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

June 6, 2024

Abstract

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
Science continuously tries to explain patterns in nature that appear in data, hidden by the noise of variation caused by thousands of influencing factors. Statistics provide us with a tool to separate noise from a real pattern that is unlikely to have occurred simply by chance. At the same time however, some statistical measures have left us overconfident to find this hidden 'treasures' and our minds seem not to be trained enough to judge the difference in many cases. In fact it turns out that humans are really bad in capturing the amount of variation that can be expected.

Introduction

A renowned researcher once explained to me the progression of a curve deriving from a dendrometer - a device attached to the bark of a tree that stood beside us. As we observed the graph on the screen extending in real-time, he provided a passionate and insightful explanation on water relations and plant physiology. However, moments later, the responsible technician entrusted me with the fact, that the device hasn't yet recorded anything meaningful and that all data displayed couldn't possibly depict anything biologically relevant. This example - and I think most of us can share similar stories - shows the incredible ability of our minds to make sense of patterns that underpin our current believes or argument. But it also underpins our blindness of how randomness or noise can look like. However, separating noise from a pattern is perhaps the most important skill of any researcher.

We all manage to go through this world with the help of models in our minds. Most of the time we are not aware of this fact but their implications are all around us. For instance when we wait for a bus we expect the waiting time to be lets say around 7±3 min or when we readjust our expected travel time after an incidence and add the variable 'traffic jam' to our internal model. In short, whenever we have an intuitive understanding of how the world works, our brain suggests basic models - mostly oversimplified and with increasing bias as soon as we enter non-linear relationships. - In many practical cases we get the chance to recalibrate our internal model (or lets call it intuition). For example when the bus after 10 min is still not there and we start realizing that we need to incorporate the variable 'traffic jam'. But in many cases we don't get this kind of feedback or at least not in an informative way because it is too complex to grasp for our minds. While daily life is arguably not a big drawback to this issue, in science it can be.

Current challenges with data interpretation in science

Famous examples of data overinterpretation or drawing wrong conclusions.

Replication crisis is not over

Too many studies are still caried out with simply too few replicates. Especially when drawing conclusions based on interactions.

Problems associated with p-values, associated thresholds and biases in reporting of relevant findings

xx
Although Fisher's concept of the p-value and the associated threshold value date back more than a century, they are still the standard in many scientific disciplines, institutes and research groups.
xx
xx

Modern statistics departs from p-values

problem with p values, many scientists continue using them, despite the ban of p-values and words like "significant" by some journals. Language change. Introduce Bayesian stats.

Here it would be nice to show a graph with the inreased usage of Bayesian approaches over time for different science fields and journals or the appearance of words like 'p-value' and 'Bayesian' or? over time to underpin that Bayesian stats is on the rise.

xx

Simulating data as an integral part of conducting science

Beyond that many Bayesian statisticians see themselves not only as simply using a different method but rather as a new philosophy of conducting science that starts already prior to an experiment or way

before looking at some real data.

We think it is time to highlight the advantages of one, in our opinion the most important step of the Bayesian approach to conduct science. And I am convinced that this step benefits to all scientists, no matter which way they proceed. In fact, we believe it would be most beneficial for students through all academic stages and should become an integral part of statistical teaching as it allows to convincingly self assess the potential of a particular dataset and to evaluate its suitability to answer a specific question of interest.
This paper aims to bridge frequentists and Baysian ways of thinking by encouraging the usage of data simulation to check assumptions, evaluate replication, define hypothesis and to assess model suitability.

What does data simulation imply?

Simulating data is not a new idea and there are excellent books demonstrating their usefulness and power as an integral approach to make statistical inference (do you know some examples beside ROS?). However, this approach still hasn't been considered/adopted by most researchers particular in some science fields and most concerning has not entered statistical teaching at most universities.

Fake data simulation is a great way to overcome these issues. Because it forces and allows us to:
- think about our currently existing model of how we see the world, clarifying our expectations
- get feedback in a valuable way so that we can calibrate our model (in this case this means our mathematical model but also our general understanding)
- get a better sense and intuition of natural variability
- think more carefully about the design of an experiment
Fake data simulation is not about directly generating the y-variable of interest with a certain distribution in mind. It is more complex but more interesting and meaningful: We think about what factors (that are truly relevant) are influencing the y-variable of interest in which way. This is important because at this point we already clarify the mathematical relationship between explanatory variables and y. These are parameters (e.g. intercept and slope). Because real data are never exactly on a line (in that case) we need to think about introducing a meaningful amount of noise and how we expect its distribution. - In other words simulating data means we have to think about a mathematical function and define our parameters so that we can produce y-values that match our expectations including a realistic amount of noise.

Examples

here we would provide some code to simulate data to give an example of how many replicates would be needed to get consistent results with a given effect size etc. or show that we can get a whole distribution of p-values if we repeat our simulation enough

xxx

xx

xxx

xx

XXX

XX

XXX

XX