1. The power of simulating data: a tool to design experiments, understand data limitations and improve scientific reasoning
2. Between noise and patterns: the power of data simulation in science to overcome perception biases
3. The power of simulating data: a tool for scientists from designing experiments to drawing/reaching reasonable conclusions
4. Between noise and patterns: Overcoming perception biases through data simulation

Frederik Baumgarten[1], Elizabeth Wolkovich, invite Andrew Gelman

May 30, 2024

Abstract

Advertise simulation (for all: frequentists, bayesian, machine learning)
What's the Problem, what do we want to save and what's the solution
- Werwolf: p-values, 'sloppy', simplified stats, over-interpretation of patterns that emerge by chance
- Baby: scientific standards, correct conclusions, knowing the potential and limits of a dataset
- Silverbullet: Data simulation through formulation of a mathematical model and playing with the parameters and the replication

1 Introduction

Opening example

Human desire for patterns even in pure noise
Confirmation bias
But noisy data
how can we ensure a standard?

Current solutions

scientific workflow (fig)
include experiments, null hypothesis testing and their limits(not meaningull, not testable, not interesting) , analysis -> conclusions

Growing evidence that this is not enough

p-values/replication crisis/overconfidence/overinterpretation/mathematical artefacts
Show famous examples. -> could expand into box

Some people hope to address this through machine learning

machines search for patterns without or less bias (or at least in a systematic/objective way)
machine learning is usually amechanistic
problem with hypothesis remain. searching for a model so it includes much of the assumptions/hypotheses expected from a useful model

Aim

We propose an updated approach focussed on simulation + show how to do it
Helps to address current gaps/limitations:
build hypothesis, then formulation of mathematical model
better design experiments
avoid overconfidence/overinterpretation and mathematical artefacts

2 Bus example with simulation workflow

Situation

model - show poisson distr. simulate

no what?

still waiting for the bus...outlayer?
update the model - assumptions
add variable traffic jam

3 Biological example - how to do it

question based + we walk through each one.

what influences y?

nitrogen

what form? linear/nonlinear, near Gaussian, Poisson

linear

What assumptions are reasonable?

for y, alpha, beta
for effect sizes (parameters)
for x data
-> pick some for this example

Simulate!

How to use this - Play!

so many ways, we highlight just a few

Power analysis

to better design experiments

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?