The power of simulating data: a tool to design experiments, understand data limitations and improve scientific reasoning

Frederik Baumgarten[1,2], EM Wolkovich[1], invite Andrew Gelman

August 28, 2024

[1] Department of Forest and Conservation, Faculty of Forestry, University of British Columbia, 2424 Main Mall Vancouver, BC Canada V6T 1Z4.

[2] Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Zürcherstr. 111, Birmensdorf 8903, Switzerland

Corresponding Author: Frederik Baumgarten; frederik.baumgarten@ubc.ca
Journal: statistical report in Ecology?

Abstract

Science continually seeks to explain patterns in nature that are often obscured by the noise of variation caused by a myriad of influencing factors. To disentangle pattern from noise, link correlation with causality, and build upon the existing body of knowledge, researchers adhere to a scientific workflow. However, concerns have been raised about several aspects of this workflow: the formulation of meaningless (null) hypotheses, insufficient sample size, poor experimental design, and the problematic usage and interpretation of p-values. These issues have led to the reproducibility/replication crisis, biased reporting of findings, over-interpretation, or incorrect conclusions, with patterns sometimes emerging from mathematical artifacts or sheer chance.

Here, we propose the integration of data simulation into the scientific workflow to address these challenges by: 1) developing meaningful and testable hypotheses through the formulation of mathematical or mechanistic models, 2) manipulating effect size, variance, sample size and replication to generate synthetic/fake datasets, 3) exploring the potential and limitations of both the statistical methods and the data obtained, and 4) drawing conclusions that can reliably build upon the existing 'body of knowledge'.

We expect that in a future dominated by powerful AI and machine learning algorithms applied to ever-larger datasets, meaningful, theoretically grounded hypotheses will be more important than ever for revealing underlying causalities. Without the understanding of causality, we may excel at making predictions within known limits, yet fail to understand the driving variables behind them.

Keywords: bayesian statistics, statistcal methods, quantitative ecology, hypothesis testing, scientific reasoning, experimental design, power analysis

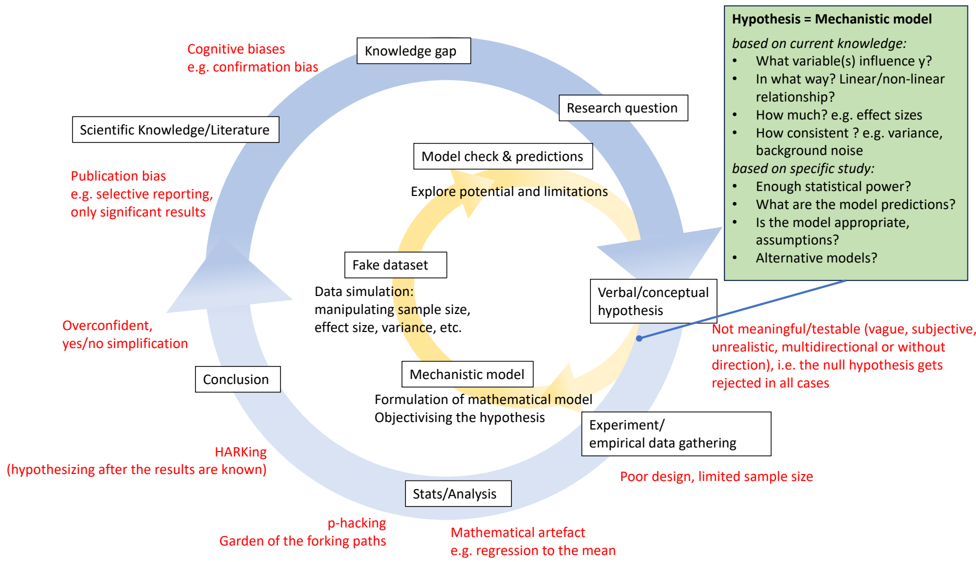

Figure 1: Schematic overview of the current scientific workflow (blue circle) with its related problems (red text) and how some of them can be anticipated with the help of data simulation (loop in yellow). The critical step is the translation of a verbal hypothesis into a mechanistic model expressed in mathematical terms. We believe that this 'extraround' fosters a better understanding of underlying processes founded in theory, clear predictions that can be evaluated and overall better science. The questions summarized in the green box should be addressed in the simulation loop before continuing on the blue circle, e.g. prior to the start of real data gathering/exploration.

1 Introduction

The constant search for patterns

A renowned researcher once explained to me the progression of a curve from a device attached to the bark of a tree next to us. As we watched the curve expand in real time on the screen, he gave a passionate and insightful explanation on the water relations of trees. A few moments later, however, the technician in charge entrusted me with the fact that the fluctuations on the graph emerged largely at random during the installation and calibration process and that the data displayed could not possibly represent anything biologically meaningful. This anecdote—which may sound familiar to many of us—illustrates 1) our tendency to interpret data in a way that supports our pre-existing beliefs, also known as confirmation bias (Nickerson, 1998), and 2) our incredible ability to recognise patterns even when there are none, or in other words, our poor intuition to judge what randomly generated data may look like. Human evolution has equipped us with a hypersensitive pattern detector and a set of cognitive biases (Tversky & Kahneman, 1974) that help us survive but get in the way of objective science. However, distinguishing between noise and pattern is perhaps the most important skill for any researcher.

How can we ensure scientific standards?

To overcome cognitive biases and objectify science, researchers usually follow a scientific workflow illustrated in Figure 1) that traditionally starts with identifying a research question and formulating a hypothesis and ends with a conclusion based on the statistical outcome given the data obtained and the hypothesis at hand (see for example Schwab *et al.* (2022)).
However, there is growing evidence that following this workflow is not enough. Serious concerns have

been raised in the last decades that jeopardize/question scientific integrity and credibility and may slow down scientific progress despite of a record breaking and still increasing publishing rate (REF). Or as Ioannidis (2005) put it in his title: “Why most published research findings are false”. Going through the traditional workflow we briefly outline some major concerns and pitfalls (see also Figure 1).

Hypothesis and predictions

Beyond the purely exploratory nature of a study, a hypothesis well-founded in current theory is essential for causal inference and reasoning (Rajtmajer *et al.*, 2022). However, studies greatly differ in the usefulness of their stated hypotheses, which can range from non-existent to too vague, subjective, untestable, multidirectional, or lacking any direction. Additionally, the common practice of testing to falsify or reject a null hypothesis has been shown to be of limited utility in most cases. This is because null hypotheses often oversimplify complex relationships into a binary outcome and are frequently rejected due to the myriad of influencing factors contributing to data variability (Rajtmajer *et al.*, 2022). Null hypotheses are particularly problematic when combined with the common practice of significance testing (NHST; null hypothesis significance testing), making them unsuitable as a cornerstone in scientific research. (Szucs & Ioannidis, 2017), see also the problems involved with p-values and ‘significance’ below.

Experimental design and sample size

Many studies are under-sampled and lack statistical power. . More precisely, they lack statistical power to detect and are prone to find statistically significant results by chance when in fact there is none (Halsey *et al.* (2015) see below), which has led to a replication crisis in many disciplines with many studies not being able to reproduce previously published results (Ioannidis, 2005; Baker, 2016; Camerer *et al.*, 2018). interactions

Problematic use of p-values

A widely recognized problem, despite its prime role in statistical inference arise from the usage of p-values (Amrhein *et al.*, 2019b, 2017; Halsey *et al.*, 2015). Many studies misinterpret ‘statistical significance’ as direct proof for both relevancy and certainty (REF). Even if p-values are interpreted correctly the multitude of tests often performed in a single study can lead to a ’fishing strategy’ also known as p-hacking, that increases the chance of getting a statistically significant result (Stefan & Schönbrodt, 2023). Finding a $p < 0.05$ during the process of exploring and analyzing data can occur largely unintentional. Most analysis can easily find hundreds of possible ways and combinations to find a potentially relevant pattern, e.g. by including/excluding variables or subgroups. Gelman & Loken (2013) illustrated this problem with the metaphor of ‘the garden of the forking paths’ pointing at the importance of study pre-registration and the awareness of the choices we make during the analysis (Rubin, 2020; **?**). Hypothesizing after the results are known (HARKing) is an additional pitfall that sometimes dangerously attaches causal reasoning to significant results (**?**). Finally, these mentioned problems feed into a reporting and publication bias (Yang *et al.*, 2023; Lin & Chu, 2018) that can substantially distort the literature as was shown for meta-analyses (Van Zwet & Cator, 2021). Overall, there is a trend to abandon p-values and to retire ‘statisitical significance’ as an outdated concept (Amrhein *et al.*, 2019a; Berner & Amrhein, 2022; McShane *et al.*, 2019; Woolston, 2015; Wasserstein *et al.*, 2019; Lee, 2016).

Mathematical artefacts

In some cases the way data are analyzed or displayed create patterns, that arise from mathematical properties, sometimes hard to understand. A famous example is the regression to the mean that is

responsible for many reported effects that were later found to be overestimated or not present at all (REF, **?**, dunninger kruger effect).
Example of decreasing sensitivity from Lizzie? An example in recent ecology research

Will AI solve these issues?

The steep rise in machine learning has propelled recent advancements of artificial intelligence (AI). Current applications excell human-level intelligence by far in many aspects (REF) offering new opportunities for science REF. Many disciplines have already benefited from these very recent advancements, eg.... However, technologies to date are limited to pattern recognition based on correlations only (Pearl, 2019). The critical ingredient lacking is the ability of causal reasoning - a feature so far limited to humans (Pearl, 2019). While it might be possible that we see further improvements towards logic and reasoning in AI systems the separation of causality from correlation will likely continue to play a pivotal role in science. Therefore
problem with hypothesis remain. searching for a model so it includes much of the assumptions/hypotheses expected from a useful model

Aim: The three steps of data simulation

Here, we propose a small addition/adjustment to the traditional scientific workflow: The integration of data simulation prior to data gathering or exploration (Figure 1). Concretely, this involves the following three steps:

1. Translation of verbal hypothesis into a mechanistic model

Models are hypothesis that we evaluate based on our data. Hence, multiple hypothesis testing means applying multiple models. To constrain and justify this process to a reasonable set of hypothesis we propose to build up mechanistic models by incorporating existing knowledge about the influencing variables and their relationships in question. Which explanatory variables are influencing y? Is it linear or non-linear? What do we know from the literature? This step nudges us to think carefully about our hypothesis in terms of model parameters—a great way to translate subjective and conceptual hypothesis to objective mathematical formulations.

2. Power analysis: Tweak the model and simulate

We don't propose a common power analysis based on some significance threshold to calculate the probability of correctly rejecting a null hypothesis when it is false (1 - β). There is nothing wrong with that approach and we even think this a good start. However, here we propose to go further: Explore your model by simulating data from it. Is it doing what you think it is? Generate your fake dataset by manipulating sample size, effect size(s), errorterms and relationships within the mechanistic model to explore how your model behaves. This is a playful, yet powerful way to design an experiment and understanding what your model does. You will find out about the potentials and limits of your model. Besides assessing the statistical power we suggest to focus on m
We believe that incorporating these three steps are vital to do better i) greatly stimulate our mechanistic understanding of underlying processes, ii) facilitate to move away from p-values and significance testing and towards evaluating model parameters such as effect sizes and error intervals, a vital tool to do better science and show how to do it
Helps to address current gaps/limitations:
build hypothesis, then formulation of mathematical model
better design experiments

In the following we go through some concrete examples and show how this procedure can look like.

Example 1: Plant growth and nitrogen

Data simulation is best shown in practice. The example presented here along with more complex ones can also be performed through the R script available in the supplement.
Every study starts with a research question. So lets start with this one: What is the influence of nitrogen fertilization on plant growth? We might want to perform an experiment on sunflowers and fertilize some with nitrogen to assess the biomass after one season. So our dependent variable y is biomass and we might have a clunky hypothesis that plants grow more when fertilized with nitrogen.

What influences y in what way?

Here we can dive deep into the literature. What do we know that influences plant growth? Let's pick nitrogen concentration for simplicity to start with. Lets further assume there is a positive linear relationship of plant growth and nitrogen concentration. So we can express this relationship in a basic linear model with an intercept a (e.g. the biomass at zero nitrogen), a slope b (e.g. the increase of biomass per 1 unit increase in nitrogen) and an error term that captures the natural variability:

```
# Building up a mechanistic model
y = a+b*x + error #model formula with x being nitrogen concentration
error = rnorm(n,0,sigma) #error distribution
```

Note that we chose an error distribution that is drawn from a normal distribution with mean 0 and a standard deviation of sigma.

Setting model parameters

This is the where things get interesting and concrete: What do you think your model parameter should look like? In other words, what biomass do we expect under no nitrogen fertilization? What is the increase in biomass per unit increase in nitrogen? And how variable is this relationship, e.g. what is the background noise? Lets put some numbers here:

```
# specifying model parameters:
a <- 30 #intercept
b <- 7 #slope (effect size)
sigma <- 5 #standard deviation of the error distribution.
```

Simulate!

In order to simulate we only need to set sample size and create a meaningful range of x values, i.e. nitrogen concentrations:

```
n <- 50 #sample size
x<-rnorm(n,10,4) #create data for nitrogen concentration
y<-a+b*x + rnorm(n,0,sigma) #generate y values
fake<-data.frame(x,y)
```

Fit your model and check the outcome

Using your fake dataset you can fit your model, check if the model parameters are in the range you expect them to be and plot them. This is basically your hypothesis as concrete as it gets.

```
fit_1<-stan_glm(y~x, data=fake) # fit the model
print(fit_1, digits=2)
plot(fake$x,fake$y, main="Data and fitted regression line")
a_hat<-coef(fit_1)[1]
b_hat<-coef(fit_1)[2]
abline(a_hat, b_hat)
```

How to use this - Play!

Going back and forward you use this tool to better understand what your model is doing. Is that a reasonable effect size? Is the sample size high enough to detect the effect with reasonable certainty? What if the background noise is much higher? You may realized quickly that this loop will give you confidence in your study design and analysis beyond a simple power analysis.

Avoiding overconfidence

play with replication while holding variance and effect size constant. p-value figure

Increasing importance in the future

Avoiding overconfidence

evergrowing Lit with AI we must learn to ask better questions
the right questions and hypothesis that are testable with current + new methods. Build up house of knowledge instead of using new pattern finding algorithms
how to integrate with AI?

stuff I did't find place yet

Wolkovich *et al.* (2024)

References

Amrhein, V., Gelman, A., Greenland, S. & McShane, B.B. (2019a). Abandoning statistical significance is both sensible and practical. Tech. Rep. e27657v1, PeerJ Inc.

Amrhein, V., Greenland, S. & McShane, B. (2019b). Scientists rise up against statistical significance. *Nature*, 567, 305–307.

Amrhein, V., Korner-Nievergelt, F. & Roth, T. (2017). The earth is flat (p > 0.05): Significance thresholds and the crisis of unreplicable research. *PeerJ*, 5, e3544.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533, 452–454.

Berner, D. & Amrhein, V. (2022). Why and how we should join the shift from significance testing to estimation. *Journal of Evolutionary Biology*, 35, 777–787.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.J. & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2, 637–644.

Gelman, A. & Loken, E. (2013). 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. *Department of Statistics, Columbia University*, 348, 3.

Halsey, L.G., Curran-Everett, D., Vowler, S.L. & Drummond, G.B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, 12, 179–185.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2, e124.

Lee, D.K. (2016). Alternatives to P value: Confidence interval and effect size. *Korean Journal of Anesthesiology*, 69, 555–562.

Lin, L. & Chu, H. (2018). Quantifying Publication Bias in Meta-Analysis. *Biometrics*, 74, 785–794.

McShane, B.B., Gal, D., Gelman, A., Robert, C. & Tackett, J.L. (2019). Abandon Statistical Significance. *The American Statistician*, 73, 235–245.

Nickerson, R.S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2, 175–220.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62, 54–60.

Rajtmajer, S.M., Errington, T.M. & Hillary, F.G. (2022). How failure to falsify in high-volume science contributes to the replication crisis. *eLife*, 11, e78830.

Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology*, 16, 376–390.

Schwab, S., Janiaud, P., Dayan, M., Amrhein, V., Panczak, R., Palagi, P.M., Hemkens, L.G., Ramon, M., Rothen, N., Senn, S., Furrer, E. & Held, L. (2022). Ten simple rules for good research practice. *PLOS Computational Biology*, 18, e1010139.

Stefan, A.M. & Schönbrodt, F.D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, 10, 220346.

Szucs, D. & Ioannidis, J.P.A. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience*, 11.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185, 1124–1131.

Van Zwet, E.W. & Cator, E.A. (2021). The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, 75, 437–452.

Wasserstein, R.L., Schirm, A.L. & Lazar, N.A. (2019). Moving to a World Beyond "p $<$ 0.05". *The American Statistician*, 73, 1–19.

Wolkovich, E.M., Davies, T.J., Pearse, W.D. & Betancourt, M. (2024). A four-step Bayesian workflow for improving ecological science.

Woolston, C. (2015). Psychology journal bans P values. *Nature*, 519, 9–9.

Yang, Y., Sánchez-Tójar, A., O'Dea, R.E., Noble, D.W.A., Koricheva, J., Jennions, M.D., Parker, T.H., Lagisz, M. & Nakagawa, S. (2023). Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology. *BMC Biology*, 21, 71.