Research Ethics and Reproducibility

Clark

2022-09-22

XKCD and Jelly Beans…

When we conduct an experiment, we want to ensure that if someone else was to conduct a similar experiment under similar conditions they would obtain a similar result. Now we shouldn't expect the results to be the exact same due to randomization and sources of unexplained variance, but in general, if we observe an effect (and make a claim) we would want this to generalize to other situations (or else what's the point of doing our study?!?)

Let's go back to XKCD. Why wouldn't this study be reproducible?

This is a clear example of what's called 'p-hacking' or 'fishing'. We are performing J tests and then reporting only the best result given the data.

This, at least to me, seems clearly not ethical. Because, mathematically, if we have done 5 different tests at $\alpha = 0.05$, the probability we've committed a Type 1 error isn't 5%. In fact we can compute it:

Now, this is easily fixed, right? We talk about this in MA376 and even MA206. In reality this doesn't actually happen that much in published research. *Most* researchers know that you shouldn't P-hack. A more nuanced issue in reproducibility is what famed statistician Andrew Gelman calls, 'A garden of forking paths'. Here I'm going to borrow from his 2013 paper with Eric Loken.

The idea behind the garden of forking paths is that whenever we are doing research we are presented with a myriad of choices that we can take. Say for instance, we have an abnormal value, or an outlier, what are we going to do with this? Say that we are analyzing data about nutrition and we want to classify high calorie and low calorie diets, how are we defining high and low calorie diets? Or another choice is to conduct our study at all! Say we have data and we look at it and it doesn't appear interesting. The choice to analyze it or not is a choice we make.

Let's peel that back a step further. Let's say we conduct an exploratory data analysis and we have potentially a large set of data and we are interested in predicting whether a diet is effective or not. Through the act of looking at the data we decide that the clearly is not a relationship between age and diet so we don't consider it in the model.

This isn't p-hacking, we didn't conduct a statistical test; but we sorta did…

Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidential election. In two studies with large and diverse samples, ovulation had drastically different effects on single versus married women. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more conservative, more religious, and more likely to vote for Mitt Romney. . . . Overall, the ovulatory cycle not only influences women's politics, but appears to do so differently for single versus married women.

Where do forking paths potentially occur here?

The above study also brings out another issue in research. Conducting a descriptive analytics (or predictive analytics) and pretending it's a diagnostic analytics. When we conduct a diagnostic analytics we are asking the question as to 'why' a phenomenon occurred. The study done above is misleading as a sample of women only completed the survey once,there were no repeated measurements, hence women were not compared at different days in their cycle.

So what can we do about this? Well, if we are doing an actual experiment we can pre-register our data collection rules and our data analysis protocol. Certainly the IRB process encourages us to do this for human research. But, for us oftentimes we aren't doing an experiment, or we are doing predictive study. Maybe we are trying to predict RASP graduates? Maybe we are trying to predict who will be successful at West Point?

Here we can try to replicate our OWN findings. How can we do this in a predictive model?

If we cannot replicate our own findings due to time or data limitations, it benefits us to produce reports that lead to replication. (From Kitzes, J. (2017))

1. Document what your code does as a whole. For simple scripts use comments to describe exactly the actions the code performs. For more complex code use a README file.
2. Focus on making your code readable. Efficient and optimized is good but prioritize making your code easy to read
   (a) Use descriptive, specific names for each object in your code
   (b) At the start of each function or method include comments that outline exactly what the function or method does
3. Comment and document as you go. You won't do it later. You just won't. I am willing to put money on it.
4. Automate repetitve tasks and avoid hardcoding when feasible. Tasks that need to be repeated by a human can be opportunities for human error to steak in. Avoid manual intervention in the workflow when possible

I will also add another note here. You should default to making your data readily available. If we truely believe our analysis we should welcome other researchers accessing our data and challenging our results. The point of research is to contribute to the field, if someone takes your work and makes it better you have done your job!

Related to this, we should also not be ashamed if our work is a replication of the work that others have done. To again quote Andrew Gelman

In the long term, I believe we as social scientists need to move beyond the paradigm in which a single study can establish a definitive result. In addition to the procedural innovations [of preregistration and mock reports], I think we have to more seriously consider the integration of new studies with the existing literature, going beyond the simple (and wrong) dichotomy in which statistically significant findings are considered as true and nonsignificant results are taken to be zero. But registration of studies seems like a useful step in any case.

What he seems to be saying here is that our attempts to replicate the results of others (or to show where their results do NOT hold true) is helpful and meaningful in science. If we are using an algorithm to do facial detection and our results show that the algorithm does not work in a specific situation, we have done a service to our field.

The last issue I want to talk about that relates to reproducibility is **Data Leakage**. From the reading last night what is data leakage?

Here we want to ensure that, if we are using a predictive algorithm, someone could actually predict off of it! The second part of Data Leakage is really, to me, just bad statistical practices. The authors talk about training models on datasets that are narrower than the population that they are intended to reflect. In other words, our *sample is not reflective of the population we are trying to make inference about.* This should sound familiar from MA376…

I want to leave you with some resources:

Lots of good videos on what reproducibility means for ML modeling

https://sites.google.com/princeton.edu/rep-workshop/

Andrew Gelman's blog that covers lots of manuscripts that do this poorly (see his 'zombies' link)

andrewgelman.com