Data and the reproducibility of research results

What are the roles of ‘data’ and ‘reproducibility’ in the scientific method?

“Data is the fundamental building block from which scientific knowledge grows; extensions of work rely on the data representing fundamental scientific truths. In this case, reproducibility is the means by which we can verify these truths and thus guarantee forward progress of science.”

What are the federal requirements on sharing data? To what extent do you think these rules are enforced?

"It requires civilian federal agencies to provide guidelines, policy and procedures, to facilitate and optimize the open exchange of data and research between agencies, the public and policymakers."

COMPARE THIS TO ANOTHER RESPONSE:

"According to Wiki the federal requirements on sharing data as stated by the America COMPETES Act 2007 'require civilian federal agencies to provide guidelines, policy and procedures, to facilitate and optimize the open exchange of data and research between agencies, the public and policymakers.' the requirements as far as I understand are stated in loose general terms and any agency can evade the act. It is not a binding requirement but the agencies would have to show valid reasons to withhold data and eliminates chances of being whimsical."

THE FIRST RESPONSE WAS AN UNATTRIBUTED DIRECT QUOTE FROM WIKIPEDIA.

LET'S SHARE SOME THOUGHTS ABOUT HOW THIS FITS IN WITH OUR PREVIOUS DISCUSSION ON PLAGIARISM.

"When a scientific book or paper is published, the authors must provide the data based on which they drew their conclusions. To my knowledge, I think these rules are well enforced since most papers I read have the source of data and the procedure of how to generate the data."

I THINK THIS IS NOW HAPPILY THE CASE IN MANY LEADING JOURNALS. IT WAS NOT, 5 YEARS AGO.

Advanced statistical methods often require sophisticated computational implementations. Should statistical researchers be expected to share their computer code on request?

"I think they should share their code on request, or at least core code. If they plan on somehow monetizing their code, or having some other reason to keep it secret, then they need to think about what claims they make in their paper, and should at least make sure some of the basic versions are available."

Why would a statistican NOT share code?

Does scientific secrecy ever have benefits to the individual? If so, how should one decide when to keep scientific secrets, bearing in mind the possible cost to scientific reputation (or lost opportunity to earn it)?

What is the difference between data and a statistical model for the data? For example, comment on the assertion "Let $y_1, \ldots, y_n$ be independent identically distributed data."

"Data is just a collection of information. A statistical model is a description of the data proposed by a researcher from his perspective. It may be true or may not. Here $y1, \ldots, y_n$ being iid is an assumption which says that the y observations are unrelated data that are generated from the same population or share the same source. This is an idea that a researcher has about data which is not necessarily true. The actual data will be the observations themselves."

"We must be careful not to confuse data with the abstractions we use to analyze them." (William James, 1842–1910)

THIS SUGGESTS DATA MAY HAVE A HIGHER ROLE THAN MODELS. BUT ONE RESPONSE POINTED OUT AN OPPOSITE VIEW:

"Statistical models for data are an additional layer of thinking/information on the data, and might be as important and valuable as the data itself. For example anyone can get info about stock market movement for free online, but stochastic models for that and estimates of parameters for those models can be extremely valuable."

(i) What sort of mathematical objects does the definition of iid apply to?

(ii) What sort of mathematical objects are data?

(iii) Can data be iid?

The remaining questions consider the following hypothetical case study:

Ben is a Statistics PhD student who has written computer code for a simulation study to test a new statistical theory and methodology which he is developing. He plans to put the results in his thesis and to publish them in a journal paper. The results of the simulations are usually consistent with his theoretical analysis. However, sometimes the code crashes, particularly when investigating more extreme values of the parameter space. Ben has checked and rechecked the code very carefully, and cannot find any error. He decides that there must be some weird numerical effect, perhaps to do with occasional extremely large or small numbers. Ben decides to report the results only in the region of the parameter space where the code never crashed.

Is Ben's course of action a reasonable balance between the necessity to make progress on his thesis and his desire to report correct results?

"It seems that Ben has put reasonable time and effort to report correct results. However, he still needs to disclosure the condition when the code crashes. If he could not find the explicit bug in his code, at least he should try to identify and introduce the range of parameters that gives the consistent results. Without such disclosures, he is trying to deceive readers by pretending that his code contains no errors."

What are the ‘data’ in this example? What is ‘reproducibility’ in this context?

“The data are the simulated data sets generated for the paper. Reproducibility is the ability to take the functions and run them on this data set on ones own computer; the random number generator seed should be saved so the exact numbers and estimates in the paper are replicated.”

Ben asks your opinion on how to proceed. What is your advice?

"I would suggest Ben double-check the code and find where the numerical effects happen. Or he could turn to expertise on numerical effects for help."

"Ben should fix the bug!"

"My advice would be mentioning the failure of the alogrithm in the report. Reporting this failure is indeed useful for future studies."