

# The two sides of the statistical war

Eric Blair

6 August 2008

There is a little war in the statistical world. Like other little wars, like Mac vs PC or Ford versus Chevy or Protestant versus Catholic, everybody who isn't on one of the teams has no idea how to differentiate between the two sides. Also, there is no resolution to the central question.

Tukey [1977, pp 1–2] gives a metaphor of the detective and the judge. The detective gathers all the evidence he can, regardless of whether the evidence will be admissible in court or whether it proves guilt or innocence. He just compiles a thick a notebook as possible and worries about sorting it out later. The judge does the sorting. She is bound by law to ignore some evidence, and is comfortable ignoring most of the detective's notebook as irrelevant to the final, narrow question before the court.

Data-oriented inquiry has a very similar division, of descriptive modeling and hypothesis testing. The descriptive modeling step simply gathers information and puts it into a human-comprehensible format. The hypothesis test uses the strict laws that you forgot from statistics class to make a more objective statistical claim.

In the in entry #237, we saw many examples of descriptive modeling: take in all the airline prices, and list all the patterns you—or a computer—can find. Find the smallest demographic/marketing subgroup who all want to vote for Obama. Observe that pesticide use has been going up with time, and cancer rates have gone up with time.

There are two steps to take from there, one of which (developing a causal link) I won't talk about until next time. The main step is the hypothesis test, wherein you come up with some means of verifying the claim that the relationship you just found is what you claimed it was.

We need those extra steps because correlations could be sheer coincidence, meaning that they may reflect a true statement about the data at hand, but we shouldn't rely on them next week, or claim that there is some causal story that made that correlation happen. Stupid coincidences happen all the time and are easy to manufacture.

The problem with all our wonderful technology, however, is that as the power of your relation-searching machinery goes up, the power of your hypothesis testing diminishes. Here are two questions:

- Randomly draw a person from the U.S. population. What are the odds that that person makes more than \$1m?
- Randomly draw 350 million people from the U.S. population. What are the odds that that wealthiest person in your list makes more than \$1m?

The odds in the second case will be much higher, because we took pains in that one to pick the wealthiest person we could. [That is, the first is a hypothesis about just data, the second is a hypothesis about an order statistic of data.]

Now say that you have a list of variables before you.

- Claim based on intuition that  $A$  is correlated to  $B_1$ . What are the odds that your claim is OK with more than 95% odds?

- Write down the best correlation between  $A$  and  $B_1, B_2, \dots, B_{1,000,000}$ . What are the odds that your best correlation is OK with more than 95% odds?

With a big enough list of variables, you are guaranteed to find a correlation (or any other model) that passes any hypothesis test you want.

You've read stories like this before: researcher inspects the data *very carefully*, eventually stumbles upon a relationship that works, thinks about how it makes sense that those two variables are related, and then publishes. With luck, it's something quirky enough to get into the NYT, Economist, or any other pop science outlet that happily reports one-off, unreplicated studies about how a crazy and unexpected variable has an important effect on the things we care about.

And that's the core of the conflict. The descriptive camp points out that it can develop badass means of testing a thousand hypotheses, and the hypothesis testing camp points out that once they do that and pick the best correlation out of a thousand, all the hypothesis tests are basically invalid until modifications are made that the descriptive kids won't bother to make.

There are a few ways by which we can have too many hypotheses. The simplest is to just have a systematic list of a few million possibilities in need of testing. If we can get a million genetic markers from a drop of blood, which we can do, then we need to correct for that as we run a million hypothesis tests. People usually do the corrections in this case.

Before moving on to the real disasters, let me note that some people reject the discussion to this point. If variables  $A$  and  $B_{2891}$  are truly and honestly correlated, then that fact is true no matter whether we ran exactly one test or ran a million. There is no Heisenberg weirdness here: observing the correlations does not change them.

However, our tests and how we interpret them are changing. A hypothesis test makes sense only in a given environment, and that environment has to include the data, how the data was gathered, cleaned, and pre-inspected, and what other tests are being run at the same time. In the cookbook-format manual, none of this gets mentioned: the recipe calls for a list of numbers, mashed into a certain statistic, compared to a certain table, and you're done. But once a human observer comes along, you're already out of the textbook.

But the people who don't quite get the concept of the multiple testing problem don't get much cred. It's subtle and easy to get wrong, but people eventually work it out. If you write a loop to run every regression of a list of twenty variables against some outcome (usually GDP or some overall productivity number), then you are guaranteed to find an excellent fit to your data, and you will have no proof that what you found is any good, and nobody will respect you.

No, that's not where the debate lies.

**Eyeballing multiple testing** Here's another way to get too many hypotheses: given a list of twenty variables, you can produce what is called a Trellis™ or lattice plot, which gives a 2-D dot plot of every variable against every other. It's not hard to put plots for twenty variables on a screen, and then scan to find the pair whose line is sharpest and shows the best correlation. Congratulations, you've just run  $20 \times 19 = 380$  hypothesis tests. When tested more formally, the correlation you just spotted is almost guaranteed to hold, even if your data is pure noise. Or you can try any of a multitude of other visualizations that will similarly allow you to see hundreds of relations at once.

The DataViz field is trendy right now. There are a few icons of the field who are working hard on self-promotion, such as Edward Tufte, whose books show how graphs can be cleaned up, chartjunk eliminated, and grainy black and white fliers from the 1970s cleaned up through the use of finely detailed illustrations in full color. John Tukey's *Exploratory Data Analysis* (cited above) is aggressively quirky, and encourages disdain for the hypothesis testing school.

These guys, and their followers, are right that we could do a whole lot better with our data visualizations, and that the stuff based on facilitating fitting the line with a straightedge should have been purged at least twenty years ago.

The underlying philosophy, however, is humanist to a fault. The claim is that the human brain is the best data-processor out there, and our computers still can't *see* a relationship among a blob of dots as quickly as our eye/brain combo can. This is true, and a fine justification for better graphical data presentation. And hey, we humans would all rather look at plots than at tables of numbers.

But the argument forgets that humans are so good at seeing relationships among blobs of dots that we often see patterns in static (there's a word for this tendency: apophenia). We look at clouds and see bunnies, or read the horoscope and think that it's talking directly to us, or listen to a Beatles song about playground equipment and think it's telling us to kill people. Given ten scatterplots, you *will* find a pattern—in fact, if a psychologist were to show you a series of ten seemingly random inkblots<sup>1</sup> and you didn't see a reasonable number of patterns in them, the psychologist might consider you to be mentally unhealthy in any of a number of ways.

Better data visualization doesn't address the problem of apophenia. In fact, following Tukey's lead, the people who focus on clean testing are characterized as not seeing the value of all these full-color diagrams. They're wearing blinders for the sake of being good Boy Scouts and not seeing the trees and grass and chirping birds around them. Conversely, the testing people generally see little value in all these full-color plots and want to go back to inferring things.

So this is the current battleground in the descriptive-versus-testing war. No side can win—there are no tests for overtesting, so this is all just intuition and opinion. We can write down in a cookbook that if your data-analysis model includes a series of ten tests, you need to make such-and-such an order-statistic correction. But how do you write into a textbook model framework that you surfed charts of the data for forty-five minutes, including eight 3D plots and two Trellis™ diagrams?

Further, both sides are necessary, and both sides have valid points. So this is a perfect recipe for sniping back and forth forever.

---

<sup>1</sup>[http://ar.geocities.com/test\\_de\\_rorschach/index.htm](http://ar.geocities.com/test_de_rorschach/index.htm)



Figure 1: If you don't see faces, you're crazy. Oh, and there's a penis and vagina in every inkblot too.

But overcharting (and defining what that means) is not where the true problem lies.

**The looming problem** After all there is a middle ground, where a person comes in with some idea of what the data will say, rather than waiting for the scatterplot of Delphi to reveal it. Then the researcher refines the original idea in dialog with the data. The closer something fits prior human beliefs, the more we are inclined to accept it, so the researcher is not on a pure fishing expedition, but is not wearing blinders to what the data has to say.

So one researcher could be reasonable—but what happens when there are thousands of reasonable researchers? When a relevant and expensive data set has been released, a large number of people will look at it. I've been to an annual conference attended by about a hundred people built entirely around a single data set, and who knows how many weren't able to fly out. With so many humans looking at the same set of numbers, *every reasonable hypothesis will be tested*. Even if every person maintains the discipline of balancing data exploration against testing, we as a collective do not.

Every person was careful to not test every option, so the order statistic problem seemed to be dodged, but the environment is not just one researcher at a computer, but thousands across the country, and collectively, a thousand hypothesis tests were run, and journals are heavily inclined to publish only those that scored highly on the tests. So it's the multiple testing problem all over again, but in the context of the hundreds or thousands of researchers around the planets studying the same topic. Try putting *that* into a cookbook description of a test's environment.

There's no short-term solution to this one.

In the next episode, I'll take this a little further.

## References

John W Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.