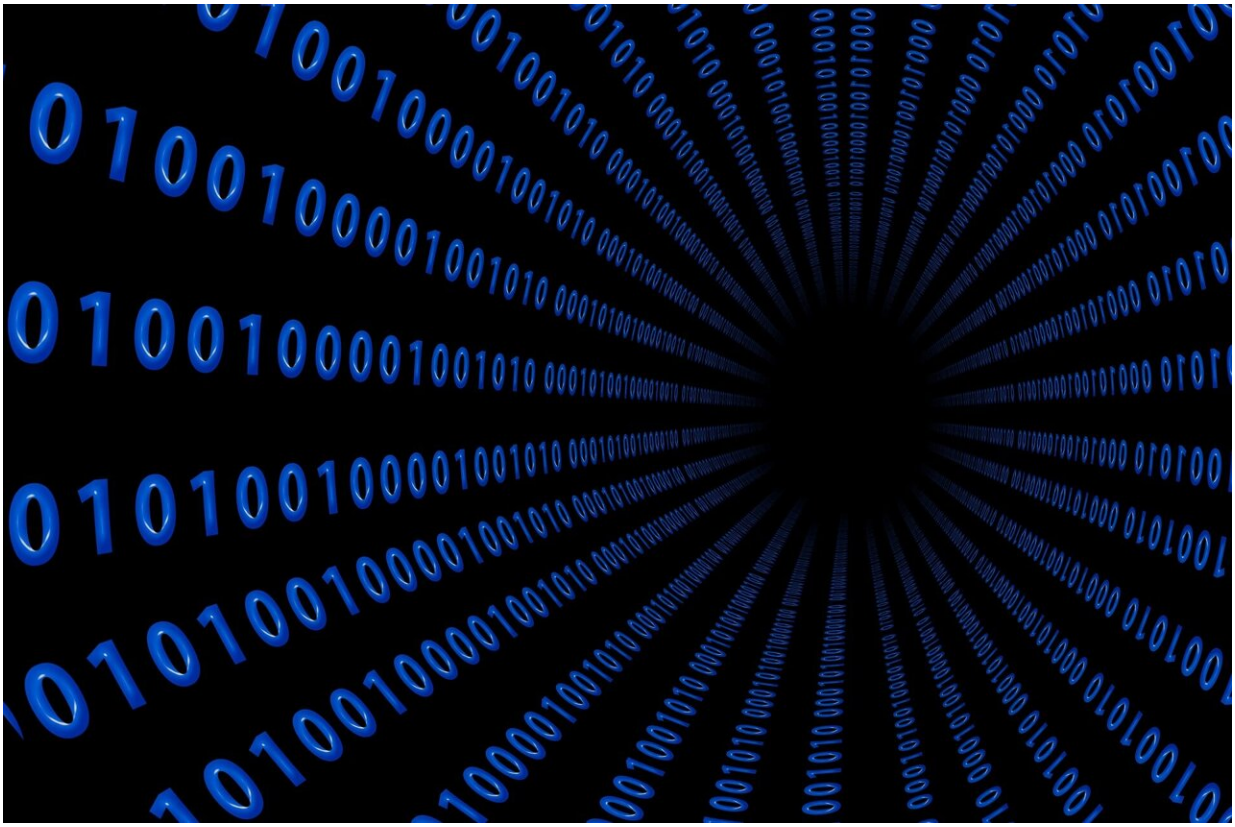
When should data scientists try a new technique?

January 26 2023, by Adam Zewe



Credit: Pixabay/CC0 Public Domain

If a scientist wanted to forecast ocean currents to understand how pollution travels after an oil spill, she could use a common approach that looks at currents traveling between 10 and 200 kilometers. Or, she could

choose a newer model that also includes shorter currents. This might be more accurate, but it could also require learning new software or running new computational experiments. How to know if it will be worth the time, cost, and effort to use the new method?

A new approach developed by MIT researchers could help data scientists answer this question, whether they are looking at statistics on ocean currents, violent crime, children's reading ability, or any number of other types of datasets.

The team created a new measure, known as the "c-value," that helps users choose between techniques based on the chance that a new method is more accurate for a specific dataset. This measure answers the question "Is it likely that the new method is more accurate for this data than the common approach?"

Traditionally, statisticians compare methods by averaging a method's accuracy across all possible datasets. But just because a new method is better for all datasets on average doesn't mean it will actually provide a better estimate using one particular dataset. Averages are not application-specific.

So, researchers from MIT and elsewhere created the c-value, which is a dataset-specific tool. A high c-value means it is unlikely a new method will be less accurate than the original method on a specific data problem.

In their proof-of-concept paper, the researchers describe and evaluate the c-value using real-world data analysis problems: modeling ocean currents, estimating violent crime in neighborhoods, and approximating student reading ability at schools. They show how the c-value could help statisticians and data analysts achieve more accurate results by indicating when to use alternative estimation methods they otherwise might have ignored.

"What we are trying to do with this particular work is come up with something that is data-specific. The classical notion of risk is really natural for someone developing a new method. That person wants their method to work well for all of their users on average. But a user of a method wants something that will work on their individual problem. We've shown that the c-value is a very practical proof-of-concept in that direction," says senior author Tamara Broderick, an associate professor in the Department of Electrical Engineering and Computer Science (EECS) and a member of the Laboratory for Information and Decision Systems and the Institute for Data, Systems, and Society.

She's joined on the paper by Brian Trippe, a former graduate student in Broderick's group who is now a postdoc at Columbia University; and Sameer Deshpande, a former postdoc in Broderick's group who is now an assistant professor at the University of Wisconsin at Madison. An accepted version of the paper is posted online in the *Journal of the American Statistical Association*.

Evaluating estimators

The c-value is designed to help with data problems in which researchers seek to estimate an unknown parameter using a dataset, such as estimating average student reading ability from a dataset of assessment results and student survey responses. A researcher has two estimation methods and must decide which to use for this particular problem.

The better estimation method is the one that results in less "loss," which means the estimate will be closer to the ground truth. Ponder again the forecasting of ocean currents: Perhaps being off by a few meters per hour isn't so bad, but being off by many kilometers per hour makes the estimate useless. The ground truth is unknown, though; the scientist is trying to estimate it. Therefore, one can never actually compute the loss of an estimate for their specific data. That's what makes comparing

estimates challenging. The c-value helps a scientist navigate this challenge.

The c-value equation uses a specific dataset to compute the estimate with each method, and then once more to compute the c-value between the methods. If the c-value is large, it is unlikely that the alternative method is going to be worse and yield less accurate estimates than the original method.

"In our case, we are assuming that you conservatively want to stay with the default estimator, and you only want to go to the new estimator if you feel very confident about it. With a high c-value, it's likely that the new estimate is more accurate. If you get a low c-value, you can't say anything conclusive. You might have actually done better, but you just don't know," Broderick explains.

Probing the theory

The researchers put that theory to the test by evaluating three real-world data analysis problems.

For one, they used the c-value to help determine which approach is best for modeling ocean currents, a problem Trippe has been tackling. Accurate models are important for predicting the dispersion of contaminants, like pollution from an oil spill. The team found that estimating ocean currents using multiple scales, one larger and one smaller, likely yields higher accuracy than using only larger scale measurements.

"Oceans researchers are studying this, and the c-value can provide some statistical 'oomph' to support modeling the smaller scale," Broderick says.

In another example, the researchers sought to predict violent crime in census tracts in Philadelphia, an application Deshpande has been studying. Using the c-value, they found that one could get better estimates about violent crime rates by incorporating information about census-tract-level nonviolent crime into the analysis. They also used the c-value to show that additionally leveraging violent crime data from neighboring census tracts in the analysis isn't likely to provide further accuracy improvements.

"That doesn't mean there isn't an improvement, that just means that we don't feel confident saying that you will get it," she says.

Now that they have proven the c-value in theory and shown how it could be used to tackle real-world data problems, the researchers want to expand the measure to more types of data and a wider set of model classes.

The ultimate goal is to create a measure that is general enough for many more data analysis problems, and while there is still a lot of work to do to realize that objective, Broderick says this is an important and exciting first step in the right direction.

More information: Brian L. Trippe et al, Confidently Comparing Estimates with the c-value, *Journal of the American Statistical Association* (2022). DOI: 10.1080/01621459.2022.2153688