Too many tests

Ben Klemens

16 March 2009

This is allegedly a blog to accompany *Modeling with Data*, so I don't feel too bad repeating its opening paragraph:

> Statistical analysis has two goals, which directly conflict. The first is to find patterns in static: given the infinite number of variables that one could observe, how can one discover the relations and patterns that make human sense? The second goal is a fight against *apophenia*, the human tendency to invent patterns in random static. Given that someone has found a pattern regarding a handful of variables, how can one verify that it is not just the product of a lucky draw or an overactive imagination?

It's the first paragraph in the book because this conflict is just that important. If you're in an inferential mindset while working on a descriptive technique, or vice versa, you'll be hopelessly confused. If you design a study so that it is as descriptive as possible, you can easily lose inferential power, and vice versa. The conflict is also a social conflict, and you'll find a lot of examples of yelling between a descriptively-oriented person on one side and an inferentially-oriented person on the other.

I'll be giving you many, many examples of how the descriptive-inferential conflict plays out, and why it's important for reading the newspaper, gathering data, and teaching. But for now, let me just clarify the point a little by giving you the most common example and the most common point of conflict: selecting the number of hypothesis tests to run.

First, here are two questions:

• Randomly draw a person from the U.S. population. What are the odds that that person makes more than $1m/year?

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

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 that A is correlated to B_1. What are the odds that your claim will pass the appropriate test with more than 95% confidence?

• Write down the best correlation between A and B_1, B_2, $\ldots$, $B_{1,000,000}$. What are the odds that your best result will pass the appropriate test with more than 95% confidence?

You can sit down at your computer and run the same correlation test in the first example with B_1 and in the second example with $B_{\rm best}$, and both tests will have the same name and produce the same format of output from your software, but you've just run two entirely different tests. Just as with the case of our wealthiest individual, the best result from a million results is very likely much better than any single result. Even a million hypothesis tests over noise are likely to find results that are very significant. [There are disciplines, techniques and tricks to mitigate the problem, but I won't get into those now. E.g., see the Bonferroni correction, on pp 318–319 of *Modeling with Data*.]

So the context of a test matters, in sometimes subtle ways. This creates friction between the descriptives and the inferentials, because the descriptives are building tools to search for the best relationships among as much data as possible, while the inferentials realize that those same tools can diminish our power to have confidence in those best relationships.

Next time, I'll talk about how this relates to some currently trendy aspects of dataviz. After that I'll reapply this abstract point to academia at large, and Freakonomics-type journalism in particular.