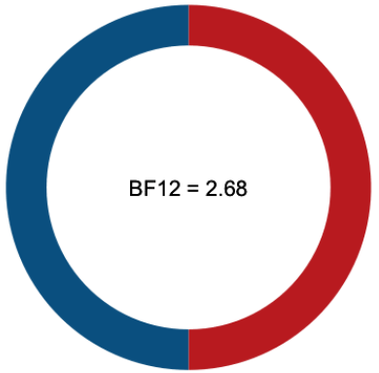


What does a Bayes factor look like?









Problems induced by the mindless use of statistics

Corroborating impossible claims

In 2011, the prestigious Journal of Personality and Social Psychology published this paper written by Daryl Bem, reporting the results of nine experiments, showing evidence in favour of the existence of precognition in humans.

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (*d*) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed.

Keywords: psi, parapsychology, ESP, precognition, retrocausation

Replicability (and reproducibility) issues

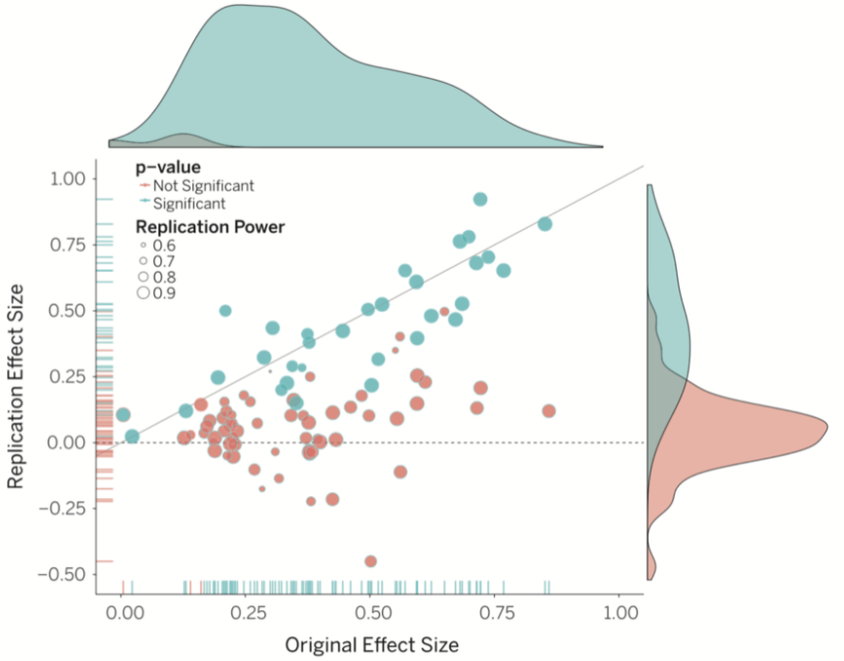
This, amongst other things, prompted an era a (healthy) methodological scepticism, which resulted in a through re-evaluation of classical findings from Psychology and experimental sciences overall. The average replication rate in Psychology was estimated to be around 30%.

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

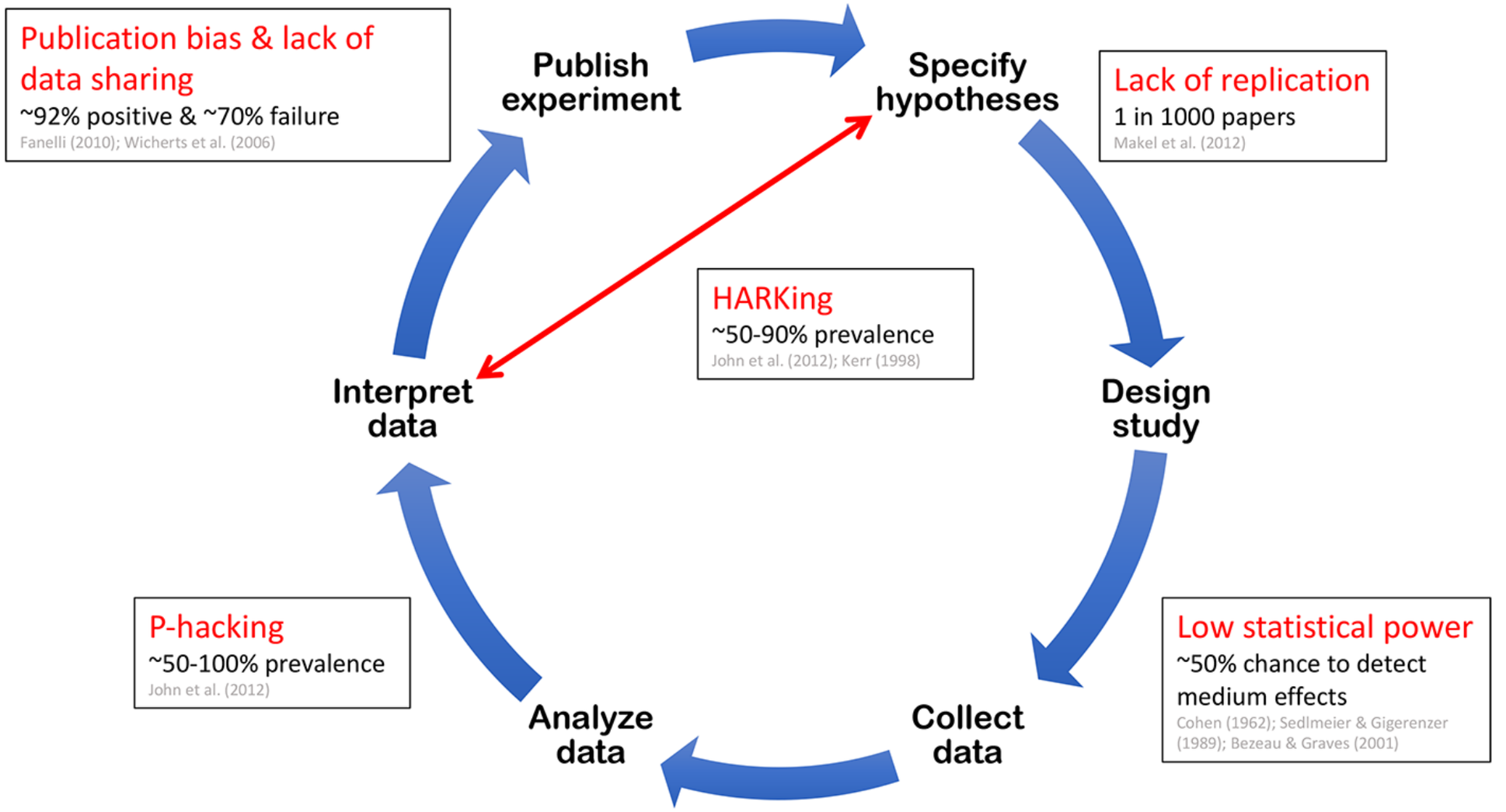
Estimating the reproducibility of psychological science

Open Science Collaboration*



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Scientific cycle and questionable research practises



Questionable research practises

Undisclosed flexibility in data collection, analysis, and interpretation dramatically increases the false positive rates (see also the “garden of forking paths” from Gelman & Loken, n.d.).

General Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
22(11) 1359–1366
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
http://pss.sagepub.com

SAGE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists’ nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

Keywords

methodology, motivated reasoning, publication, disclosure

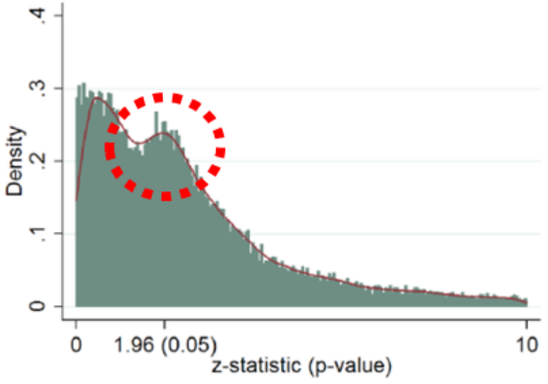
P-hacking

Pressure to produce (e.g., publish), together with widespread misunderstanding of basic concepts in (philosophy of) statistics, have practical/dramatic consequences on the published literature (Figure from Data colada).

Economics

Brodeur et al (AEJ:A, in press)

"Star Wars: The empirics strike back"

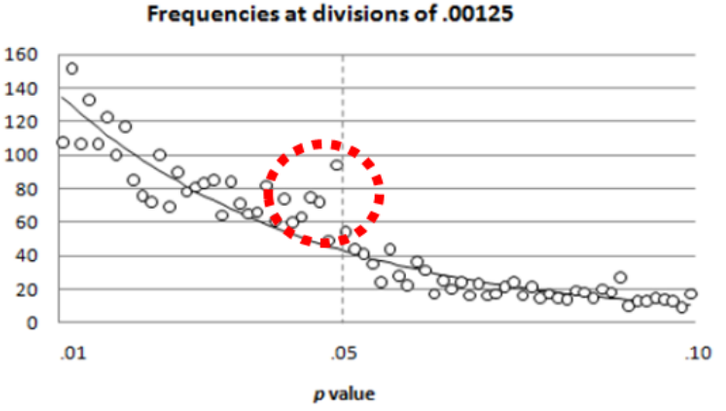


(b) De-rounded distribution of z-statistics.

Psychology

Masicampo Lalande (QJEP, 2012)

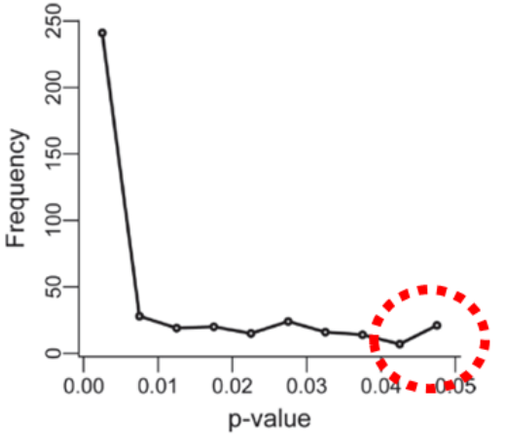
"A peculiar prevalence of p values just below .05"



Biology

Head et al (PLOS Biology 2015)

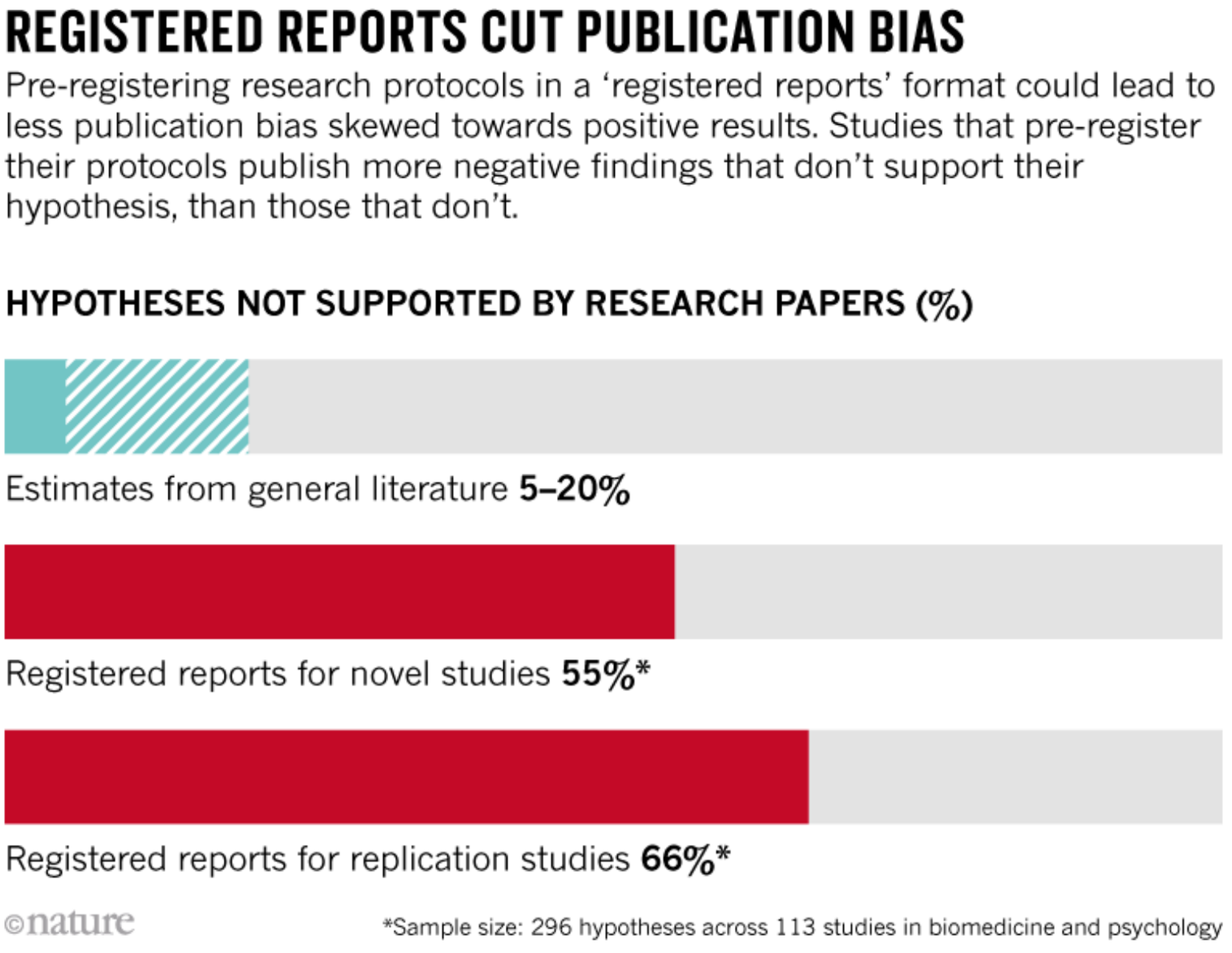
"Extent and Consequences of P-Hacking in Science"



Registered reports: Reducing publication bias

Publication bias is defined as the selective publication of findings based on the obtained p-value (or another statistical index). How can we avoid this biased reporting? A simple idea is not to base the accept/reject decision on a paper on the statistical significance of the results, but rather on the theoretical relevance and methodological rigour of the study (figure from https://www.cos.io/initiatives/registered-reports).

Registered reports: Reducing publication bias



The ATOM guidelines

Do not say “statistically significant”

In 2019, The American Statistician published a special issue on *Moving to a World Beyond “p<.05”*, with the intention to provide new recommendations for users of statistics (e.g., researchers, policy makers, journalists). This issue comprises 43 original papers aiming to provide new guidelines and practical alternatives to the “mindless” use of statistics. In the accompanying editorial, Wasserstein et al. (2019) provide a first practical recommendation.

> “
>
> We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term “statistically significant” entirely. Nor should variants such as “significantly different”, “p < 0.05,” and “nonsignificant” survive, whether expressed in words, by asterisks in a table, or in some other way.

ATOM guidelines

Then, they summarise their practical recommendations in the form of the **ATOM** guidelines:

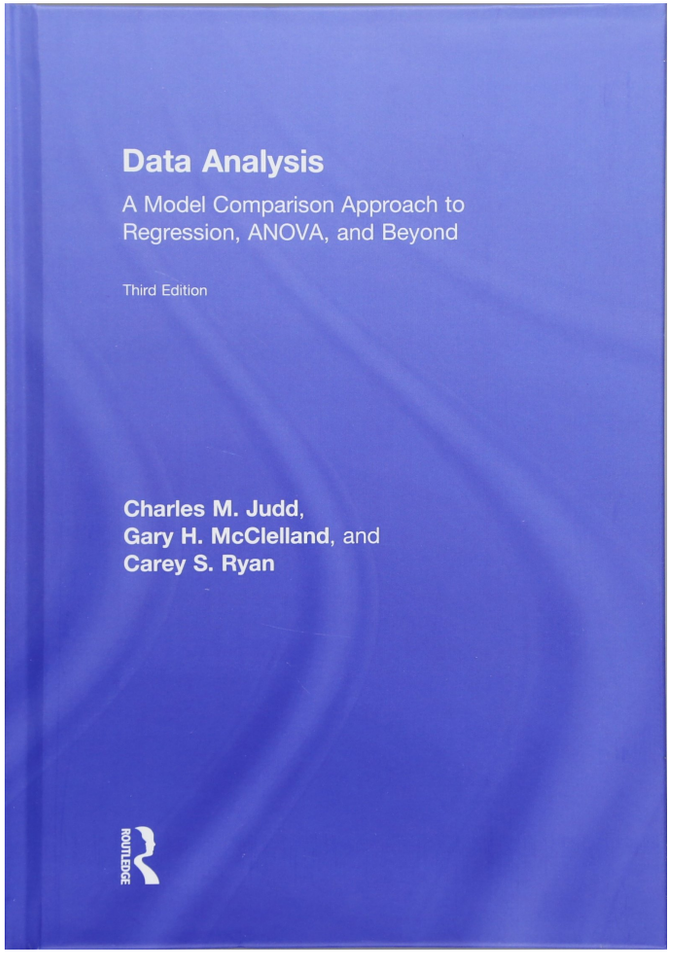
- **A**ccept uncertainty: we must “countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error” (Calin-Jageman & Cumming, 2019).
- Be **T**houghtful: we clearly distinguish between confirmatory (preregistered) and exploratory (non-preregistered) statistical analyses. We routinely evaluate the “validity” of the statistical model and we are suspicious of statistical defaults.
- Be **O**pen: we try to be exhaustive in the way we report our analyses and we beware of short-cuts than could hinder important information to the reader.
- Be **M**odest: we recognise that there is no unique “true statistical model” and we discuss the limitations of our analyses and conclusions. We also recognise that scientific inference is much broader than statistical inference and we try not to conclude anything from a single study without the warranted uncertainty.

How to move forward: A model comparison (and model criticism) approach

Common statistical tests are model comparisons

First insight: Common statistical “tests” (e.g., t-test, ANOVA) can be restated as comparisons of regression models.

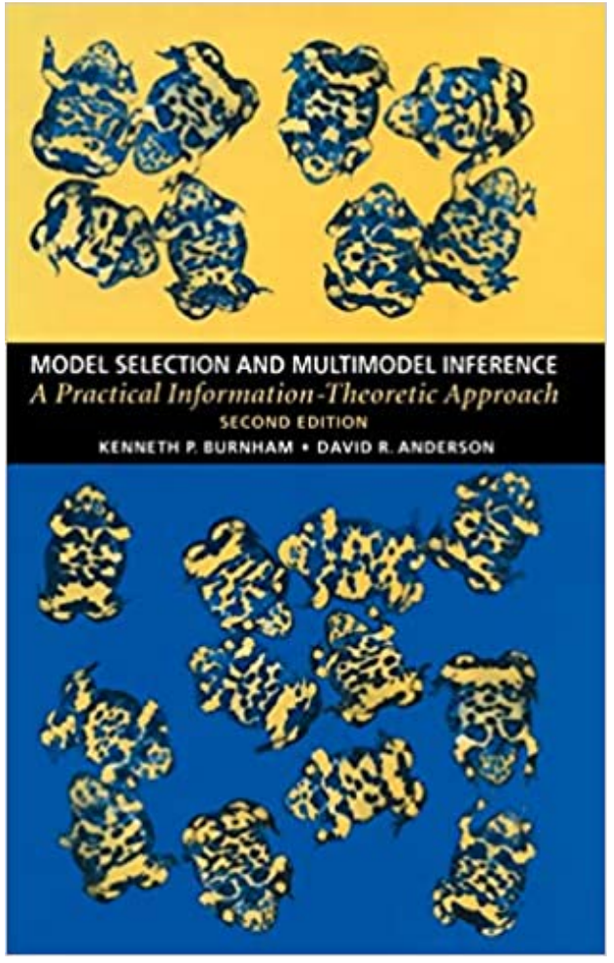
Instead of (or in supplement to) binary conclusions, we also consider how sounds are the models we are comparing, given the phenomenon at hand.



Model comparison and out-of-sample predictive accuracy

Second insight: Instead of comparing unrealistic models (e.g., the “null hypothesis” and the unspecified/default “alternative hypothesis” models), let’s compare interesting models, embodying theoretical hypotheses of interest.

General steps of the model selection approach usually consist in establishing a set of R relevant models, ranking these models (and attributing them weights) using an information criterion, and choosing the best model from the model set to make an inference from this best model. Alternatively, one can make inference from a weighted average of the models’ predictions (aka model averaging or multimodel inference).



Model comparison and out-of-sample predictive accuracy

Hirotugu Akaike noticed that the negative log-likelihood of a model + 2 times its number of parameters was approximately equal to the **out-of-sample deviance** of a model...

$$\text{AIC} = \underbrace{-2\log(\mathcal{L}(\hat{\theta}|\text{data}))}_{\text{in-sample deviance}} + 2K$$

In-sample deviance: how bad is a model to explain the current dataset (the dataset that we used to fit the model).

Out-of-sample deviance: how bad is a model to explain a **future** dataset issued from the same data generating process (the same population).

Philosophical oecumenism: Statistical toolbox

Different statistical tools rest on different philosophical frameworks and aim to answer different questions.

Quantifying the relative evidence for a hypothesis/model

↳ Use Bayes factors or likelihood ratios (do not use p-values for this).

Making decisions while controlling error rates in the long-run

↳ Use NHST & p-values (à la Neyman-Pearson) (do not use Bayes factors for this).

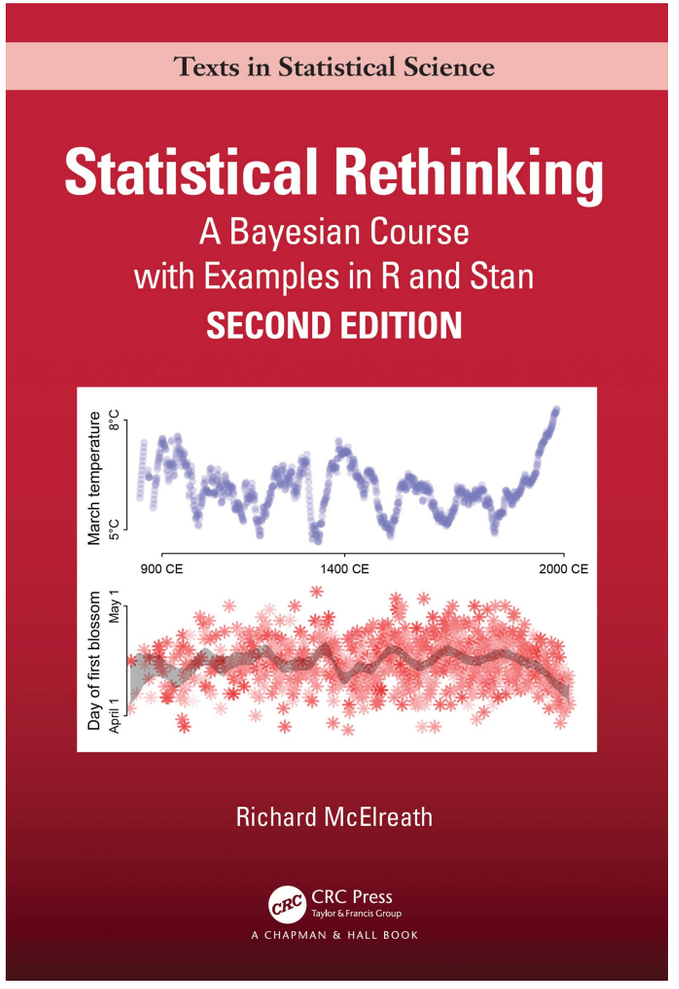
Comparing the (out-of-sample) predictive abilities of models

↳ Use information criteria (e.g., AIC, WAIC).

Towards a principled workflow: Statistical rethinking

Making use of the toolbox and pushing further the statistical modelling and model comparison approach. The focus is on building models, validating them (both against prior knowledge and new observations), comparing them, and using them for prediction and/or inference.

A full course on Bayesian statistical (thinking and) modelling is available freely on Youtube, see the Github repository for more details: https://github.com/rmcelreath/stat_rethinking_2022.



Applying this to empirical data in cognitive sciences

Bayesian workflow*

Andrew Gelman† Aki Vehtari‡ Daniel Simpson§ Charles C. Margossian†
Bob Carpenter¶ Yuling Yao† Lauren Kennedy‖ Jonah Gabry†
Paul-Christian Bürkner** Martin Modrák††

2 Nov 2020

Abstract

The Bayesian approach to data analysis provides a powerful way to handle uncertainty in all observations, model parameters, and model structure using probability theory. Probabilistic programming languages make it easier to specify and fit Bayesian models, but this still leaves us with many options regarding constructing, evaluating, and using these models, along with many remaining challenges in computation. Using Bayesian inference to solve real-world problems requires not only statistical skills, subject matter knowledge, and programming, but also awareness of the decisions made in the process of data analysis. All of these aspects can be understood as part of a tangled workflow of applied Bayesian statistics. Beyond inference, the workflow also includes iterative model building, model checking, validation and troubleshooting of computational problems, model understanding, and model comparison. We review all these aspects of workflow in the context of several examples, keeping in mind that in practice we will be fitting many models for any given problem, even if only a subset of them will ultimately be relevant for our conclusions.

Statistics > Methodology

[Submitted on 29 Apr 2019 (v1), last revised 28 Feb 2020 (this version, v3)]

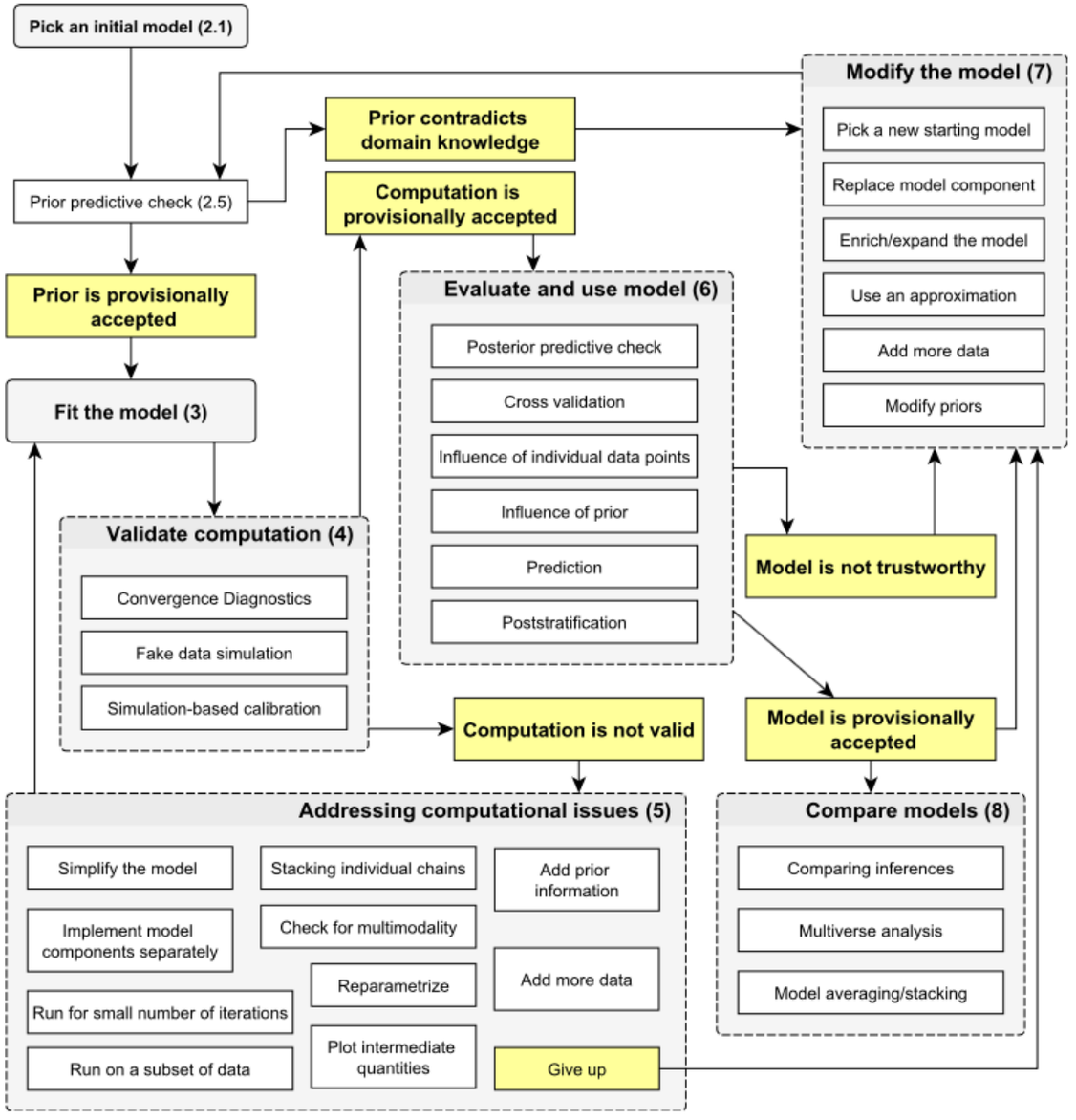
Toward a principled Bayesian workflow in cognitive science

Daniel J. Schad, Michael Betancourt, Shravan Vasishth

Experiments in research on memory, language, and in other areas of cognitive science are increasingly being analyzed using Bayesian methods. This has been facilitated by the development of probabilistic programming languages such as Stan, and easily accessible front-end packages such as brms. The utility of Bayesian methods, however, ultimately depends on the relevance of the Bayesian model, in particular whether or not it accurately captures the structure of the data and the data analyst's domain expertise. Even with powerful software, the analyst is responsible for verifying the utility of their model. To demonstrate this point, we introduce a principled Bayesian workflow (Betancourt, 2018) to cognitive science. Using a concrete working example, we describe basic questions one should ask about the model: prior predictive checks, computational faithfulness, model sensitivity, and posterior predictive checks. The running example for demonstrating the workflow is data on reading times with a linguistic manipulation of object versus subject relative clause sentences. This principled Bayesian workflow also demonstrates how to use domain knowledge to inform prior distributions. It provides guidelines and checks for valid data analysis, avoiding overfitting complex models to noise, and capturing relevant data structure in a probabilistic model. Given the increasing use of Bayesian methods, we aim to discuss how these methods can be properly employed to obtain robust answers to scientific questions. All data and code accompanying this paper are available from this https URL.

Comments: 75 pages, 19 figures
Subjects: **Methodology (stat.ME)**
Cite as: arXiv:1904.12765 **[stat.ME]**
(or arXiv:1904.12765v3 **[stat.ME]** for this version)

Applying this to empirical data in cognitive sciences



A brief summary

Sensible data analysis is not a linear process. A typical workflow involves the following steps:

- Think hard about a (or several) plausible data-generating process(es).
- Think hard about the available prior knowledge and encode it into your model(s).
- Validate these assumptions using simulation (e.g., prior predictive checking).
- If deemed appropriate, fit the model(s) and update prior knowledge using the Bayesian machinery.
- Assess the validity of the model(s) using simulation (e.g., posterior predictive checking).
- Compare various interesting and competing models.
- Make inference about interesting quantities (using as many statistical indexes as needed).
- This is *not* a linear process, feedback loops are often needed between these steps.

Further resources

The special issue on “Statistical Inference in the 21st Century: A World Beyond p < 0.05”: https://www.tandfonline.com/toc/utas20/73/sup1.

Everything is fucked: The syllabus, https://hardsci.wordpress.com/2016/08/11/everything-is-fucked-the-syllabus/.

Some examples of ATOMised reporting of statistical modelling (from my own work): https://pubs.asha.org/doi/abs/10.1044/2018_JSLHR-S-18-0006, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0233282, https://journals.sagepub.com/doi/abs/10.1177/0956797619900336.

Introduction to the Meehlian Corroboration-Verisimilitude theory of science: https://www.barelysignificant.com/post/corroboration1/ and https://www.barelysignificant.com/post/corroboration2/.

The materials of my doctoral course on Bayesian statistical modelling (in French): https://www.barelysignificant.com/IMSB2022/.

Take-home messages

Don'ts

- Do not say "statistically significant" (but you can use p-values to control error rates).
- Do not dichotomise or trichotomise statistical results.

Dos

- Read, digest, and teach some philosophy of statistics and statistical modelling (vs. testing).
- Accept uncertainty. Be thoughtful, open, and modest.

lnalborczyk lnalborczyk https://osf.io/ba8xt www.barelysignificant.com

References

Calin-Jageman, R. J., & Cumming, G. (2019). The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known. *The American Statistician*, *73*(sup1), 271–280. https://doi.org/10.1080/00031305.2018.1518266

Campbell, D. T. (1990). The meehlian corroboration-verisimilitude theory of science. *Psychological Inquiry*, *1*(2), 142–147. https://www.jstor.org/stable/1448769

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Gelman, A., & Loken, E. (n.d.). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*∗. 17.

Gigerenzer, G. (1993). *The superego, the ego, and the id in statistical reasoning* (pp. 311–339). Lawrence Erlbaum Associates, Inc.

Lakatos, I. (1976). *Falsification and the methodology of scientific research programmes* (S. G. Harding, Ed.; pp. 205–259). Springer Netherlands. https://doi.org/10.1007/978-94-010-1863-0_14

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in r and stan*. CRC Press/Taylor & Francis Group.

McElreath, R. (2020a). *Statistical rethinking: A bayesian course with examples in r and stan* (2nd ed.). Taylor; Francis, CRC Press.

McElreath, R. (2020b). *Statistical rethinking: A bayesian course with examples in r and stan* (2nd ed.). Taylor; Francis, CRC Press.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

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*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1986). *What Social Scientists Don't Understand* (D. W. Fiske & R. A. Shweder, Eds.; p. 24). Chicago: University of Chicago Press.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. *What If There Were No Significance Tests?*, 393–425. http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=72FF987997EFB5F0602B02E1A2E04E40?doi=10.1.1.693.9583

Pollard, P., & Richardson, J. T. (1987b). On the probability of making type i errors. *Psychological Bulletin*, *102*(1), 159–163. https://doi.org/10.1037/0033-2909.102.1.159

Pollard, P., & Richardson, J. T. (1987a). On the probability of making type i errors. *Psychological Bulletin*, *102*(1), 159–163. https://doi.org/10.1037/0033-2909.102.1.159

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Collabra*, *2*(1), 6. https://doi.org/10.1525/collabra.28

Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology*, *45*, 19–27. https://doi.org/10.1016/j.newideapsych.2017.01.002

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond " $p < 0.05$". *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913