Newsweek

Lies, Damned Lies And …

In the Wild West that is genome research, statisticians are the new sheriffs in town.

Sharon Begley
NEWSWEEK
Updated: 2:32 PM ET Jul 12, 2008

Just to get the obvious out of the way, no science should be burdened with a memorable dis from Disraeli (as popularized by Mark Twain), and statisticians have had it up to here with being tagged as more nefarious than liars and damned liars. Especially when the invention of new techniques, and the application of tried-and-true tools to new problems, lets them play quite a different role. Done right, statistical analysis can ferret out lies, or at least mistakes.

Consider: one of the field's more significant recent inventions is the "false discovery rate" procedure. It's a way to tell when a claimed discovery, such as whether a certain gene raises your risk of some disease, is worth the pixels it's announced in. Such claims form the basis for a growing DNA-testing industry and, not incidentally, life-and-death decisions.

But that's not what biostatistician Michael Schell of the H. Lee Moffitt Cancer Center & Research Institute will be unveiling next month. Schell studies such important questions as whether particular cancer therapies help patients. That will be front and center at the annual meeting of the American Statistical Association and other stats groups in Denver, as will research on how to make election polls more accurate (measuring the intensity of voters' views is crucial), for instance, and which voters get asked for ID (55 percent of blacks do, and 45 percent of whites, Stephen Ansolabehere of MIT will report). Schell will be taking a busman's holiday from biomedical stats, however, returning to what hooked him on statistics when he was a boy: baseball. In this case, why does Coors Field offer the greatest home-field advantage of any park in baseball history?

The difference between the Rockies' percentage of home wins and road wins is .169, greater even than that of the infamously home-friendly Fenway Park (.108). How come? After using stats to rule out everything else they could think of, Schell and Dan Ayers of Vanderbilt University hit on the spread between the Rockies' batting average at home and away: 40 points, when the league average is 20. "It's a huge difference," says Schell, and enough to account for the team's home/away winning spread. The reason the Rockies hit so much better at home, he suspects, is that the thin air at Coors Field—the reason balls fly out of the park like pigeons at a dog run—affects pitches. It keeps curveballs from breaking as much as they do in, say, Chicago. The Rockies adjust their swings for that; visiting players don't. But the Rockies then have trouble in parks where pitches do break. Presto: a big home/away batting-average spread, and Coors's home-field edge.

Statistical analysis derives much of its power from tools that let you isolate a bunch of factors that might explain something and test which ones really matter. This "regression analysis" is so basic, you learn it in a first-year course. (If pollsters had used it, they might not have trumpeted the silly notion last week that owning a pet makes voters lean toward John McCain. In fact, race, age and marital status make you more or less likely to own a pet, and influence political leanings. Pets are irrelevant.) Regression can also explain the home-field advantage throughout sports, Jeffrey Simonoff and Gary Simon of NYU have found. They ruled out such usual suspects as batting last and the toll that travel takes on a visiting team. Instead, crowd support and familiarity with local conditions—such as the thin air at Coors Field, or Fenway's Green Monster—explain the effect, they find.

Traditional tools of statistics work fine when analyzing mere thousands of data points. But "in the new century, we've had a whole new class of problems, with 100,000 times more data than we used to see," says Bradley Efron of Stanford University. That is especially true in genomics, which spews out data like there's no tomorrow. In the Wild West that is genome research, statisticians are the new sheriffs in town.

Thanks to new tools like the false-discovery-rate technique, they've repeatedly shot down claims. In searching for genes that raise your risk of developing some disease, for instance, geneticists applied a traditional test: is a link to a gene less than 5 percent likely to be a coincidence? Unfortunately, when you're testing thousands of genes, that lets too many false positives through. "Most of the literature [linking a gene to a disease] is riddled with false discoveries," warns Fred Wright of the University of North Carolina. Something to consider before forking over thousands of dollars to have your DNA analyzed by companies marketing the tests to consumers.

Geneticists now acknowledge that scores of claims about disease and DNA are wrong, and have tightened their criteria for what's likely to be a real link. Unfortunately, "the process of applying a stricter threshold [to keep out false links] leads you to overestimate" what proportion of the risk of getting a disease is due to DNA rather than, say, lifestyle, says Wright. At next month's meeting, he and UNC colleagues will unveil a new statistical tool "to tamp down that inflation and get more realistic estimates of the magnitude of a gene's effect." In applying it to some highly publicized studies, such as one linking genes to diabetes, he has knocked down genes' importance by as much as one third. If Disraeli were alive, he might amend his line to "lies, damned lies, revealed by statistics."

URL: http://www.newsweek.com/id/145865