(Reproduced from http://bulletin.imstat.org/2012/11/terences-stuff-does-it-work-in-practice/)

Terence's Stuff: Does it work in practice?

Many of us are in the business of inventing novel statistical methods or improving upon existing ones. That is, we try to find new ways of answering a question, in context, with some kind of data. The question, the context or the data may be new, or all three may be familiar, and we want to do better than others who have gone before us. Our goal is to discover a method that will *work in practice*. How can we tell?

Papers summarizing research of this kind typically take the following form. After reviewing previous work, they describe their new method. Next, they might apply it to *simulated* data that mimics the *real*data for which their method was designed. If so, they will compare their results with the (known) *truth*, and also with the results obtained using other methods. They need to do well in these comparisons to be publishable. Finally, they might apply their new method to some *real* data. If so, they must then decide how good their answers are, and whether they are as good as those obtained using other methods.

All of this sounds straightforward, yet there are pitfalls. Let's call the first approach the simulation test, and the second the real data test, and look at each a little more. Remember, the goal is to determine how well a method *works in practice*.

The first point I want to make is this: *simulated data is not real data*. Most simulations use idealized models which we know don't accurately describe real data. We hope the difference doesn't matter in practice. We learn from simulations (approximately) how a method works in theory, without theory. The simulation test can only be as good as our ability to simulate data that embodies *all* features of the real data *relevant* to the performance of our method in practice. But we will never know what these are. Simulation can be very helpful for theory and methods development, but not for learning how a method works in practice.

The second point I want to make is this: few real datasets come equipped with truth that will permit us to assess the effectiveness of our new methods. At times they will, and when they do, we should treasure them. For example, with prediction in time-series, or with classification using a training and test set, you can apply your method to new data for which the answers are known. (I'll pass over the harder—but important—question of whether the data you use to demonstrate that your new method works adequately reflects the data to which your method will be applied in the future.) Suppose that you are estimating regression coefficients or testing hypotheses in the social sciences. In how many such contexts will you know the true value of the coefficients or the true status of the hypotheses? In this situation you will probably rely on indirect arguments, surrogates for the truth, or criteria that are necessary but not sufficient, to suggest that your method works on real data. You might consider the sizes or the signs of the regression coefficients, and argue that they are plausible. Or you might look at a histogram of p-values, and show that it is fairly uniform, apart from a spike near zero. Or you might apply your model to independent datasets and seek consistent results.

I don’t think we can expect easy answers to the question, “Does it work in practice?” Indeed, it may be as hard as deciding, “In what circumstances can we pass from this observed association to a verdict of causation?” That is, we may want to satisfy several criteria, and even then not be sure. We should seek agreement on what these criteria might be, and I’d put positive and negative controls and replication high on the list. Specially created datasets can help.

Positive and negative controls? Biologists have been dealing with these issues for decades. They have a highly-developed framework for determining whether a method works in practice, one which makes extensive use of controls of different kinds, including (context-specific) positive and negative controls. Definitions and an introduction to their approach can be found in *Experimental Design for Biologists* by David V. Glass. We statisticians have a lot to learn from them.

Let me end by recounting a recent experience. I was talking to statisticians about removing *unknown* unwanted variation from microarray data using negative controls (variables with regression coefficients known or expected to be zero). When I described how to use positive controls (variables with regression coefficients known or expected to be non-zero) to see if our method worked in practice, my audience looked lost (mea culpa). I then asked them, “How do we tell whether a statistical method works in practice?” The first, and loudest, answer was, *“By simulation!”*