File 20101020.0700: Notes from Reading Group this week:

I introduced a paper from 2006 by Gelman and Stern called 'The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant' (The American Statistician 60(4), pp. 328–331. I had hoped that it would provoke a good discussion of experimentation and statistical significance in software engineering and security research.

In the room were John Lyle, Dr Flechais, Cornelius, Ronald, Shamal, Anbang and myself on video. I began by showing a graph from the Cosmic Background Explorer (COBE) satellite in 1992 that measured 2.725 K blackbody radiation observed over the whole sky by the Far Infrared Absolute Spectrometer (FIRAS), a liquid helium-cooled sensor. [Actually, the graph never showed up because screen sharing did not work, but I emailed it out afterwards.] The graph is interesting because the error bars are 400 sigma wide. Along the entire curve, the difference between predicted and actual values is so small as to be invisible. This is the most precise and accurate measurement of a real system that has ever been done.

So why, I asked, are error bars almost never shown on measurements in software enginnering research? Part of the reason, we are told, is because experiments in software engineering are expensive. That brings up an interesting point, though; the LHC is an expensive experiment too. It cost more to construct than even a large software project. But the LHC, as an experimental device, is used to perform millions of experiments over the lifetime of the equipment, probably. I wonder what the amortised cost of each experiment is? (CERN issues a lot of reports; we could probably calculate it from their financials and their annual reports.) Looked at that way, a software project costing £1MM that can only be used once for the purpose of experimentation starts to look more expensive than even specialised scientific apparatus. Someone should do the sums and write a paper about this.

To get the discussion started, I offered the following definition (from Krantz, 1999): 'an effect is a change in the central tendency of a distribution of measurements, with identical measurements subject to unexplained error.' I stated, for the purpose of argument, that in software engineering a good outcome is often considered to be coming within 20 percent of your predicted cost, schedule and functionality, 80 percent of the time. Ivan immediately pointed out that these are all process metrics. The difference between what we can measure in software engineering and what civil engineers measure is the difference between measuring a sick patient's temperature and knowing that they have bacterial pneumonia. Fever is a symptom, not the problem, just as cracks in a dam are indicative of insufficiently strong materials, or unexpected stress leading to strain. Mechanical engineers have a wealth of strength-of-materials knowledge to build on. Software engineers seemingly are always implementing new bespoke materials. 'We should always try to measure the property we wish to know directly, rather than measuring another property from which the property we wish to know can be inferred' (Kletz, 1999, p.89).

From the paper by Krantz (1999), there are three schools of thought in psychology, when it comes to statistics. Freudians believe that nothing happens by chance, so to them every event is significant. By contrast, the conventional use of statistics in psychological research (which is after all, as John pointed out, today mostly applied statistics) is to use statistical significance tests to distinguish real effects from happenstance. There is a third school of thought in psychology which advocates designing experiments in such a clever way as to maximise the signal to noise ratio so that the effects are obvious. Ivan noted that the last option only front-loads the analysis problem. I countered that it puts the burden on the originator, thereby removing the necessity for having a sophisticated understanding of statistics [what this paper is about] from a much larger number of readers. If that is not such a good end in itself, at least it removes a large number of potential sources of error.

John suggested looking a the provenance community; they are all about reliable data. I pointed out a long quotation from a book by Melton (1962) in which the editors of journals used to replicate results themselves before they would publish; that is never done anymore. The standard of evidence, said Dr Flechais, should be to publish enough information to allow others to replicate it. Then others can publish three more papers showing that they were able to replicate your results, or not.

Cornelius mentioned that it is possible to choose your statistics in such a way as to get any statistical result you need; this provoked a discussion of lying with statistics.

Ivan: at least in SecHCI, the standards for statistical rigour are nonexistent. What is the most meaningful way to compare two pieces of software? Shamal posed a question: what is the background radiation of the internet? Is it changing? Dr Flechais said he would love to see his experimental design. The notion of statistical significance, said Dr Flechais, is a social construct. It has be be within a finite budget. It cannot be arbitrarily high. Ninety-five percent is the accepted level of confidence, so we have no option—we must aim for that as the standard for publishing in this discipline.

We examined Figure 2 in the paper and discussed the presentation of evidence. Error bars, everyone agreed, are clearly a superior way of presenting the information, better than the obfuscatory method used by Blackman et al. (1988). 'If you want high confidence values, and frankly you should to have good science, then you have to choose the areas you measure.' Dr Flechais would not go below 95 percent.

The point of this paper is to correct a misunderstanding of the meaning of statistical significance by people who use statistics but are not statisticians. 'Even large changes in significance levels can correspond to small, nonsignificant changes in the underlying quantities' (Gelman and Stern, 2006). This is different from the more common error of misunderstanding what confidence intervals mean, and is completely separate from the question of what significance level is acceptable. The authors present three examples, two drawn from published studies, demonstrating that the authors of those studies claimed that differences in statistical significance were themselves significant, where examination of the underlying data does not support that assertion. The relation between the effect size and the standard error must be taken into account also. 'One way of interpreting lack of statistical significance is that further information might change one's decision recommendations' [ibid]. '...one should look at the statistical significance of the difference rather than the difference between their significance levels' [loc. cit.].

I posed a philosophical question: should you publish results that say, 'we were not able to achieve a 95 percent confidence interval, but we can calculate what we did get, and here it is'? Dr Flechais replied that it is always valuable to know what methods are good and which are bad. He put me on the spot a bit, asking if the question relates to my confirmation report. I hesitated, but then forged ahead recklessly and replied that that a few months ago I would have answered yes (and in fact that is what I was publicly saying in the question-and-answer period after my paper presentation in France). But since then, I made a conscious effort to improve the methods of my thesis by moving in the direction of a numerical simulation whose parameters I control more than I could control the vissicitudes of an ethnomethodological study. My new study, based on a simple but complete abstract model, grounded by a proof linking it to an old and well-trusted economic theory, implemented through a physical analogue suited to numerical integration, and finally tied to mathematics that are again solvable to arbitrary precision (although not necessarily having a closed form solution!) is my attempt to stay on the straight and narrow with respect to statistical significance. I can perform experiments on my model to a 95 percent confidence interval easily. My abstract model is general enough to describe all the accreditor, developer and certifier behaviour I have observed in two case studies. That is the real direction in which I have made progress.

What this paper says is, if you have a statistical measure that does not reach a 95 percent confidence level, then that tells you something. It tells you that you are not measuring the right things, or you are not using the right tools.

Dr Flechais then asked point blank if I am doing what Kletz said not to. How do I know that I am measuring the value of interest directly, and not some hopefully correlated value in my physical analogue? I replied that the argument lies in the apparent commonality of behaviour seen between the oscillation of risks under multiple bids in the model and the sequence of events observed in two different case studies.

Last thing: maybe measuring the only thing you can measure is a good thing. The discussion moved to the Common Criteria, and whether CC evaluations are comparable. I stipulated that they are not, even when evaluated to the same Protection Profile and EAL, because of the evaluated configuration problem. The CC itself has evolved radically, from version 2.1 to 2.3 to 3.0 to 3.1. They are actively trying to make it better, and they are, but it makes historical comparisons difficult.

References