In the past two decades, with the advent of the Internet, the ubiquity of data generation and collection, and the prevalence of microarray and related “-omic” technologies, Robert Tibshirani’s work developing statistical tools has gained importance. The origins of his research can be partly traced to August 1998, when Tibshirani returned to Stanford University as a professor with dual appointments in the departments of health research and policy and statistics. He had graduated from Stanford with a doctorate in statistics in 1984 and spent the intervening years researching and teaching at the University of Toronto in his native Canada. His return came at a fortuitous time. Researchers at Stanford had developed DNA microarrays, which allow for the simultaneous study of tens of thousands of genes, to investigate the mechanisms of cancer as well as patients’ responses to treatment. Microarrays produce vast amounts of data, and Tibshirani and colleagues developed a technique called “significance analysis of microarrays” (SAM), which revealed patients’ reactions to radiation treatment for skin cancer. PNAS recently spoke to Tibshirani, who was elected to the National Academy of Sciences in 2012, about his current research.

**PNAS:** What got you interested in statistics and its applications?

**Tibshirani:** In high school, I enjoyed math, and in college, math and computer science. But statistics allowed me to apply math and computation in other scientific contexts. After college, I worked at the Princess Margaret Cancer Hospital in Toronto trying to figure out what treatments worked best for patients. That experience got me interested in cancer research, in general, and the application of statistics to clinical trials. That was in 1979, well before the Internet, big computation, and microarrays.

**PNAS:** Two techniques that you developed, SAM and the lasso method, have been widely deployed in diverse fields. How do these techniques help interpret data?

**Tibshirani:** When you’re measuring thousands of genes, for example, you need to know whether the differences in a measurement that you see between control and experimental groups are significant and not the result of random fluctuations. You typically do this by computing a $P$ value. If the $P$ value is very small, you can say that this result is not likely to have happened by chance. But the challenge is: What do you do when you have not one but thousands of measurements to compare? That’s what SAM does: It compares your results to what you would get with your data if you scrambled the labels, providing an estimate of the false discovery rate.

SAM looks at individual features like gene expression. Lasso, on the other hand, is more sophisticated because it allows you to look at a combination of features and use them to build a predictive model. Let’s say you’re giving patients a particular drug and want to know for whom it will work. You’re looking at patients’ gene expression but are also considering their age, sex, family history, and other health measurements. Lasso can build a rule that predicts whether a patient will respond or not. It delivers a “sparse” regression model by sorting through a large number of features and forming a combination that predicts a response. The resulting model might also hint at an underlying disease process.

In today’s science, we have far more features than we can measure, so the lasso is very useful. I invented it, actually, before my return to Stanford, but it didn’t receive much attention until computation got faster and our ability to collect large data sets got easier.

**PNAS:** Your Inaugural Article (1) discusses statistical learning and selective inference. Can you explain the importance of these concepts?

**Tibshirani:** We have a problem in science now: We can generate large amounts of data, but we tend to “cherry pick” for the strongest associations. Let’s look at how this plays out in practice: Say you have one study with a low $P$ value that demonstrates that a particular medication is effective. That’s research a journal will publish. But what if there were 99 other studies that show that this drug is ineffective—but with mixed $P$ values? Those are results that journals are not necessarily interested in, so they don’t get published. It’s called the “file drawer” problem, and it’s essentially a selection problem. We’re selecting for low $P$ values, but we need to know if the results are truly significant or if they’re occurring by chance. This is a major challenge for science in general.

So the effects of selection can exaggerate the strengths of relationships. The same kind of selection bias occurs within a given statistical analysis, when the researcher searches for the most significant patterns in his dataset. The methods I described in my article help you determine which of these results are worthy of further investigation. We’re in the process of producing free, public domain software to help scientists do this.

**PNAS:** How do you envision the future of big data?

**Tibshirani:** It’s a very exciting time to be a statistician. The traditional methods that many of us learned as graduate students are insufficient to analyze modern data, which is dynamic and interconnected. Social networks, for instance, produce millions of interactions each second, and the observed data are far from a random sample; this makes it difficult to interpret. I also think the healthcare system in the United States, which is not centralized like it is in many other nations, provides interesting statistical challenges, especially in genome-wide association studies, where you have potentially three billion measurements per patient. Sorting through these complexities and finding the strongest signals—that’s going to be part of my challenge for the remainder of my career.

---


This is a QnA of a recently elected member of the National Academy of Sciences to accompany the member’s Inaugural Article 10.1073/pnas.1507583112.