HEADS AND REDS: THE HUMAN TENDENCY TO SEE PATTERNS IN RANDOM DATA

Andrew Oswald, University of Warwick www.andrewoswald.com

Empirical evidence is more important than theory. But when I was young I did not appreciate how easy it is to find exciting, but illusory, statistical patterns in data. We all -- I include myself -- need humility when we do empirical work. This is especially true if we use small data sets of less than 1000 observations. So these days I try to ask myself: (i) can I check my exciting discovery by making sure that it is there within subsamples of my own data, by splitting the sample into men and women, or young and old, or before-1980 and after-1980? (ii) did I come up with my theory ex post, after already seeing the data? (iii) have I, without realising it, searched across lots of possible empirical patterns before stumbling on my exciting finding? Unfortunately, if we subconsciously pre-search for patterns then we cannot apply conventional statistical significance levels when we hit upon an exciting discovery in the data.

Here is an illustration.

With a collaborator, I am doing experiments. We have a theory that we call Time of the Day Effects. We believe that the time of day has important consequences. I am Andrew and I run my lab. She is called Amanda and runs hers. I am working on coin-tossing -- heads and tails. She is working on the spin of a roulette wheel, with only two colours -- red and black. I throw a coin each morning 6 times; then the same in the afternoon: 12 throws a day. I do this for a week, so sample size is 84. In the other experiment, Amanda is spinning her roulette wheel. She also does it 6 times in the morning, and 6 in the afternoon – for 7 days. Our total observations are therefore 168. We agree to collaborate on any finding in either experiment, whatever it turns out to be, and to send a jointly authored paper to the prestigious journal, the Journal of Scientific Discoveries.

How likely are Andrew and Amanda to be able to write a paper with a time-of-the-day effect that is statistically significant at the 2% level (p< 0.02)?

The probability of throwing a dice 6 times in a row and getting a head each time is one half to the power 6. Write this as (0.5)^6 = 1/64. Hence the probability of this event is less than 2%. So what is the chance that, if I search across all my data, there will be at least one morning or afternoon with a run of a head or a tail? It is 1 – probability there will be neither a Heads Run nor a Tails Run.

Well, there are two types of run, one for heads and one for tails. So the probability of no Heads-or-Tails Run for my experiment during the week is (31/32)^14 = 0.64. *Therefore 36% of the time we will be able to write a paper finding some version of "Heads come up on Wednesday afternoons".* But Amanda is also working in her lab, and also generating data. The probability that EITHER Amanda or I find a result is

1 – probability there will neither a Heads-or-Tails Run nor a Red-or-Black Run.

The probability that there will be neither is (31/32)^28 = 0.41.

Thus 59% of the time we will be able to write a paper proving, in a way that greatly exceeds the ninety-five confidence level, some version of "Heads come

up on Wednesday afternoons" or "Reds occur on Saturday mornings" ...Yet our paper will be wrong. The pattern is an illusion caused by too much searching.

Say we extend our theory to Day of the Year Effects. Say that referees tell us we need to enforce a 0.001 statistical-significance level. We now throw the coin ten times every day for a whole year, and also spin the wheel ten times. The chance of a head coming down ten times in a row is 1/1024. Because there are 365 days in a year, the chance that neither Amanda nor I get any run of 10 in a single day is thus 1 – (511/512)^730 = 0.24.

Hence, 76% of the time we will be able to write a paper proving, at the 0.001 level of statistical significance, some version of Coins Come Down Heads on March 27th… Yet our new paper will be wrong. Again, we have subconsciously searched too much.

When they start to look at data, human beings speedily discard theories and patterns that do not work. Without even being aware of it, they dream up new theories. They latch on to exciting results they had not forecast or expected.

Humans' minds work so flexibly that they can see convincing patterns where there are none. If quizzed by sceptics in seminars, researchers reply: "But my result is significant at the 1% level". This is a pervasive problem; we are all prone to the error, and it is a mistake to be haughty about it. But independent replication is the only convincing check on a finding.

To try to guard against problems, I find (i), (ii), and (iii) helpful, and cautiously recommend them. *18 November 2008, Zurich*

Acknowledgements

I thank Bert Van Landeghem, Steve Stillman and Michael Wolf for helpful discussions. I thank Rainer Winkelmann and his group for their kind hospitality.

Suggested References

- Afshartous, D. and Wolf, M. (2007). 'Avoiding data snooping in multilevel and mixed effects models', Journal of the Royal Statistical Society: Series A, 170, 1035-1059.
- Cumming, G. (2008). 'Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better.' Perspectives on Psychological Science, 3, 286-300.
- Ioannidis, J. P.A. (2005). 'Why most published research findings are false.' PLoS Medicine, 2, 696-701.
- Ioannidis, J. P.A. (2008). 'Why most discovered true associations are inflated.' Epidemiology, 19, 640-648.
- Oswald, AJ and Winkelmann, R. (2011). 'Red wine tastes best at 9pm on November the 18th' Journal of Scientific Discoveries, April 1, 100-98.
- Sterne, JA and Davey Smith, G. (2001). 'Sifting the evidence — What's wrong with significance tests.' British Medical Journal, 322, 226-231.