economist.com

Are results in top journals to be trusted?

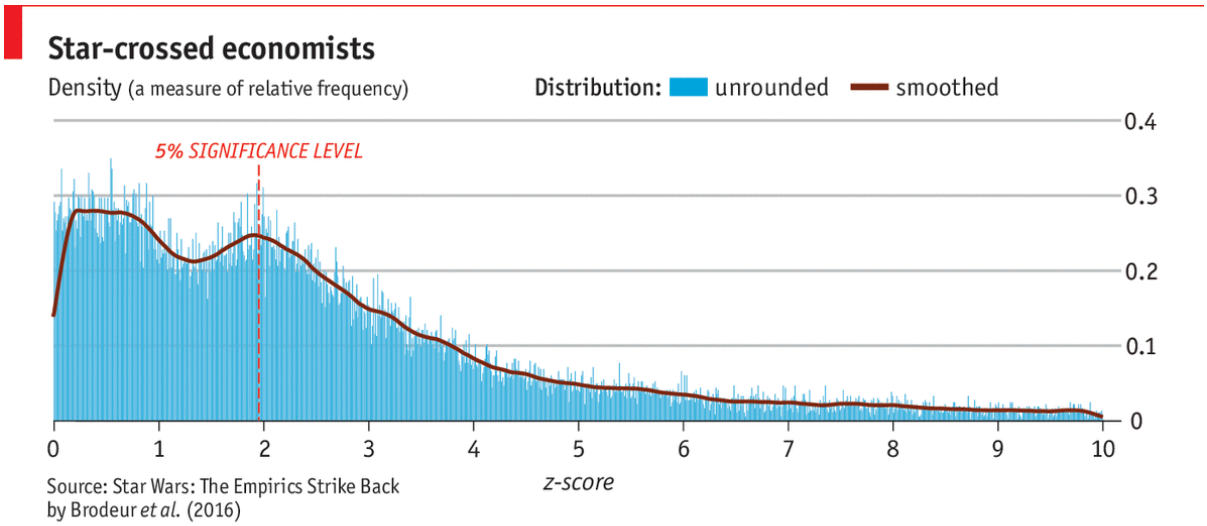
Free exchange | 1 hour 37 mins ago

Are results in top journals to be trusted?

PUBLICATION bias in academic journals is nothing new. A finding of no correlation between sporting events and either violent crime or property crime may be analytically top class, but you couldn't be blamed, frankly, for not giving a damn. But if journal editors are more interested in surprising or dramatic results, there is a danger

that the final selection of published papers offer a distorted vision of reality.

This should skew the distribution of published results, towards more 'significant' findings. But a paper just published in the *American Economic Journal* finds evidence of a different sort of bias, closer to the source. Called "Star Wars, the empirics strike back", it analyses 50,000 papers published between 2005 and 2011 in three top American journals. It finds that the distribution of results (as measured by z-score, a measure of how far away a result is from the expected mean) has a funny double-humped shape (see chart). The dip between the humps represents "missing" results, which just happen to be in a range just outside the standard cut-off point for statistical significance (where significance is normally denoted with starsthough the name may also be something to do with a film recently released—file under 'economists trying to be funny'). Their results suggest that among the results that are only just significant, 10-20% have been fudged.



One explanation is that if a result shows up as significant at the 5% significance level (the industry standard) then researchers crack open the champagne and move on to making economics jokes. But if the result is tantalisingly close to a positive result then perhaps the researchers will fiddle a bit with their method...and celebrate their nice publisher-friendly result. Yanos Zylberbe, one of the paper's authors, explains that in economics it is difficult to conduct controlled experiments, which ultimately gives a lot of freedom to researchers to tweak their methods. Sometimes researchers are tweaking because they want to find the best way of estimating an effect, but sometimes it's in the search for a significant effect. The distinction might be hazy, even in their own minds.

The paper does look at the results split into subgroups, and there seem to be some factors that are associated with a less humpy distribution (which could suggest less fudging). Although the overall pattern holds across all three prestigious journals the paper considers (the *American Economic Review*, the *Quarterly Journal of*

Economics and the *Journal of Political Economy*), papers by older researchers and ones describing randomised control trials have less marked humps—though they are still there. This is worrying for those trying to interpret and communicate the latest research, as it is impossible to tell if there has been foul play in any individual study. But more fundamentally it is worrying for the profession and policymakers making decisions based on economic evidence; fiddling and running multiple, slightly different tests on the same data rapidly sucks meaning from the reported size and accuracy of the final results.

Various solutions have been proposed. One is to publish 'pre-analysis plans', where researchers say how they will do their analysis before they actually do it. Another is to encourage more replication. A new NBER working paper by Marcel Fafchamps and Julien Labonne suggests another, related, method. The idea is that researchers send their data to a third party, who randomly splits the data sample in half. The researchers do their analysis based on the first dataset, finalise their method, and submit for publication. If and when the paper is accepted, the same analysis is carried out on the second sample, and the unadulterated results published. If the initial result only showed up because of manipulation, then the chances of the same result in the second sample are relatively low. To avoid the embarassment of a non-result, researchers should be stricter with themselves when it comes to tweaking their results. When sample sizes are small, this fix is difficult, as halving the sample saps power from tests. But in a world of big data, it could

work. The bigger barrier might be getting career-conscious researchers to sign up.